Annals of Computer Science and Information Systems
Volume 43

Processing of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 14–17, 2025. Kraków, Poland



Marek Bolanowski, Maria Ganzha, Leszek Maciaszek,
Marcin Paprzycki, Dominik Ślęzak (eds.)

PTI

IEEE

Annals of Computer Science and Information Systems, Volume 43

Series editors:
Maria Ganzha (Editor-in-Chief),
Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland
Leszek A. Maciaszek,
Wrocław University of Economics, Poland and Macquarie University, Sydney, Australia
Marcin Paprzycki,
Systems Research Institute, Polish Academy of Sciences, Warsaw and Management Academy, Warsaw, Poland
Dominik Ślęzak,
University of Warsaw, Poland
Marek Bolanowski,
Rzeszow University of Technology, Rzeszów, Poland

Senior Editorial Board:
Wil van der Aalst,
RWTH Aachen University, Netherlands
Enrique Alba,
University of Málaga, Spain
Marco Aiello,
University of Stuttgart, Germany
Mohammed Atiquzzaman,
University of Oklahoma, USA
Christian Blum,
Artificial Intelligence Research Institute (IIIA-CSIC), Spain
Jan Bosch,
Chalmers University of Technology, Sweden
George Boustras,
European University Cyprus, Cyprus
Barrett Bryant,
University of North Texas, USA
Rajkumar Buyya,
University of Melbourne, Australia
Chris Cornelis,
Ghent University, Belgium
Robertas Damaševičius,
Kaunas University of Technology / Vytautas Magnus University, Lithuania
Hristo Djidjev,
Los Alamos National Laboratory, Los Alamos, NM, USA and Institute of Information and Communication Technologies, Sofia, Bulgaria
Włodzisław Duch,
Nicolaus Copernicus University, Toruń, Poland
Schahram Dustdar,
Research Division of Distributed Systems at the TU Wien, Austria and part-time ICREA research professor at UPF, Spain
Hans-George Fill,
University of Fribourg, Switzerland
Ulrich Frank,
Universität Duisburg-Essen, Germany

Ana Fred,
Department of Electrical and Computer Engineering, Instituto Superior Técnico (IST—Technical University of Lisbon), Lisbon, Portugal
Giancarlo Guizzardi,
University of Twente, Netherlands
Francisco Herrera,
University of Granada, Spain
Mike Hinchey,
University of Limerick, Ireland
Janusz Kacprzyk,
Systems Research Institute, Polish Academy of Sciences, Poland
Irwin King,
The Chinese University of Hong Kong, Hong Kong
Michael Luck,
King's College London, United Kingdom
Ivan Luković,
University of Belgrade, Serbia
Marjan Mernik,
University of Maribor, Slovenia
Michael Segal,
Ben-Gurion University of the Negev, Israel
Andrzej Skowron,
University of Warsaw, Poland
John F. Sowa,
VivoMind Research, LLC, USA
George Spanoudakis,
University of London, United Kingdom

Editorial Associates:
Katarzyna Wasielewska,
Systems Research Institute Polish Academy of Sciences, Poland
Paweł Sitek,
Kielce University of Technology, Poland

TEXnical editor: Aleksander Denisiuk,
University of Warmia and Mazury in Olsztyn, Poland

Promotion and Marketing: Anastasiya Danilenka,
Warsaw University of Technology, Poland

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)

Marek Bolanowski, Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Dominik Ślęzak (eds.)

2025, Warszawa, Polskie Towarzystwo Informatyczne

2025, New York City, Institute of Electrical and Electronics Engineers

Annals of Computer Science and Information Systems, Volume 43

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)

ART: ISBN 978-83-973291-7-1, IEEE Catalog Number CFP2585N-ART
Web: ISBN 978-83-973291-6-4

ISSN 2300-5963
DOI 10.15439/978-83-973291-6-4

© 2025, Polskie Towarzystwo Informatyczne
Ul. Solec 38/103
00-394 Warsaw, Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover art:
Paulina Wiewióra,
Elbląg, Poland (https://tonieziemia-otwartapracownia.pl/)

Also in this series:

Volume 45: Communication Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN 978-83-973291-9-5**
Volume 44: Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN 978-83-973291-8-8**
Volume 42: Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering **ISBN 978-83-973291-5-7**
Volume 41: Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-973291-0-2, ISBN USB: 978-83-973291-1-9**
Volume 40: Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-9-1, ISBN USB: 978-83-969601-0-8**
Volume 39: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-6-0, ISBN USB: 978-83-969601-7-7, ISBN ART: 978-83-969601-8-4**
Volume 38: Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering, **ISBN WEB: 978-83-969601-5-3**
Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**
Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2**
Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**
Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**
Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4**
Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0**

DEAR Reader, it is our pleasure to present to you the Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS 2025), which took place on September 14-17, 2025, in Kraków, Poland.

FedCSIS 2025 was chaired by Jarosław Wąs. Moreover, Tomasz Hachaj was the Chair, while Marian Bubak, Marek Grzegorowski and Łukasz Rauch, were the Co-Chairs of the Organizing Committee.

This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute of Polish Academy of Sciences, The Faculty of Mathematics and Information Science Warsaw University of Technology, The Faculty of Electrical and Computer Engineering of the Rzeszów University of Technology and The Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering AGH in cooperation with The Faculty of Metals Engineering and Industrial Computer Science AGH, The Faculty of Materials Science and Ceramics AGH, and Centre for Computational Personalised Medicine SANO.

FedCSIS 2025 was technically co-sponsored by IEEE Poland Section, IEEE Poland Section Computer Society (Gdańsk) Chapter, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Serbia and Montenegro Section Computational Intelligence Society Chapter, IEEE Serbia and Montenegro Section Young Professionals Affinity Group, Committee of Computer Science of the Polish Academy of Sciences and Mazovia Cluster ICT.

FedCSIS 2025 was organized in collaboration with the Strategic Partner QED Software, and sponsored by Intel+Lenovo, Jupiter as well as MDPI Electronics, MDPI Applied Sciences and MDPI AI journals. Moreover, FedCSIS 2025 has been conducted under Honorary Patronages of Professor Jerzy Lis, Rector of the AGH University of Kraków and of Aleksander Miszalski, Mayor of Krakow, as well as under patronages of the Ministry of Digital Affairs of the Republic of Poland, Polish Artificial Intelligence Society (PSSI), Forum Akademickie and Naukowe Towarzystwo Informatyki Ekonomicznej. Finally, media patronage was provided by Krakow.pl, TVP Info, TVP3 Kraków, and Kraków Convetion Bureau.

This year, the structure of the conference remained the same as last year. FedCSIS 2025 had a single Main Track with 5 Topical Areas, which was supported by 12 Thematic Sessions and a Competition. Such structure emphasizes the integrity of the conference and its closeness to the issues that are crucial for the world around us. Here, we recognize the fact that today (in 2025) it is difficult to envision research (and its applications) without an intelligence component. Reflecting this, all five Topical Areas, which constitute the Main Track, represent various aspects of Intelligence Systems. Moreover, the Thematic Sessions provide further insights into selected areas of Intelligence Systems, approached from different perspectives.

In this context, Proceedings that we present to you, consist of six parts. Part 1 contains Invited Contributions. Part 2 collects Main Track full papers (arranged alphabetically, according to the last name of the first author, with the Topical Area represented in the metadata). Part 3 contains Main Track short papers (again, arranged alphabetically, with the Topical Area represented in the metadata). Part 4 contains full papers, originating from the Thematic Sessions (again, arranged alphabetically, according to the last name of the first author, with the name of the Thematic Session stated in the metadata). Part 5 collects short papers from all Thematic Sessions. Finally, Part 6 is devoted to contributions originating from the Competition.

Keeping this in mind, let us now introduce Keynote Speakers, the remaining Invited Contributions, and the five Topical Areas of the FedCSIS 2025 Main Track.

I. Invited Contributions

FedCSIS 2025 invited four keynote speakers to deliver lectures providing a broader context for the conference participants.

- Damaševičius, Robertas, Kaunas University of Technology, Lithuania
  Keynote title: AI-Driven Innovations in Brain Cancer Research
- Dustdar, Schahram, TU Wien, Austria
  Keynote title: Active Inference for Distributed Intelligence in the Computing Continuum
- Jonker, Catholijn, TU Delft (main affiliation), Leiden University, Vrije Universiteit Amsterdam, Netherlands
  Keynote title: Hybrid Human-AI Intelligence to Strengthen the Reflective and Learning Capacity of Organisations
- Plank, Barbara, LMU Munich, Germany
  Keynote title: Human-centered LLMs for Inclusive Language Technology

Moreover, four past FedCSIS keynote speakers have been invited to prepare special contributions, which refer to the core focus of the conference series. These were:

- Atiquzzaman, Mohammed, University of Oklahoma, USA
  Contribution title: Q-ID: A Reinforcement Learning Framework for Adaptive Intrusion Detection
- Blum, Christian, Artificial Intelligence Research Institute, Spain
  Contribution title: Optimizing the Optimizer: An Example Showing the Power of LLM Code Generation
- Luković, Ivan, University of Belgrade, Serbia
  Contribution title: New Education Challenges in Profiling Digital Experts for a Digital Economy Era
- Skowron, Andrzej, Systems Research Institute Polish Academy of Sciences, Poland
  Contribution title: Interactive Granular Computing: Toward Computing Model for Complex Intelligent Systems

At the time, when you are reading this text, videos of the keynote presentations and of invited contributions, delivered

during the FedCSIS 2025 conference, are already available on the official conference website (www.fedcsis.org). We warmly encourage you to visit the website and watch these recordings to gain additional insights and perspectives shared by distinguished speakers.

Finally, as a part of official Conference Opening, a special presentation, entitled: *Paths to Zero Emission Computing—Reducing Energy Consumption, and carbon emissions in HPC and AI environments*, was delivered by Tikiri Wanduragala, Technology Leader Lenovo Infrastructure Solutions Group (ISG), Lenovo UK and Ireland. An extended abstract, outlining main pints of this presentation can be found in the Position papers of FedCSIS 2025 volume, published in the Annals of Computer Science and Information Systems book series.

II. Advanced Artificial Intelligence in Applications

This Topical Area covers a wide range of core aspects of intelligence systems, with a particular focus on AI. Nowadays, AI is usually perceived as closely related to the data, therefore, the scope of this Topical Area includes, among others, elements of machine learning, data science, and big data processing, with important emerging aspects such as interactive learning and human-centered AI, as well as interpretable learning, explainable AI, and trustworthiness. Furthermore, since the realm of AI is far richer, the ultimate goal of this Topical Area is to show relationships between all of the currently pursued AI subareas, emphasizing a cross-disciplinary nature of various research branches. From the perspective of the general scope of FedCSIS, this Topical Area embraces AI methods and examples of their applications in different practical fields.

This Topical Area was curated by:

- Artiemjew, Piotr, University of Warmia and Mazury, Poland
- Chelly Dagdia, Zaineb, University of Lille, France
- Corizzo, Roberto, American University, USA
- Zdravevski, Eftim, Ss. Cyril and Methodius University, Northern Macedonia

III. Computer Science & Systems

This Topical Area covers technical (or applicable) aspects of computer science and related disciplines. The Topical Area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science, and to issues of interest to distributed, smart, data-oriented, multimedia and network systems.

The Topical Area is oriented on the research where the computer science meets the real world problems, real constraints, simulations and processing, model objectives, etc., in order to deliver Intelligence Systems. However the scope is not limited to applications, we all know that all of them were born from the innovative theory developed in the laboratory. We want to show the fusion of these two worlds. Therefore one of the goals for the Topical Area is to show how the idea is transformed into application, since the history of modern science shows the most successful research experiments had their continuation in the real world.

This Topical Area was curated by:

- Ferrando, Angelo, University of Modena and Reggio Emilia, Italy
- Pawłowski, Wiesław, University of Gdańsk and Systems Research Institute, Polish Academy of Sciences, Poland
- Wasielewska-Michniewska, Katarzyna, Systems Research Institute, Polish Academy of Sciences, Poland

IV. Network Systems & Applications

Modern network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services, and applications. This results in numerous active research areas oriented towards various technical, scientific, and social aspects of network systems and applications. The rapid development of computer networks including wired and wireless networks observed today is evolving, dynamic, and multidimensional. On the one hand, network technologies are used in the majority of areas that make human life easier and more comfortable. On the other hand, the rapid need for network deployment brings new challenges in network management and network design, which are reflected in hardware, software, services, and security-related problems. Every day, a new solution in the field of technology and applications of computer networks is released. New network vulnerabilities, threats and attacks are also emerging. This Topical Area is devoted to emphasizing up-to-date topics in networking systems and technologies by covering problems and challenges related to intensive multidimensional network developments. Yet, it is inclusive and spans a wide spectrum of networking-related topics.

This Topical Area was curated by:

- Armando, Alessandro, University of Genova, Italy
- Furtak, Janusz, Military University of Technology, Poland
- Suri, Niranjan, Institute of Human and Machine Cognition, United States

V. Information Technology for Business & Society

This Topical Area emphasizes Intelligence Systems as a central theme while embracing a comprehensive perspective on the role of various technologies, Management Information Systems (MIS), and Decision Support Systems (DSS) in addressing contemporary challenges faced by businesses, organizations, and society. It seeks to provide a platform for exploring how advanced technologies–such as big data, data mining, machine learning, IoT, blockchain, cloud computing, and social networks–interact with and contribute to the continuous improvement of processes, decision-making, and innovation in diverse domains, including business, government, and social sectors.

The aim is to comprehensively understand how Intelligence Systems and related technologies can be effectively leveraged within MIS and DSS to drive innovation, support strategic

goals, and create value across various sectors and domains. This focus goes beyond the technical dimensions, integrating socio-technical and management perspectives to explore how these systems can be designed, implemented, and managed to meet real-world needs. This includes addressing challenges in organizational processes, decision-making frameworks, and strategies for achieving sustainable and continuous improvement.

This Topical Area was curated by:

- Cano, Alberto, Virginia Tech University, USA
- Dias, Gonçalo, University of Aveiro, Portugal
- Miller, Gloria, maxmetrics, Germany
- Naldi, Maurizio, LUMSA University, Italy
- Wątróbski, Jarosław, University of Szczecin, Poland
- Ziemba, Ewa Wanda, University of Economics in Katowice, Poland

VI. Software, System & Service Engineering

The Topical Area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. It investigates both established traditional approaches and modern emerging approaches to large software production and evolution.

For decades, it is still an open question in software industry, how to provide fast and effective software process and (intelligent) software services, and how to come to the software systems, embedded systems, autonomous systems, intelligence systems, or cyber-physical systems that will address the open issue of supporting information management process in many, particularly complex organization systems. Even more, it is a hot issue how to provide a synergy between systems in common and software services as mandatory component of each modern organization, particularly in terms of IoT, Big Data, Data Science, Artificial Intelligence, Machine Learning, and Industry 4.0 paradigms.

In recent years, we are the witnesses of great movements in the area of software, system and service engineering (S3E). Such movements are both of technological and methodological nature. By this, today we have a huge selection of various technologies, tools, and methods as a discipline that helps in a support of the whole information life cycle in organization systems. Despite that, one of the hot issues in practice is still how to effectively develop and maintain complex systems from various aspects, particularly when software components are crucial for addressing declared system goals, and their successful operation. It seems that nowadays we have great theoretical potentials for application of new and more effective approaches. However, it is more likely that real deployment of such approaches in industry practice is far behind their theoretical potentials.

This Topical Area was curated by:

- Luković, Ivan, University of Belgrade, Serbia
- Kolukısa Tarhan, Ayça, Hacettepe University, Turkey
- Popović, Aleksandar, University of Montenegro, Podgorica, Montenegro

VII. FedCSIS 2025 Competition—Predicting Chess Puzzle Difficulty, Second Edition

FedCSIS 2025 Challenge was entitled: Predicting Chess Puzzle Difficulty—Second Edition. It was the 11th data science challenge, organized within the scope of FedCSIS conference series. The goal was to build a model to predict the difficulty (measured as Lichess rating) of given chess puzzles. The 2025 competition was organized by:

- Ślęzak, Dominik, QED Software and University of Warsaw, Poland
- Ślęzak, Michał, Polish-Japanese Academy of Information Technology, Poland
- Świechowski, Maciej, Grail Team, Poland
- Zyśko, Jan, University of Warsaw, Poland

This year, 42 teams participated in the competition. They, collectively, submitted 1185 solutions. Team members represented 16 different countries from around the world, with largest number of representatives coming from Poland, Germany, United States and Singapore. After evaluation, the following contributions, found also in these proceedings, discuss the winning solutions:

- First place: Sebastian Björkqvist, Estimating the Difficulty of Chess Puzzles by Combining Fine-Tuned Maia-2 with Hand-Crafted and Engine Features
- Second place: Tyler Woodruff, Luke Imbing, Marco Cognetta, The bread emoji Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge
- Third place: Szymon Miłosz, Pretraining Transformers for Chess Puzzle Difficulty Prediction

VIII. Professor Zdzisław Pawlak Award

The above-described five Topical Areas of the FedCSIS Main Track reflect the five fundamental aspects of understanding, developing, and applying Intelligence Systems. This topical integrity is emphasized by the Professor Zdzisław Pawlak Award, presented in four categories: Best Paper, Young Researcher, Industry Cooperation, and International Cooperation. Here, note that although Professor Zdzisław Pawlak has been often recognized as "the father of Polish AI", his research achievements have gone far beyond AI itself, in particular toward AI applications and Intelligence Systems as we understand them. Accordingly, for this award contributions from the Main Track and from all Thematic Sessions are considered. This year, the following contributions have been awarded:

- In the category *Best Paper*: Preiß, Niklas and Westner, Markus, From Agents to Copilots: A Systematic Review of Digital Assistant Technology Adoption in Proprietary Productivity Software
- In the category *Young Researcher*: Lorenz, Alisa, A New Dimension of Acceptance? Introducing Perceived Public Value as Extension of UTAUT in the Smart City Context
- In the category *Industry Cooperation*: Sadel, Jakub, Tarasiewicz, Tomasz, Kowaleczko, Paweł, Ziaja, Ma-

ciej, Kostrzewa, Daniel, Benecki, Paweł and Kawulok, Michał, Keypoint-based metric for evaluating image super-resolution quality
- In the category *International Cooperation*: De Maeyer, Wannes, Van Overberghe, Steven, Cornelis, Chris and Restrepo Lopez, Mauricio, Estimating the entropy of covering-based rough set approximation operators

Young Researcher Award was sponsored by the MDPI Applied Sciences Journal, International Cooperation Award was sponsored by the MDPI AI Journal, Industry Cooperation Award was sponsored by the MDPI Electronics Journal, while the Best Paper Award was sponsored by the Mazovia Branch of Polish Information Processing Society.

Here, let us also note that Professor Zdzisław Pawlak Awards have been presented to authors of best papers since 2006.

IX. Statistics

Each contribution, found in this volume, was refereed by at least two referees and the acceptance rate of regular full papers was approximately 18.66% (47 accepted contributions, out of a total of 209 submissions representing 48 countries and 6 continents). Since we live in times when "data is the new oil", the long-term trend of acceptance of regular full papers is depicted in Figure 1. Note that this year the lowest, thus far, acceptance rate for regular full papers has been recorded.

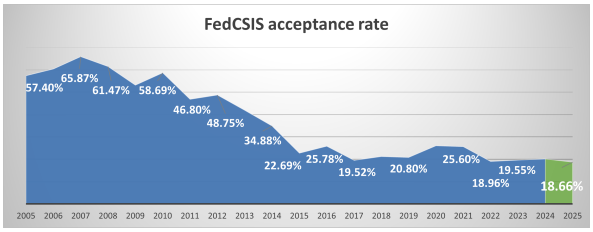


Figure 1. Acceptance rate for the regular full papers for the FedCSIS conference series since 2005 (when FedCSIS predecessor was organized for the first time).

X. Committees

The Senior Program Committee of FedCSIS 2025 consisted of:

- Arciuch, Artur, Military University of Technology, Poland
- Armando, Alessandro, University of Genova, Italy
- Artiemjew, Piotr, University of Warmia and Mazury in Olsztyn, Poland
- Bicevska, Zane, University of Latvia, Latvia
- Bumberger, Jan, Helmholtz-Centre for Environmental Research – UFZ, Germany
- Bylina, Beata, Maria Curie-Sklodowska University, Poland
- Bylina, Jarosław, Marie Curie-Sklodowska University, Poland
- Cano, Alberto, Virginia Commonwealth University, United States
- Charvat, Karel, Czech Center for Science and Society, Czech Republic
- Chelly Dagdia, Zaineb, University of Lille, France
- Chmielarz, Witold, University of Warsaw, Poland
- Cooper, Anthony-Paul, Durham University, United Kingdom
- Corizzo, Roberto, American University, United States
- Coronato, Antonio, Università Giustino Fortunato, Italy
- Cyganek, Bogusław, AGH University of Krakow, Poland
- Dan, Daniel, MODUL University, Austria
- Debar, Hervé, Télécom SudParis, Institut Mines-Télécom, France
- Dias, Gonçalo, University of Aveiro, Portugal
- Dörpinghaus, Jens, Federal Institute for Vocational Education and Training (BIBB), Germany
- Felkner, Anna, NASK—National Research Institute, Poland
- Fernandes De Muylder, Cristiana, Universidade FUMEC, Brazil
- Ferrando, Angelo, University of Modena and Reggio Emilia, Italy
- Fidanova, Stefka, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- Fornés Leal, Alejandro, Universitat Politècnica de València, Spain
- Franczyk, Bogdan, University of Leipzig, Germany
- Frank, Urlich, Universität Duisburg-Essen, Germany
- Furtak, Janusz, Military University of Technology, Poland
- Häckl, Dennis, University of Leipzig, Germany
- Helmrich, Robert, Federal Institute for Vocational Education and Training (BIBB), Germany
- Ioannidis, Sotirios, Technical University of Crete, Greece
- Iwanowski, Marcin, Warsaw University of Technology, Poland
- Janousek, Jan, Czech Technical University, Czech Republic
- Kadobayashi, Youki, Nara Institute of Science and Technology, Japan
- Kobyliński, Łukasz, Institute of Computer Science, Polish Academy of Sciences, Poland
- Kolukısa Tarhan, Ayça, Hacettepe University, Turkey
- Kubis, Marek, Adam Mickiewicz University, Poland
- Kwaśnicka, Halina, Wroclaw University of Technology, Poland
- Lacalle Úbeda, Ignacio, Universitat Politècnica de València, Spain
- Leyh, Christian, Technische Hochschule Mittelhessen (THM) – University of Applied Sciences, Germany
- Lirkov, Ivan, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- Luković, Ivan, University of Belgrade, Serbia
- Martinelli, Massimo, Institute of Information Science and Technologies, National Research Council of Italy, Italy
- Milašinović, Boris, University of Zagreb, Croatia
- Militzer-Horstmann, Carsta, WIG2 Institute for Health

Economics and Health Service Research, Germany
- Miller, Gloria, maxmetrics, Germany
- Moroni, Davide, Institute of Information Science and Technologies, National Research Council of Italy, Italy
- Mucherino, Antonio, IRISA, University of Rennes, France
- Naeem, Muddasar, Università Giustino Fortunato, Italy
- Nakayama, Minoru, Institute of Science Tokyo, Japan
- Naldi, Maurizio, LUMSA University, Italy
- Pałys, Tomasz, Military University of Technology, Poland
- Paszkiewicz, Andrzej, Rzeszow University of Technology, Poland
- Pawłowski, Wiesław, University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland
- Popović, Aleksandar, University of Montenegro, Montenegro
- Procházka, Ales, University of Chemistry and Technology, Czech Republic
- Rangel Henriques, Pedro, University of Minho, Portugal
- Reinhold, Olaf, Social CRM Research Center, Germany
- Rishiwal, Vinay, Mahatma Jyotiba Phule Rohilkhand University, India
- Ristić, Sonja, University of Novi Sad, Serbia
- Rizvi, Syed Tahir Hussain, Wenn AsA, Sandnes, Norway
- Slivnik, Boštjan, University of Ljubljana, Slovenia
- Śluzek, Andrzej, Warsaw University of Life Sciences—SGGW, Poland
- Sołtysik-Piorunkiewicz, Anna, University of Economics in Katowice, Poland
- Speckesser, Stefan, University of Brighton, United Kingdom
- Stanescu, Liana, University of Craiova, Romania
- Stpiczyński, Przemysław, Maria Curie-Sklodowska University, Poland
- Suri, Niranjan, Institute of Human and Machine Cognition, United States
- Świątkowska, Joanna, Deputy Secretary General at European Cyber Security Organisation, Belgium
- Tiemann, Michael, Federal Institute for Vocational Education and Training (BIBB), Germany
- Tolvanen, Juha-Pekka, MetaCase, Finland
- Varanda Pereira, Maria Joao, Instituto Politecnico de Braganca, Portugal
- Wasielewska-Michniewska, Katarzyna, Systems Research Institute, Polish Academy of Sciences, Poland
- Wątróbski, Jarosław, West Pomeranian University of Technology, Poland
- Wróblewska, Anna, University of Technology, Poland
- Zaharie, Daniela, West University of Timisoara, Romania
- Zdravevski, Eftim, Ss. Cyril and Methodius University, North Macedonia
- Ziemba, Ewa, University of Economics in Katowice, Poland

The Program Committee of FedCSIS 2025 consisted of:

- Abromavičius, Vytautas, Vilnius Gediminas Technical University, Lithuania
- Adil, Muhammad, Institute of Marine Sciences CNR, Italy
- Ahmad, Muhammad Ovais, Karlstad University, Sweden
- Amodio, Pierluigi, Universita' di Bari, Italy
- Anastassi, Zacharias, ASPETE School of Pedagogical and Technological Education, Greece
- Arabas, Piotr, NASK—National Research Institute, Poland
- Asselborn, Thomas, University of Hamburg, Germany
- Awuni Kolog, Emmanuel, University of Ghana, Ghana
- Azad, Mohammad, Jouf University, Saudi Arabia
- Azevedo, Ana, CEOS.PP, ISCAP, P.PORTO, Portugal
- Bacco, Manlio, ISTI-CNR, Italy
- Badica, Costin, University of Craiova, Romania
- Badica, Amelia, University of Craiova, Romania
- Badr, Georges, Holy Spirit University of Kaslik, Lebanon
- Baldán Lozano, Francisco Javier, University of Málaga, Spain
- Banach, Richard, University of Manchester, United Kingdom
- Barišić, Ankica, ActiveEon, France
- Belciug, Smaranda, University of Craiova, Romania
- Ben-Assuli, Ofir, Ono Academic College, Israel
- Beuran, Razvan, Japan Advanced Institute of Science and Technology, Japan
- Białas, Andrzej, Research Network Łukasiewicz—Institute of Innovative Technologies EMAG, Poland
- Bicevskis, Janis, University of Latvia, Latvia
- Bieniek, Mariusz, University of Maria Curie Skłodowska, Poland
- Bjeladinović, Srđa, University of Belgrade, Serbia
- Boehm, Karsten, FH KufsteinTirol—University of Applies Science, Austria
- Borkowski, Boleslaw, Warsaw University of Life Sciences, Poland
- Bountakas, Panagiotis, Sphynx Technology Solutions AG, University of Piraeus, Switzerland
- Brdjanin, Drazen, University of Banja Luka, Bosnia and Herzegovina
- Brezovan, Marius, University of Craiova, Romania
- Bridova, Ivana, University of Zilina, Slovakia
- Brzoza-Zajęcka, Ada, AGH University of Krakow, Poland
- Cabri, Giacomo, Università di Modena e Reggio Emilia, Italy
- Calpe Maravilla, Javier, University of Valencia, Spain
- Carbone, Roberto, Security & Trust Unit, FBK, Italy
- Carchiolo, Vincenza, Universita di Catania, Italy
- Casalino, Gabriella, University of Bari Aldo Moro, Italy
- Chen, Haiming, Chinese Academy of Sciences, China
- Chiru, Costin, University Politehnica of Bucharest, Romania
- Christozov, Dimitar, American University in Bulgaria, Bulgaria
- Cicconetti, Claudio, IIT-CNR, Italy

- Ciucci, Davide, Università di Milano-Bicocca, Italy
- Colantonio, Sara, ISTI-CNR, Italy
- Corizzo, Roberto, American University, United States
- Cosulschi, Mirel, University of Craiova, Romania
- Coviello, Giuseppe, Politecnico Di Bari, Italy
- Cybulski, Piotr, Military University of Technology in Warsaw, Poland
- Czachorski, Tadeusz, IITiS PAN, Polish Academy of Sciences, Poland
- Czarnacka-Chrobot, Beata, Warsaw School of Economics, Poland
- Czarnul, Pawel, Gdansk University of Technology, Poland
- Damasevicius, Robertas, Vytautas Magnus University, Poland
- de Lara, Juan, Universidad Autonoma de Madrid, Spain
- Defigueiredo, Felipeaugustopereira, Ghent Univerity, Belgium
- Derezinska, Anna, Warsaw University of Technology, Poland
- Dery, Lihi, Ariel University, Israel
- Dimitrieski, Vladimir, Faculty Of Technical Sciences, Serbia
- Dionysiou, Antreas, University of Cyprus, Cyprus
- Dudycz, Helena, Wroclaw University of Economics, Poland
- Dutta, Arpita, National University of Singapore, Singapore
- Eisenbardt, Monika, Univeristy of Economics Katowice, Poland
- Eisenbardt, Tomasz, University of Warsaw, Poland
- Engelbrecht, Andries, University of Stellenbosch, South Africa
- Erata, Ferhat, Yale University, United States
- Faber, Łukasz, AGH University of Krakow, Poland
- Fajsi, Angela, University of Novi Sad, Serbia
- Faonio, Antonio, Eurecom, France
- Fertalj, Krešimir, University of Zagreb, Croatia
- Fialko, Sergiy, Cracow University of Technology, Poland
- Fischer, Andreas, f-bb, Germany
- Frommholtz, Ingo, MODUL University, Austria
- Frommholz, Ingo, MODUL University, Austria
- Fuentes, Alvaro, Jeonbuk National University, South Korea
- Furnell, Steven, University of Nottingham, United Kingdom
- G.-Tóth, Boglárka, University of Szeged, Hungary
- García-Mireles, Gabriel, Universidad de Sonora, Mexico
- García, Laura, Universidad Politécnica de Cartagena, Spain
- Garitano, Iñaki, Mondragon Unibertsitatea, Spain
- Gawkowski, Piotr, Warsaw University of Technology, Poland
- Gburzyński, Paweł, Vistula University, Poland
- Gepner, Paweł, Warsaw University of Technology, Poland
- Gerhards, Christian, Federal Institute for Vocational Education and Training (BIBB), Germany
- Geri, Nitza, The Open University of Israel, Israel
- Gezici Geçer, Bahar, Hacettepe University, Turkey
- Gheisari, Mehdi, Islamic Azad University, Iran
- Gimenez, Pierre-François, Inria, France
- Gołuchowski, Jerzy, Katowice University of Economics
- Gomolińska, Anna, University of Białystok, Poland
- Góźdź, Marek, Maria Curie-Skłodowska University, Poland
- Grabara, Dariusz, University of Economics in Katowice, Poland
- Gravvanis, George, Democritus University of Thrace, Greece
- Grochla, Krzysztof, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland
- Gurgen, Tugba, Hacettepe University, Turkey
- Halawi, Leila, Embry Riddle Aeronautical Univ, United States
- Hamel, Oussama, University Blida, Algeria
- Hammi, Badis, Telecom-Sud Paris, France
- Hanslo, Ridewaan, University of Johannesburg, South Africa
- Hapka, Aneta, Koszalin Univeristy of Technology, Poland
- Harezlak, Katarzyna, Silesian University of Technology, Poland
- Hasso, Hussein, Fraunhofer FKIE, Germany
- Heidler, Tobias, WIG2 Institute for Health Economics and Health System Research, Germany
- Hein, Kristine, Federal Institute for Vocational Education and Training (BIBB), Germany
- Helsingius, Mika, Finnish Defence Research Agency, Finland
- Hernes, Marcin, Wrocław University of Economics and Business, Poland
- Hodoň, Michal, University of Zilina, Slovakia
- Homa, Jarosław, Silesian University of Technology, Poland
- Horváth, Zoltán, Eotvos Lorand University, Hungary
- Hosobe, Hiroshi, Hosei University, Japan
- Houben, Timo, Helmholtz Centre for Environmental Research—UFZ, Germany
- Hrach, Christian, Institute for Applied Informatics (InfAI), Germany
- Hussain, Asad, Reading University, United Kingdom
- Ienco, Dino, IRSTEA, France
- Ilia, Panagiotis, Cyprus University of Technology, Cyprus
- Ionescu, Mihaela, University of Medicine and Pharmacy of Craiova, Romania
- Ivanovic, Mirjana, University of Novi Sad, Serbia
- Jakovljević, Nikša, University of Novi Sad, Serbia
- Janicki, Ryszard, McMaster University, Canada
- Janicki, Artur, Warsaw University of Technology, Poland
- Janiszewski, Marek, NASK—National Research Institute, Poland
- Janousek, Jan, Czech Technical University, Czech Republic

- Jarzebowicz, Aleksander, Gdansk University of Technology, Poland
- Jaskóła, Przemysław, NASK—National Research Institute, Poland
- Jassem, Krzysztof, Adam Mickiewicz University, Poland
- Jelonek, Dorota, Czestochowa University of Technology, Poland
- Johnsen, Frank, Norwegian Defence Research Establishment (FFI), Norway
- Jovancevic, Igor, University of Montenegro, Montenegro
- Jovanovik, Milos, Ss. Cyril and Methodius University in Skopje, North Macedonia
- Kalibatiene, Diana, VGTU, Lithuania
- Kaloyanova, Kalinka, University of Sofia, Bulgaria
- Kamola, Mariusz, NASK—National Research Institute, Poland
- Kanciak, Krzysztof, Military University of Technology, Poland
- Kapczyński, Adrian, Polish Information Processing Society, Poland
- Kardas, Geylani, Ege University International Computer Institute, Turkey
- Karpinska, Marzena, University of Massachusetts Amherst, United States
- Kasprzak, Włodzimierz, Warsaw University of Technology, Poland
- Keir, Paul, University of the West of Scotland, United Kingdom
- Kelner, Jan, Military University of Technology, Poland
- Kenett, Yoed, Israel Institute of Technology, Israel
- Khan, Adil, Xi'an Eurasia University, China
- Khlif, Wiem, FSEGS, Tunisia
- Kibilda, Jacek, Virginia Tech, United States
- Klapp, Iftach, Agriculture research Organization—Volcani Institute, Israel
- Kobylinski, Andrzej, Warsaw School of Economics, Poland
- Kołodziej, Joanna, Krakow University of Technology, Poland
- Kononova, Anna V., LIACS, Leiden University, Netherlands
- Kordić, Slavica, Faculty of Technical Sciences, Serbia
- Kosmopoulos, Dimitrios, University of Patras, Greece
- Kossowska-Lai, Magdalena, Jagiellonin University, Poland
- Kosta, Sokol, Aalborg University, Denmark
- Kozak, Jan, University of Economics in Katowice, Poland
- Krajsic, Philippe, University of Leipzig, Germany
- Krdzavac, Nenad, Leibniz Information Centre for Science and Technology (TIB), Germany
- Kucaba-Pietal, Anna, Rzeszów University of Technology, Poland
- Kucharska, Edyta, AGH University of Krakow, Poland
- Kurasova, Olga, Institute of Mathematics and Informatics, Lithuania
- Kurzyna, Marcin, UMCS, Poland
- Kwasnicka, Halina, Wroclaw University of Technology, Poland
- Kwater, Tadeusz, PANS Jaroslaw, Poland
- L'Ecuyer, Pierre, University of Montreal, Canada
- La Cascia, Marco, University of Palermo, Italy
- Laborde, Romain, Université de Toulouse, France
- Lasek, Piotr, University of Rzeszów, Poland
- Laskov, Lasko, New Bulgarian University, Bulgaria
- Lastovetsky, Alexey, University College Dublin, Ireland
- Leka, Blerta, Agricultural University of Tirana, Albania
- Lerga, Jonatan, Faculty of Engineering & Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Croatia
- Levillain, Olivier, Samovar, Télécom SudParis, Institut Polytechnique de Paris, France
- Lewandowski, Piotr, NASK—National Research Institute, Poland
- Ligeza, Antoni, AGH University of Krakow, Poland
- Li, Tianrui, Southwest Jiaotong University, China
- Li, Shujun, University of Kent, United Kingdom
- Ljubić, Sandi, University of Rijeka, Croatia
- Lloret, Jaime, Universitat Politècnica de València, Spain
- Ltaief, Hatem, King Abdullah University of Science and Technology, Saudi Arabia
- Lucas-Estañ, M. Carmen, University Miguel Hernandez of Elche, Spain
- Luszczek, Piotr, University of Tennessee Knoxville, United States
- Machado, José, University of Minho, DI, ALGORITMI/LASI, Portugal
- Majdik, András, HUN-REN SZTAKI—Hungarian Research Network, Hungary
- Mangroliya, Meetkumar Pravinbhai, University of Koblenz, Germany
- Manso, Marco, PARTICLE LTD, Portugal
- Marchiori, Massimo, UNIPD and EISMD, Italy
- Marcińczuk, Michał, Samurai Labs, Poland
- Marciniak, Jacek, Adam Mickiewicz University, Poland
- Marciniak, Małgorzata, Institute of Computer Science, Polish Academy of Sciences, Poland
- Marcinkowski, Bartosz, University of Gdansk, Poland
- Marciszewska, Barbara, Gdynia Maritime University, Poland
- Marghitu, Daniela, Auburn University, United States
- Markakis, Evangelos, Hellenic Mediterranean University, Greece
- Marko, Oskar, BioSense Institute, Serbia
- Martín-Navarro, Alicia, INDESS. University of Cádiz, Spain
- Martínez López, Pablo E., Universidad Nacional de Quilmes, Argentina
- Matson, Eric, Purdue University, United States
- Medeišis, Artūras, Vilnius Gediminas Technical University, Lithuania
- Mele, Valeria, University of Naples Federico II, Italy
- Melzer, Sylvia, University of Hamburg, Germany

- Meneses, Claudio, Universidad Católica del Norte, Chile
- Merkel, Manuel, University of Stuttgart, Germany
- Micota, Flavia, West University of Timisoara, Romania
- Mignone, Paolo, Università degli studi di Bari, Italy
- Mihaescu, Marian Cristian, University of Craiova, Romania
- Mihajlov, Martin, Jozef Stefan Institute, Slovenia
- Milašinović, Boris, University of Zagreb, Croatia
- Mildorf, Tomas, University of West Bohemia, Czech Republic
- Milella, Annalisa, CNR-STIIMA, Italy
- Millham, Richard, Durban University of Technology, South Africa
- Misra, Nn, Dublin Institute of Technology, Ireland
- Misra, Sanjay, Institute For Energy Technology, Norway
- Mocanu, Mihai, University of Craiova, Romania
- Modoni, Gianfranco, STIIMA-CNR, Italy
- Mollenhauer, Hannes, Helmholtz Centre for Environmental Research—UFZ, Germany
- Moraca, Slobodan, Faculty of Technical Sciences, Serbia
- Moshkov, Mikhail, KAUST, Saudi Arabia
- Mozgovoy, Maxim, University of Aizu, Japan
- Mullins, Roisin, University of Wales Trinity Saint David, United Kingdom
- Mumm, Rebekka, WIG2 Institute for Health Economics and Health System Research, Germany
- Muñoz, Andres, Universidad de Cádiz, Spain
- Muszyńska, Karolina, University of Szczecin, Poland
- Mykowiecka, Agnieszka, Institute of Computer Science, PAS and Polish-Japanese Academy of Information Technology, Poland
- Myszkowski, Pawel, Wroclaw University of Science and Technology, Poland
- Nguyen, Khuong An, University of London, United Kingdom
- Niewiadomska-Szynkiewicz, Ewa, Warsaw University of Technology, Poland
- Ota, Daniel, Fraunhofer FKIE, Germany
- Ozkan, Necmettin, Gebze Technical University, Turkey
- Oztutus, Ferhat, Hakkari University, Turkey
- Paliwoda-Pękosz, Grażyna, Krakow University of Economics, Poland
- Palma, Raul, Poznan Supercomputing and Networking Center, Poland
- Palmigiano, Alessandra, the Vrije Universiteit Amsterdam, Netherlands
- Paluszyński, Wiesław, Polish Information Processing Society, Poland
- Pandey, Dr. Rajiv, Amity University, India
- Pankowska, Malgorzata, University of Economics in Katowice, Poland
- Pataricza, András, Budapest University of Technology and Economics, Hungary
- Patterson, Alfred, University of Turku, Finland
- Pazienza, Andrea, NTT DATA Italia SpA & A3K Srl, Italy
- Peña, Josem, CSIC, Spain
- Petcu, Dana, West University of Timisoara, Romania
- Peukert, Hagen, University of Hamburg, Germany
- Pires, Ivan Miguel, Instituto de Telecomunicações, ESTGA, Universidade de Aveiro, Portugal
- Popescu, Elvira, University of Craiova, Romania
- Popović, Aleksandar, University of Montenegro, Montenegro
- Pop, Florin, Politehnica Bucharest, Romania / ICI Bucharest, Romania
- Porta, Marco, University of Pavia, Italy
- Po, Laura, Universitá di Modena e Reggio Emilia, Italy
- Przybyła-Kasperek, Małgorzata, University of Silesia in Katowice, Poland
- Ptaszynski, Michal, Kitami Institute of Technology, Japan
- Quessette, Franck, PRISM—University of Versailles, France
- Ramanna, Sheela, University of Winnipeg, Canada
- Rangel Henriques, Pedro, University of Minho, Portugal
- Rauch, Jan, Prague University of Economics and Business, Czech Republic
- Rechavi, Amit, Ruppin Academic Center, Israel
- Reformat, Marek, University of Alberta, Canada
- Reiser, Thomas, University of Koblenz, Germany
- Riedinger, Constanze, HTWG Konstanz, Germany
- Ristic, Sonja, University of Novi Sad, Serbia
- Rojek, Krzysztof, Czestochowa University of Technology, Poland
- Rossi, Bruno, Masaryk University, Czech Republic
- Roszczyk, Radosław, Warsaw University of Technology, Poland
- Rozevskis, Uldis, University of Latvia, Latvia
- Rytel, Marcin, NASK—National Research Institute, Poland
- Sachdeva, Shelly, National Institute of Technology Delhi, India
- Salvetti, Ovidio, Institute of Information Science and Technologies, CNR, Italy
- Sancho, Paula, University of Cádiz, Spain
- Saracino, Andrea, Scuola Superiore Universitaria Sant'Anna di Pisa, Italy
- Saraiva, João, University of Minho, Portugal
- Sari, Riri Fitri, University of Indonesia, Indonesia
- Sawerwain, Marek, University of Zielona Góra, Poland
- Schreiber, Celine, University Leipzig, Germany
- Schreiner, Wolfgang, Johannes Kepler University Linz, Austria
- Scozzari, Andrea, CNR ISTI, Italy
- Segedinac, Milan, Faculty of Technical Scieneces, Serbia
- Shah, Syed Ihtesham Hussain, Vrije University Amesterdam, Netherlands
- Shaska, Tony, Oakland University, United States
- Shen, Kao-Yi, Chinese Culture University, Taiwan
- Sidje, Roger B., University of Alabama, United States
- Sierra, Jose Luis, Universidad Complutense de Madrid, Spain

- Sifaleras, Angelo, University of Macedonia, Greece
- Sikorski, Marcin, Gdansk University of Technology, Poland
- Silva, Lincoln, Departamento de Tecnologia da Informação e Educação em Saúde—Faculdade de Ciências Médicas da UERJ, Brazil
- Skarmeta, Antonio, Universidad de Murcia, Spain
- Skórzewski, Paweł, Adam Mickiewicz University, Poland
- Skubalska-Rafajłowicz, Ewa, Wrocław University of Science and Technology, Poland
- Slivnik, Boštjan, University of Ljubljana, Slovenia
- Sosnowski, Łukasz, Systems Research Institute, Polish Academy of Sciences, Poland
- Sowański, Marcin, Samsung Research Poland, Poland
- Söylemez, Mehmet, Hacettepe University, Turkey
- Stanczyk, Urszula, Silesian University of Technology, Poland
- Stark, Sandra, Leipzig University, Germany
- Stavness, Ian, University of Saskatchewan, Canada
- Steiner, Petra, Federal Institute for Vocational Education and Training (BIBB), Germany
- Štěpánek, Lubomír, Faculty of Medicine, Charles University; Faculty of Informatics and Statistics, University of Economics, Czech Republic
- Stoean, Ruxandra, University of Craiova, Romania
- Stoica, Cosmin, University of Craiova, Romania
- Stoicu-Tivadar, Lăcrămioara, "Politehnica" University Timisoara, Romania
- Stój, Jacek, Silesian University of Technology, Poland
- Strzelecki, Artur, University of Economics in Katowice, Poland
- Stutzer, Franziska, WIG2 Institute for Health Economics and Health System Research, Germany
- Subbotin, Sergey, National University "Zaporizhzhia Polytechnic", Ukraine
- Suraj, Zbigniew, University of Rzeszów, Poland
- Swacha, Jakub, University of Szczecin, Poland
- Świechowski, Maciej, QED Software, Poland
- Sysko-Romańczuk, Sylwia, Warsaw University of Technology, Poland
- Syta, Jakub, Akademia Marynarki Wojennej w Gdyni, Poland
- Szantoi, Zoltan, European Space Agency, Italy
- Szczech, Izabela, Poznan University of Technology, Poland
- Szczuka, Marcin, The University of Warsaw, Poland
- Szmit, Maciej, University of Lodz, Poland
- Szumski, Oskar, University of Warsaw Faculty of Management, Poland
- Szustak, Lukasz, Czestochowa University of Technology, Poland
- Ślęzak, Michał, Polish-Japanese Academy of Information Technology, Poland
- Taglino, Francesco, IASI-CNR, Italy
- Telek, Miklos, Budapest University of Technology and Economics, Hungary
- Tomczyk, Łukasz, Jagiellonian University, Poland
- Tomović, Savo, University of Montenegro, Montenegro
- Töreyin, Behçet Uğur, Istanbul Technical University, Turkey
- Trabert, Thomas, Helmholtz Centre for Environmental Research—UFZ, Germany
- Trocan, Maria, Institut Supérieur d'Électronique de Paris, France
- Tsihrintzis, George A., University of Piraeus, Greece
- Tudruj, Marek, Institute of Computer Science, Polish Academy of Sciences, Poland
- Udristoiu, Anca, University of Craiova, Romania
- Ullah, Ubaid, Deusto University Bilbao, Spain
- Ullah, Zaib, Università Telematica Giustino Fortunato, Italy
- Unland, Rainer, University of Duisburg-Essen, ICB, Germany
- Valle, Marcos Eduardo, Universidade Estadual de Campinas, Brazil
- Varanda Pereira, Maria João, Instituto Politécnico de Bragança, Portugal
- Vardanega, Tullio, University of Padua, Italy
- Vasiliadis, Giorgos, Hellenic Mediterranean University and FORTH-ICS, Greece
- Vega-Rodríguez, Miguel A., University of Extremadura, Spain
- Vescoukis, Vassilios, National Technical University of Athens, Greece
- Vogiatzis, Chrysafis, University of Illinois, United States
- Vo, Bich Khue, University of Finance—Marketing, Viet Nam
- Vuković, Željko, Faculty of Technical Sciences, Serbia
- Wahid, Khan Ferdous, Airbus Group, Germany
- Walters, Kamila, University of Brighton, School of Business and Law, United Kingdom
- Wątróbski, Jarosław, West Pomeranian University of Technology, Poland
- Wawer, Aleksander, Institute of Computer Science, Polish Academy of Sciences, Poland
- Wecel, Krzysztof, Poznan University of Economics and Business, Poland
- White, Martin, University of Sussex, United Kingdom
- Wiesmayr, Bianca, Johannes Kepler University Linz, Austria
- Wilcockson, Thomas, Loughborough University, United Kingdom
- Winnige, Stefan, Federal Institute for Vocational Education and Training (BIBB), Germany
- Wojszczyk, Rafał, Koszalin University of Technology, Poland
- Wrona, Konrad, NCIA, Poland
- Wyrzykowski, Roman, Czestochowa University of Technology, Poland
- Xenakis, Christos, University of Piraeus, Greece
- Yasir, Muhammad, IREA-CNR, Italy
- Zargayouna, Mahdi, Université Gustave Eiffel, France

- Zaza, Gianluca, University of Bari Aldo Moro, Italy
- Zborowski, Marek, University of Warsaw, Poland
- Zielinski, Zbigniew, Military University of Technology, Poland
- Ziemba, Paweł, University of Szczecin, Poland
- Zurutuza, Urko, Mondragon Unibertsitatea, Spain

ACKNOWLEDGMENTS

In conclusion, let us emphasize that delivery of FedCSIS 2025 required a dedicated effort of many people. We would like to express our warmest gratitude to all Topical Area Curators, Thematic Session organizers, members of the FedCSIS 2025 Senior Program Committee and members of the FedCSIS 2025 Program Committee (a total of more than 600 individuals), for their hard work in attracting and reviewing all submissions. We thank the authors of papers for their great contribution to the theory and practice of Computer Science and Intelligence Systems. We are grateful to Keynote and Invited Speakers for sharing their knowledge and experiences with the participants. Last, but not least, we acknowledge, one more time, Jarosław Wąs, Tomasz Hahaj, Łukasz Rauch, Anna Smyk, Anna Stolarczyk Piotrowska, Marian Bubak and Marek Grzegorowski, and their Team, Anastasiya Danilenka and Paweł Szmeja, as well as a fantastic group of student helpers. We are very grateful for your efforts!

We also hope to meet you again for the 21st Conference on Computer Science and Intelligence Systems (FedCSIS 2026) which will take place in Riga, Latvia, on August 23-26, 2026.

Co-Chairs of the FedCSIS Conference Series

Bolanowski, Marek, *Rzeszów University of Technology, Poland*

Ganzha, Maria, *Warsaw University of Technology, and Systems Research Institute Polish Academy of Sciences, Poland*

Maciaszek, Leszek, (Honorary Chair), *Macquarie University, Australia and Wrocław University of Economics, Poland*

Paprzycki, Marcin, *Systems Research Institute Polish Academy of Sciences, Poland*

Ślęzak, Dominik, *QED Software and University of Warsaw, Poland*

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 14–17, 2025. Kraków, Poland

TABLE OF CONTENTS

MAIN TRACK

MAIN TRACK INVITED CONTRIBUTIONS

Active Inference in the Distributed Computing Continuum **1**
Schahram Dustdar

Multiscale MoE: A Mixture of Experts Framework with Attention-Driven Multi-Scale Learning for Brain Tumor Classification **3**
Ameer Hamza, Robertas Damaševičius

SHARE: An Interactive Learning System for Improving Legal Consultation Services and Training **13**
Stephanie Tan, Wendy Aartsen, Dicky van Hamersveld, Maarten van der Sanden, Catholijn Jonker

New Education Challenges in Profiling Digital Experts for a Digital Economy Era **25**
Ivan Luković, Marija Đukić, Dragan Stanojević

Q-ID: A Reinforcement Learning Framework for Adaptive Intrusion Detection **35**
Maisha Maliha, Mohammed Atiquzzaman

Human-centered LLMs for Inclusive Language Technology: The Need to Embrace Variation Holistically in NLP **43**
Barbara Plank

Optimizing the Optimizer: An Example Showing the Power of LLM Code Generation **47**
Camilo Chacón Sartori, Christian Blum

Interactive Granular Computing: Toward Computing Model for Complex Intelligent Systems **59**
Andrzej Skowron, Andrzej Jankowski, Soma Dutta

MAIN TRACK REGULAR PAPERS

Adoption and Evaluation of Mobile Gaming Applications for Pain Management in Paediatric Oncology: A Culturally Contextualised TAM-Based Framework and Mixed-Methods Pilot Study in Saudi Arabia **73**
Samah Almaghrabi, Natalia Beloff, Martin White

Digital Transformation in Saudi Public Universities: A Novel Framework for Adoption Drivers and Impact Analysis **87**
Saleh Z Alshehri, Natalia Beloff, Martin White

CADM: An LSTM-Based Model for Detecting Creative Accounting in Time-Series Data from Saudi-Listed Companies **99**
Maysoon Bineid, Natalia Beloff, Anastasia Khanina, Martin White

Comparison of Large Language Models Supporting the Polish Language in Terms of Faithfulness in Retrieval-Augmented Generation Applications **111**
Marcin Blachnik, Jakub Chmielewski

A Framework for Model-Driven AI-Assisted Generation of IT Project Management Plan and Scope Documents **121**
Jānis Rihards Blazevičs, Oksana Nikiforova, Oscar Pastor

A New Perspective of Associative Memories for Partial Patterns **133**
Soma Dutta, Jaewon Shin

Hybrid mutation MTD solution with dedicated SDN agents **145**
Mateusz Fronczyk, Mariusz Rawski

Towards a Framework for Systematic API Migrations **155**
Nikolas Jíša, Robert Pergl

Toward Conversational Decision Support Systems: Integrating LLMs in the Operations Research Methodology **165**
Mariusz Kaleta

Slurm plugin for HPC operation with time-dependent cluster-wide power capping **175**
Alexander Kammeyer, Florian Burger, Daniel Lübbert, Katinka Wolter

Assigning scientific texts to existing ontologies **185**
Lukáš Korel, Martin Holeňa

Estimating the entropy of covering-based rough set approximation operators **195**
Wannes De Maeyer, Steven Van Overberghe, Chris Cornelis, Mauricio Restrepo López

Simultaneous pursuit of accountability for regulatory compliance, financial benefits, and societal impacts in artificial intelligence (AI) projects **207**
Gloria Miller

Deep Differentiable Logic Gate Networks Based on Fuzzy Łukasiewicz T-norm **219**
Chan Duong Nguy, Piotr Wasilewski

Segmentation and Process Assignment of Semi-Structured Event Logs **231**
Piotr Przymus, Krzysztof Rykaczewski, Janusz Zieliński, Łukasz Mikulski

Addressing Skills Shortages through Low-Code/No-Code: Skill Profiles and Implementation Challenges **243**
Paul Rozbitski, Christian Leyh, Ching-Ho Ip

Keypoint-based metric for evaluating image super-resolution quality **255**
Jakub Sadel, Tomasz Tarasiewicz, Pawel Kowaleczko, Maciej Ziaja, Daniel Kostrzewa, Pawel Benecki, Michal Kawulok

A Comparative Study of LSTM Efficiency vs. Transformer Power for Localized Time Series Forecasting **265**
Twisampati Sarkar, Henry Chu

Real-time Container Tracking and Damage Detection at Seaports Using Deep Learning **277**
Sotiris Vasileiadis, Sheraz Aslam, Kyriacos Orphanides, Alessandro Cassera, Eduardo Garro Crevillen, Alvaro Martinez-Romero, Michalis Michaelides, Herodotos Herodotou

MAIN TRACK SHORT PAPERS

A Lightweight Optimization Approach to the Single-Person Pose Estimation Pipeline in RGB-D Cameras **287**
Vytautas Abromavičius, Rytis Maskeliūnas

Simple Idea Discovery in a Minimalist LLM Architecture Implementation **291**
Robert Chihaia, Maria Trocan, Florin Leon

Data Interoperability Using Smart Data Models and NGSI-LD for the Norwegian Agrifood Sector **297**
Rustem Dautov, Simeon Tverdal, André Skoog Bondevik, Svein Arild Frøshaug, Vera Szabo, Jan Robert Fiksdal

Automatic Parcel Damage Recognition Module for an Inspection Robot **303**
Wiktor Goszczyński, Szymon Wałęga, Janusz Chmiel, Bartłomiej Gawęda, Katarzyna Grobler-Dębska

Length-Preserving Hair Simulation via Bézier Curves and Artificial Neural Network **309**
Kali Gürkahraman, Ahmet Firat Yelkuvan, Rukiye Karakis

StarCraft strategy learning refinement using replay snapshotting **315**
Štefan Krištofík, Michaela Hanková

Applying Evolutionary Techniques to Enhance Graph Convolutional Networks for Node Classification: Case Studies **321**
Maciej Krzywda, Szymon Lukasik, Amir H Gandomi

AI classifier of defects in Artworks captured by active infrared thermography **327**
Peter Malík, Martin Orlej, Branislav Fajčák, Massimo Rippa

Attention-Based Multi-Task Learning and PPO Reinforcement Learning for Explainable Blood Glucose Prediction **333**
Sarmad Maqsood, Egle Belousovienė, Rytis Maskeliūnas

Forest-Inspired Reinforcement Learning Based On Nature Ecosystem Feedback Mechanisms **339**
Rytis Maskeliūnas, Robertas Damaševičius

Out-Of-Distribution Is Not Magic: The Clash Between Rejection Rate and Model Success **345**
Itay Meiri, Ran Dubin, Amit Dvir, Chen Hajaj

Digital Twin Micro-Services Architecture to Support a Scenario-based Integration of Smart Mobility Services and Simulations **351**
Damien Nicolas, Christophe Feltus, Ruben Antonio De Jesus Marques

Enhancing API documentation by inter-endpoint dependency graphs **357**
Panagiotis Papadeas, Dimitrios Gerokonstantis, Christos Hadjichristofi, Vassilios Vescoukis

Hybrid U-Net segmentation of vessels in fundus eye images **363**
Lesław Pawlaczyk

Enhancing Low-Cost Air Quality Sensors with AI for Smart Green Routing **369**
Laura Po, Martina Casari, Federica Rollo, Matteo Angelinelli, Giorgio Pedrazzi, Chiara De Pascali, Luca Nunzio Francioso, Roberta Turra

Metaheuristics for Rolling Stock Cyclic Job Scheduling Problems with Maintenance **375**
Radoslaw Rudek

Symbolic vs Black-Box Explanations: A Model-Driven Approach Using Grammatical Evolution **381**
Dominik Sepioło, Antoni Ligęza

Comparative Analysis of AI and Traditional Software GitHub Repositories Using Process Mining **387**
Oguzhan Tasci, Tugba Gurgen Erdogan

Challenges in Evaluating OSS Quality: Results from SLR on Quality Evaluation Tools **393**
Aslı Taşgetiren, Ayça Kolukısa Tarhan

Anomaly Detection for Unmanned Surface Vehicles Based on a Multi-Modal Bayesian Generative Model **399**
Micheale Tekulu, Ali Krayani, Pamela Zontone, Lucio Marcenaro, Francesco Caprile, Antonino Masaracchia, Carlo Regazzoni

Unified Data-Driven Prediction of Photovoltaics Output from Weather and Geographic Data Across Diverse Systems **405**
Joanna Wójcicka, Tomasz Hachaj

Evaluating Effectiveness of Nonlinear Dimensionality Reduction in Hedge Funds' Returns Forecasting **411**
Milica Zukanović, Aleksa Radosavčević, Ana Poledica, Pavle Milošević, Ivan Luković

Encouraging Sustainable Fashion Choices Through a Mobile Application **417**
Bianka Šimková

THEMATIC SESSIONS

Preface to Thematic Sessions **425**

THEMATIC SESSIONS REGULAR PAPERS

Multi-level Optimization-based Ensemble Machine Learning for Efficient Crop Yield Prediction in Saudi Arabia **429**
Raya Aldawoud, Zohra Sbaï

Leveraging Large Language Models for Automated Export Control Screening: Evaluating LLMs Framework **441**
Salem Alotaibi, Alexei Lisitsa, Antony McCabe, Joanna MacSween

AraXLM: Evaluating Arabic Diacritization Tools for Cross-Language Plagiarism Detection **451**
Mona Alshehri, Natalia Beloff, Martin White

Hybrid Approaches for Pneumonia Detection in X-rays: Combining CNNs and ML Classifiers **461**
Gabriele S. Araújo, Omar Andres C. Cortes, Olaf Reinhold, Fabio M. F. Lobato

Treating OCR Output as a Language (TOOL) – Improving OCR Output with Seq2Seq Translation **471**
Thomas Asselborn, Magnus Bender, Ralf Möller, Sylvia Melzer

Blackcurrant Plantations Monitoring Using Drones **479**
Jānis Bičevskis, Reinis Odītis, Ivo Odītis, Zane Bičevska

A Serious Game for Teaching Entrepreneurship and Strategic Management **489**
Boyan Bontchev, Sia Tsolova, Dessislava Vassileva

ELEVATE-AI: Evaluation of Learning Environments Via Assessment Tools Enhanced by AI **499**
Linda Burchiellaro, Francesco Faenza, Claudia Canali

Bridging Global Language Models and Local Spatial Data: The JackDaw Approach to Context-Aware Agriculture and Rural Planning **507**
Karel Charvat, Stein Runar Bergheim, Raul Palma, Adam Aron Rynkiewicz, Matúš Botek, Alexander Kovalenko, Pavel Kordík, Antonín Kubíček, Markéta Kollerová, Šárka Horáková

While-guard Synthesis by Abstract Static Analysis and CHC Solving **519**
Aleksandar S. Dimovski

Machine-Readable by Design: Language Specifications as the Key to Integrating LLMs into Industrial Tools **531**
Alexander Fischer, Louis Burk, Ramin Tavakoli Kolagari, Uwe Wienkop

Forecasting consent in organ donation: early assessment of machine-learning techniques **543**
Arianna Freda, Davide Maestosi, Maurizio Naldi, Gaia Nicosia, Andrea Pacifici

The Power of Preemptions in Scheduling on Shareable Resources **553**
Omer Lapidot, Tami Tamir

From Agents to Copilots: A Systematic Review of Digital Assistant Technology Adoption in Proprietary Productivity Software **565**
Niklas Preiß, Markus Westner

Vehicle Routing under Complex Access-to-Energy Constraints **577**
Alain Quilliot, Hélène Toussaint

An Application of Tensors in the Stochastic Reaction Diffusion Master Equation **587**
Md Mustafijur Rahman, Roger B. Sidje

Leveraging Business-IT Collaboration: Effective Design and Challenges of the IT Business Partner Role **595**
Constanze Riedinger, Christopher Rentrop, Melanie Huber, Stephan Zimmermann

Optimizing Deep Learning for Cotton Leaf Disease Detection Using Meta-Heuristic Feature Selection Algorithms **605**
Naeem Ullah, Francisco Martínez-Álvarez, Ivanoe De Falco, Giovanna Sannino

Implementation of random number generator service with IoT device **615**
Rafał Wojszczyk, Aneta Hapka, Kacper Akdağ-Ochnik

The price of customer presence in Attended Home Delivery with Customer Availability Profiles **625**
Roberto Zanotti, Daniele Manerba, Renata Mansini

THEMATIC SESSIONS SHORT PAPERS

Adiabatic Quantum Computing for the Subset Sum Problem: Preliminary Studies **635**
Cesar Freitas Bernardes, Pedro Castellucci, Douglas Gonçalves, Eduardo Duzzioni, Antonio Mucherino

Cognitive-Aware Peer Assessment: Design Implications from a Classroom Deployment **641**
Naama Bouskila, Lihi Dery

Machine learning for survival analysis: a comparative study on intensive care unit (ICU) patient data and simulations **647**
Lukáš Boček, Lubomír Štěpánek

Coverage of OpenStreetMap data at firm level. Assessing the suitability for social science representative firm samples and found data firm analyses **653**
Christian Gerhards

A Multithreaded Java-Based Video Encoder for Multicore Systems **659**
Beata Bylina, Maciej Okoń

reVISION: A Polish Benchmark for Evaluating Vision-Language Models on Multimodal National Exam Data **665**
Michał Ciesiółka, Filip Graliński

Applications and Challenges of Artificial Intelligence in Educational Course Design and Delivery **675**
Daniel Dan, Anna Wróblewska, Bartosz Grabek, Michał Taczała, Minoru Nakayama

Towards the automated classification of German job titles according to KldB **681**
Ralf Dorau, Kristine Hein

Do LLMs dream of antique hermeneutics? Critical remarks on automated text interpretation **687**
Jens Dörpinghaus, Michael Tiemann

Knitwear Production Scheduling **693**
Stefka Fidanova, Jelenko Stanchov, Maria Ganzha

AI-MTD: Zero-Trust Artificial Intelligence Model Security Based on Moving Target Defense **699**
Daniel Gilkarov, Ran Dubin

Entangled by Design: A structured Overview of Management Challenges concerning AI Adoption in Organizations **705**
Tim Klos, Alessandra Zarcone, Christopher Rentrop, Constanze Riedinger, Niculin Prinz, Melanie Huber

An application of BURA solver to fractional super-diffusion problems **715**
Nikola Kosturski, Ivan Lirkov, Marcin Paprzycki

An algorithm for Direct Construction of all Pareto Optimal Biobjective Minimum Spanning Trees **721**
Lasko Laskov, Marin Marinov

Towards Optimal Train Routing Using Microscopic Simulation on Moving Block Controlled Networks **727**
Severin Lochschmidt, Stefan Engels, Robert Wille

A New Dimension of Acceptance? Introducing Perceived Public Value as Extension of UTAUT in the Smart City Context **733**
Alisa Lorenz

Towards Game Level Generation Through LLM and GAN **739**
Filip Martinović, Danijel Mlinarić, Juraj Dončević, Agneza Krajna, Ivica Botički

DivCrypt: A Structured Framework for Validating Cryptographic Implementations **747**
Artur Misztal

NPDP programming for RISC multi-core processors **753**
Marek Palkowski, Mateusz Grużewski

Bias in Classical Life Tables Under Censoring: A Comparative Study With Kaplan-Meier Estimation and Actuarial Estimation Using Real and Simulated Data **759**
Lubomír Seif, Ondřej Vít, Lubomír Štěpánek

An RL Agent to Find Minimum Energy in a Tensegrity Representing a Cell **765**
Mustafa Shah, Arsenio Cutolo, Muddasar Naeem, Muhammad Waris, Musarat Abbas

AI Software Security for Smart Environment in a Dynamically Changing Knowledge Management Strategy **771**
Anna Sołtysik-Piorunkiewicz, Małgorzata Pańkowska, Stanisław Stanek, Karsten Böhm, Mariusz Żytniewski

A New Optimization Method for evaluating Sobol' Sensitivity Indices **777**
Venelin Todorov, Velichka Traneva, Stoyan Tranev, Mihai Petrov, Slavi Georgiev, Fatima Sapundzhi

Illumination-targeting Data Augmentation for Monochrome Images **783**
Andrzej Śluzek, Piotr Stachura

Improved upper bounds on the shortest watchman route in simple polygons: Dependence on reflex vertices and triangulation strategies **789**
Lubomír Štěpánek

A statistical hypothesis test for primality based on random divisor sampling: Principles, properties, adaptive design, and algorithmic analysis **795**
Lubomír Štěpánek

COMPETITIONS

DATA MINIG COMPETITION

Estimating the Difficulty of Chess Puzzles by Combining Fine-Tuned Maia-2 with Hand-Crafted and Engine Features **801**
Sebastian Björkqvist

A Multi-Stage Framework for Chess Puzzle Difficulty Prediction **807**
Ling Cen, Jiahao Cen, Malin Song, Zhuliang Yu

Multi-Modal Deep Learning with Residual and Structure-Guided Refinement for Chess Puzzle Difficulty Prediction **813**
Junlin Chen, Cenru Liu, Yujie Gao

A Stacking-Based Ensemble Approach for Predicting Chess Puzzle Difficulty **819**
Alan Liang, Cenzhi Liu, Kai Wang, Ethan Liu

Hybrid Boosting and Multi-Modal Fusion for Chess Puzzle Difficulty Prediction **825**
Ming Liu, Junye Wang, Yinghan Hu, Xiaolin Yang, Defu Lin

Pretraining Transformers for Chess Puzzle Difficulty Prediction **831**
Szymon Miłosz

The bread emoji Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge **837**
Tyler Woodruff, Luke Imbing, Marco Cognetta

Multi-Source Feature Fusion and Neural Embedding for Predicting Chess Puzzle Difficulty **843**
Haitao Xiao, Daiyuan Yu, Xuegang Wen, Le Chen, Kun Fu

FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests **849**
Jan Zyśko, Michał Ślęzak, Dominik Ślęzak, Maciej Świechowski

Author Index **855**

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 1–2
DOI: 10.15439/2025F4345
ISSN 2300-5963 ACSIS, Vol. 43

Active Inference in the Distributed Computing Continuum

Schahram Dustdar,
TU Wien, Austria

***Abstract*—Distributed applications now span sensors, edge nodes, fog clusters, and hyperscale clouds. Meeting service-level objectives across this "Distributed Computing Continuum" persistently fails when management is reactive, centralized, and blind to uncertainty. I argue for *predictive equilibrium* as the control objective and for a concrete diagnostic: the Kullback–Leibler divergence between a system's expected and observed causal behavior under perturbations, each modeled with a Bayesian network. This perspective draws from predictive regulation in neuroscience and the fluctuation–dissipation view of equilibrium in physics, and it sets the stage for *antifragility*—systems that get better because they were stressed, not despite it.**

I. Introduction

THE practical difficulty of continuum computing is not merely scale or heterogeneity but *nonstationarity*: workloads, topologies, energy states, and network conditions drift faster than operators can retune thresholds. Threshold rules act only after damage is visible; they presume that the future resembles the past and that a central brain has enough signal to decide well for everyone. In real deployments, devices often act with partial information, and local decisions interact in nontrivial ways, which makes steady Service Level Objective (SLO) compliance brittle. A different organizing principle is needed—one that regulates for stability while learning in the face of novelty.

A helpful lens comes from biology. Bodies do not maintain temperature or glucose by passively waiting for deviations; they *predict* and prepare. **Homeostasis** is reactive stabilization; **allostasis** generalizes it to **predictive regulation**, adjusting internal states in anticipation of expected demands. The **Free Energy Principle** reframes this as minimizing prediction error so organisms preserve their form and functions. Transported to engineered systems, the analogue is clear: components should anticipate loads, not just recover from violations, and success should be judged by how well predictions align with reality when the world is nudged.

The second lens is physical. In many systems, the relationship between internal fluctuations and responses to small perturbations is codified by the **Fluctuation–Dissipation Theorem**; departures from that relationship mark distance from equilibrium. Recent studies show how FDT violations reveal nonequilibrium brain dynamics. For computing, the physics is an analogy rather than a derivation, but it is a useful one: if a deployment's internal model predicts how a nudge will propagate and the observed response matches, we are near equilibrium; if prediction and observation disagree, we are not.

Putting these lenses together yields a working definition: **predictive equilibrium** is the condition in which a continuum system's internal model accurately represents its behavior under perturbations, such that deviations remain small enough to meet goals. This is not a static point but an active property sustained through ongoing modeling, anticipation, and reorganization. It entails dynamic balance (local parts keeping global SLOs on track), continuous reconfiguration (structure in service of goals), and predictive consistency (forecasts match outcomes under realistic nudges). Equilibrium so defined is the platform upon which antifragility can emerge.

To operationalize the idea, represent the deployment as a **Bayesian network** whose nodes are salient metrics and whose directed edges encode probabilistic dependencies within and across tiers. Choose a small, actionable vocabulary—latency percentiles by tier, queue depths, utilization, link RTT/loss, battery state, SLO flags—so the model remains interpretable and fast to update. Then specify a **perturbation** with clear semantics (bandwidth throttle, dependency delay, CPU cap, synthetic burst). Apply it twice: once to the model to produce an **expected** network (updating the affected conditionals) and once to the live system or a faithful twin to produce an **observed** network. The **KL divergence** between these distributions is the "distance-from-equilibrium" signal. Small divergence indicates predictive alignment; a spike signals a breakdown that warrants model revision, policy change, or both.

Why privilege KL divergence over raw SLOs? A latency breach only states that a guarantee failed. KL divergence localizes *why* predictions broke: edges may rewire (emergent dependency), weights may shift (weaker coupling), or noise may increase (stochasticity rose), all of which are visible in the comparative structure and parameters of the two networks. Moreover, the same divergence that warns of impending SLO drift

doubles as a *learning* signal for improving the internal model—a property traditional thresholds lack. Conceptually, this aligns with active-inference accounts where action is chosen to balance pragmatic value and information gain; minimizing KL between expected and observed behavior is precisely minimizing that mismatch.

An architecture built around predictive equilibrium emphasizes **local modeling, safe perturbation, and hierarchical synthesis**. Edge devices maintain compact, incremental BNs over their local neighborhoods and expose an "expectation API" that answers *What should happen here if bandwidth drops by 20% for 30 seconds?* Fog nodes orchestrate small, isolated perturbation campaigns (canaries, synthetic traffic), fit the observed BN, and compute divergence, while also enacting reorganizations—operator placement, replication factors, routing. A cloud-level meta-controller aggregates divergence summaries and, crucially, modulates the exploration–exploitation balance. Here the physics analogy pays off again: use an **FDT-inspired signal** as a *modulator* on the active-inference drive, transiently biasing decisions toward information gain when prediction degrades, then annealing back as equilibrium returns.

Consider a city-scale video analytics deployment. A fog controller periodically reduces uplink bandwidth by fifteen percent for a handful of cameras. The expected BN predicts mild queue growth and a compensatory frame-rate dip, preserving tail latency. Instead, observations reveal that a co-located GPU tenant injects bursty contention, creating a new dependency from that tenant's utilization to the canary's latency. KL divergence jumps. The controller updates the model (adding the edge), raises isolation for GPU tenants in that cell, and adjusts placement to avoid co-scheduling bandwidth-sensitive pipelines with bursty neighbors. In the next probe, divergence falls and the new policy holds. The system did not merely "heal"; it learned a structural lesson and retained it—an instance of antifragility in miniature.

Antifragility is not a slogan but a design criterion: systems should improve *because* stress exposes mismatches. In Taleb's sense, antifragile entities gain from volatility. In our setting, the gain is a better causal model and a sharper policy encoded in the BN and the controller; perturbations become training data that increase predictive fidelity and coordination skill. Equilibrium supplies the safety rails—keeping the system within acceptable performance bands while it harvests information from controlled nudges. There are limits. This is a **conceptual** framework; practical feasibility and cost remain to be demonstrated at scale. Perturbations must be safe and ethically scoped. BN learning must be sparse, incremental, and bounded in overhead. Multi-objective trade-offs (latency vs. energy vs. privacy), partial observability, and compliance constraints complicate modeling and may require task-specific variables and priors. Yet these are engineering questions, not theoretical roadblocks. A sensible validation path is to start in a digital twin, calibrate divergence thresholds to the system's natural variability, and then graduate to canary slices in production.

In sum, predictive equilibrium reframes continuum operations as a dialogue between model and world. Neuroscience supplies the instinct—prepare rather than react—and physics offers a ruler for distance from equilibrium. With Bayesian networks to represent expectations and KL divergence to measure their failure modes, we obtain both an early-warning gauge and a learning gradient. That combination is the essence of antifragility in engineered systems: *stay stable enough to learn, and learn enough to become more stable the next time*. In our previous work [1-5] some of the aspects above are detailed.

References

[1] S. Dustdar, V. Casamayor Pujol, and P. K. Donta, "On Distributed Computing Continuum Systems," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4092–4105, Apr. 2023, doi: 10.1109/TKDE.2022.3142856.

[2] V. Casamayor Pujol, B. Sedlak, Y. Xu, P. K. Donta, and S. Dustdar, "DeepSLOs for the Computing Continuum," in *Proceedings of the 2024 Workshop on Advanced Tools, Programming Languages, and PLatforms for Implementing and Evaluating algorithms for Distributed systems*, in ApPLIED'24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 1–10. doi: 10.1145/3663338.3663681.

[3] B. Sedlak, V. C. Pujol, P. K. Donta, and S. Dustdar, "Equilibrium in the Computing Continuum through Active Inference," *Future Gener. Comput. Syst.*, May 2024, doi: 10.1016/j.future.2024.05.056.

[4] V. Casamayor Pujol, B. Sedlak, P. K. Donta, and S. Dustdar, "On Causality in Distributed Continuum Systems," *IEEE Internet Comput.*, vol. 28, no. 2, pp. 57–64, Mar. 2024, doi: 10.1109/MIC.2023.3344248.

[5] V. C. Pujol, B. Sedlak, T. Salvatori, K. Friston, and S. Dustdar, "Distributed Intelligence in the Computing Continuum with Active Inference," May 30, 2025, *arXiv*: arXiv:2505.24618. doi: 10.48550/arXiv.2505.24618.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 3–12
DOI: 10.15439/2025F0901
ISSN 2300-5963 ACSIS, Vol. 43

Multiscale MoE: A Mixture of Experts Framework with Attention-Driven Multi-Scale Learning for Brain Tumor Classification

Ameer Hamza, Robertas Damaševičius
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania Email: ameer.hamza@ktu.edu, robertas.damasevicius@ktu.lt

***Abstract*—The impact of brain tumors as a global health concern is due to their aggressive behavior, high mortality, and complexities in their diagnosis. While MRI remains the gold standard for identifying, monitoring, and detecting brain tumors, automated classification methods encounter many complications with respect to the diverse morphologies of tumors, similarities in their imaging features, and the potential variability in imaging conditions.CNNs can capture spatial hierarchies, but cannot generalize effectively and ViTs rely on the context to characterize the image modalities which means that, whilst they address some deficiencies of CNNs, they require extensive data and computational resources. To remedy some of the issues that each approach presents, we present multiscale MoE that leverages CNNs and attention-oriented modules. The proposed architecture uses multi-scale feature extraction, channel-spatial attention, and dynamic expert routing, which adequately collects tumor-specific features efficiently. We applied two different publicly available datasets, namely the Bangladesh Brain Cancer MRI and Figshare Brain Tumor dataset. For the Bangladesh dataset, the proposed model achieved overall accuracy of 96.92% and for FigShare dataset, the highest results achieved 96.42% accuracy. In contrast to state-of-the-art models, multiscale MoE achieved the highest testing accuracy 96.14%, and the lowest Brier score 0.0603. The proposed model has shown to have balanced classification results across the tumor classes and reduced the number of false predictions whilst maintaining efficient computational performance and thus has the potential to provide a valuable resource for clinical practice with respect to real-time applications.**

I. Introduction

BRAIN tumors, whether they are primary or metastatic, are a major global health challenge because of their high mortality rates and difficult clinical presentations [1]. Symptoms of brain tumors vary broadly sometimes dependent on location, size, or growth speed and can include debilitating headaches, seizures, neurological deficits, cognitive deficits, and problems with speech or vision [2]. Global health data suggests that this burden is increasing: in 2020, an estimated 308,000 new brain and CNS tumors were diagnosed around the world which led to approximately 251,000 deaths [3]. In the United States, the estimated number of new malignant brain tumors for 2023 to 2025 is expected to reach nearly 24,000 new cases per year with the projection of approximately 18,330 brain tumor related deaths in 2025. Despite advances in their treatment, the overall five-year survival rate is still below 35%, with even lower rates for more aggressive brain tumor types like glioblastoma at close to 7%. This clearly highlights the need for better methods for diagnosis and treatment [4].

Brain tumors are usually approached with a multi-modality approach involving surgical excision, chemotherapeutics, and radiotherapy, which is additionally accompanied with corticosteroids to control intracranial pressure, and anticonvulsant medications to control seizures [5]. Currently, MRI imaging is the gold standard in brain tumor imaging, used for identifying, localizing, and monitoring tumors because it is non-invasive and has greater spatial resolution than the other imaging modalities, and it gives the surgeon an idea of structural and functional information regarding tumor heterogeneity [6]. The advent of automated imaging approaches, driven by MRI precision, has prompted an interest in a collection of imaging variants to assist radiologists in classifying brain tumors based on multimodal MR imaging datasets and to provide images a radiologist would have a unique understanding of and the stages could potentially be determined as well [7].

AI and deep learning specifically, has arisen as a powerful remedy to this issue with unique advancements in assessing the vast amounts of medical images we produce each day [8]. CNNs are the current state-of-the-art for classifying brain tumors as they are able to learn spatial hierarchies and differences based on the spatial relationships between features in MRIs [9]. Recently, ViTs have emerged as a promising rival by modelling global dependencies and self-attention across patch data. Regardless, both approaches address limitations [10]. CNNs rely on intensive training with finite and homogeneous training sets and their performance is symptomatic of their training process, as their ability to classify imaged tumors are less generalizable across imaging conditions[11]. ViTs struggle as they require large amounts of training data and rigorous computational resources to offset their limitations. They also depend upon well-performing patch models over large regions of space and subject to varying spatial biases, which is more frequently defined differently in medical domains because of variability and size [12].

Another problem is the need to develop explainable AI models, which could provide the explanation and justification of its decision [13], [14].

In order to overcome these challenges and limitations, the Mixture of Experts (MoE) framework serves as a promising

path forward because it combines multiple sub-model, allowing each expert to focus on different features of the input such as shape of tumor, texture, and surrounding tissue context. In MoE architecture, each expert focuses on a different feature representation, while a gating mechanism selects the best expert for a given input and its properties. MoE can even manage inter-patient differences, irregular morphologies of tumors, and inconsistencies in medical imaging that often complicate single model approaches. By integrating the advantages of both CNNs and transformers in to a modular framework, the MoE framework can improve classification and generalizability and can improve interpretability of decision making processes. Thus, the Mixture of Experts approach not only addresses the various issues regarding current approaches to automated brain tumor classification but also opens up the possibility for more reliable, rigorously, sensitive and clinically useful AI-supported diagnostic tools. In this work, we proposed a multiscale MoE for the precise classification of brain tumors.

The contributions of this study are as follows:

- A novel Multiscale MoE architecture consist of five experts, multiscale feature extraction, channel-spatial attention, and dynamic expert routing mechanism has been proposed.
- Comprehensive evaluation has been conduction based on two benchmark MRI datasets based on calibration, and inference speed comapred to the SOTA methods.
- Expert utilization analysis revealing dataset-specific specialization patterns, improving interpretability of decision-making.

II. Related Works

In recent years, a number of researchers have proposed several methods for using MRI scans to detect brain cancers [15]. These methods include both conventional machine learning algorithms and deep learning models [16]. Here is a presentation of the pertinent research on brain tumor identification utilizing brain tumor datasets [17]. For example Muhammad et al.[18] presented a comprehensive examination of the various grades used to classify brain tumors. A detailed explanation was given of the procedures used to categorize brain tumors (BTCs), including preprocessing the tumor, determining deep learning features, and classification. They discussed the particular limitations and achievements of the existing deep learning techniques for Bitcoin. The importance of transfer learning for deep learning feature extraction was also covered.

Narmatha et al. [19] presented a methods for classification and segmentation using a fuzzy brain-storm optimization algorithm. With this approach, the target brain cluster's greatest priority is provided by the storm optimization. To find the best answer, the fuzzy procedure is iterated several times. An accuracy of 93.85% is given for the experimental procedure, which was conducted using the BRATS2018 dataset.

Sajad et al.[20] presented a CNN-based multimodal tumor categorization method. They first separated the tumor regions in the MRI data using CNN. They then performed an extensive data augmentation to train an efficient CNN model. Later, they improved the CNN model that was already trained using improved brain data. Tumors were classified using the final layer of the presented method, and it was shown that improved data yielded superior results on the selected datasets.

Mzoughi et al. [21] presented a method for making neuroradiology simple. This study's primary goal was to use volumetric 3D MRI to detect brain tumors. The authors classified the tumors using a Multiscale 3D CNN architecture to increase process efficiency. By using tiny kernels, the suggested approach can lessen the weight of both local and global information. Additionally, the data augmentation technique is used to improve model training. Ultimately, they used experimental findings to demonstrate the effects of data augmentation.

III. Proposed MultiScale MoE

Convolutional Neural Networks have been used for image classification tasks for more than a decade now, demonstrating great potential in various domains particularly in medical imaging. Despite their good performance, there were some issues in traditional CNNs which needed to be addressed such as their inability to extract global patterns, handle diverse features and focus on important regions. Although these issues were addressed individually through attention modules and multi-scale architectures in different studies but there is no comprehensive model that can tackle all the issues together. In this paper, we proposed a new model which can handle all these limitations single handedly by incorporating attention modules, multi scale feature extraction block and Multi scale Mixture of experts in a single architecture to ensure efficient and robust image classification for diverse feature maps.

A detailed diagram of proposed architecture is shown in Figure 1.

Stage 1 begins with initial feature extraction using a 7×7 convolution, batch normalization, ReLU activation, and max pooling, followed by the first MoE block consisting of five parallel expert networks and a dynamic router that assigns adaptive weights to each expert's output. The aggregated expert features are refined through a Channel Attention module, which applies both max-pooling and average-pooling operations to emphasize important channels, and a Spatial Attention module, which uses mean and max pooling across spatial dimensions to highlight critical spatial regions. An Auxiliary Classifier is connected at the end of this stage to provide intermediate supervision.

The proposed model starts with an initial feature extraction block which lays the foundation for deep feature extraction. This block accepts an input of size $(224 \times 224 \times 3)$ and passes it to a 7×7 convolutional layer with a stride of 2 to process low-level features such as textures and edges. It is followed by a batch normalization layer β to normalize the features for stabilized training and a ReLU activation function σ to introduce non-linearity in order to capture complex features. After that, a max pooling layer M_p with a kernel size of 3×3 and a stride of 2 is applied to reduce the spatial dimensions

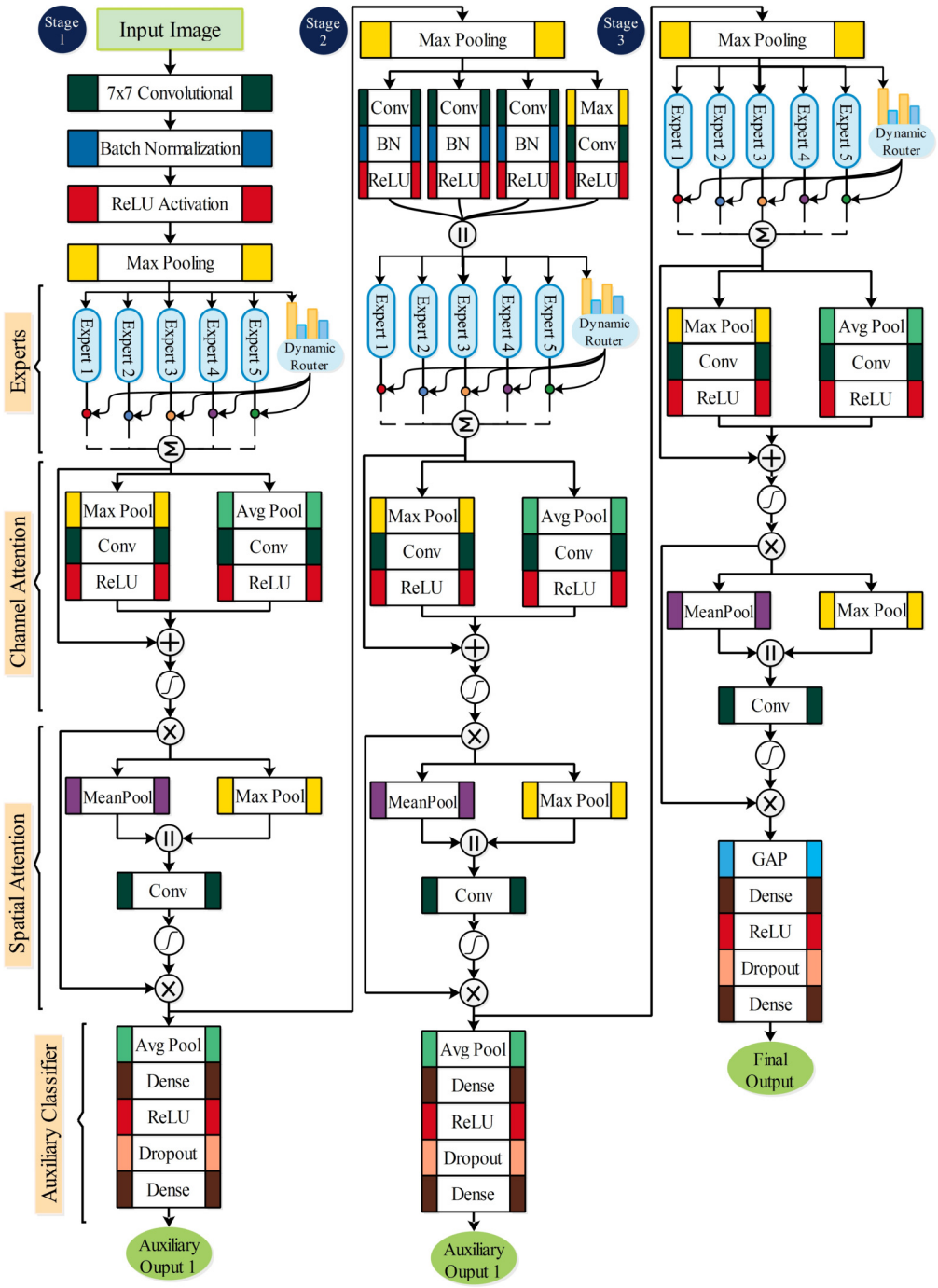


Fig. 1. Detailed architecture of the proposed Multiscale Mixture of Experts (MoE) framework for brain tumor classification. The model operates in three sequential stages, each comprising specialized processing modules.

from 224×224 to 56×56. However, after this block, the number of channels is increased from 3 to 64 for deep feature processing. Mathematically, this block can be represented as:

$$Z_1 = M_p\left(\sigma\left(\beta\left(W_1 * X\right) + b_1\right)\right) \tag{1}$$

Here, X represents the input tensor, $*$ represents the convolutional operation, W_1 represents the convolutional weights, and b_1 represents the bias term.

Stage 2 introduces a multiscale feature extraction block with parallel convolutional layers of varying kernel sizes and pooling operations to capture features at different spatial resolutions. The output is processed by a second MoE block with its dynamic router, followed again by channel and spatial

attention modules and a second auxiliary classifier.

Mixture of Experts: After initial feature extraction, the feature map is passed to the first block of the Mixture of Experts to extract complex features. This block is composed of 5 expert blocks and a dynamic router which assigns weights to each expert block. An expert block is a convolutional neural network that contains several convolutional, batch normalization, and activation layers to extract specialized features based on the input image.

In an expert block, the feature map is first passed through two 3×3 convolutional layers to extract features, where each convolutional layer is followed by a batch normalization layer to normalize the activations. A ReLU activation function is applied at the end to introduce non-linearity. All the expert blocks share the same architecture but have different weights that were initially assigned during training, making them specialized for different types of tumors. An expert block can be defined as:

$$E_i = \sigma\left(\beta\left(W^{''} * \beta\left(W^{'} * Z_1 + b^{'}\right) + b^{''}\right)\right) \tag{2}$$

Here, $W^{'}$ and $W^{''}$ represent the weights of both convolutional layers, and E_i represents the output of the i-th expert block.

On the other hand, a router is composed of an adaptive average pooling layer to summarize spatial information and two dense layers that produce a probability distribution over the 5 expert blocks. A ReLU activation is also applied between both dense layers. It can be defined as:

$$R_i = \psi\left(W_2 \cdot \sigma\left(W_1 \cdot A_p(Z_1)\right)\right) \tag{3}$$

Here, A_p represents adaptive average pooling, W_1 and W_2 represent the weights of the dense layers, and ψ represents the SoftMax activation function which converts the router weights into a probability distribution.

In the MoE block, the input is passed to each expert block and the dynamic router at the same time. The router analyzes the image and assigns weights to each expert block according to their specificity, ensuring that the most relevant experts receive the maximum weightage. The input is processed by each expert individually, and their outputs are multiplied with their respective weights and then added to generate the final feature representation. It can be defined as:

$$\hat{Y} = \sum_{i=1}^{5} R_i \cdot E_i(Z_1) \tag{4}$$

The block diagrams of the expert block and dynamic router are shown in Figure 2 and Figure 3, respectively.

Stage 3 incorporates deeper feature representations via a third MoE block, followed by attention modules and a Final Classifier consisting of global average pooling (GAP), fully connected layers, ReLU activation, dropout regularization, and a final dense layer for tumor type prediction. The architecture's modular design allows dynamic expert utilization, multi-scale feature capture, and attention-guided refinement, enabling robust and efficient classification across diverse MRI tumor appearances.

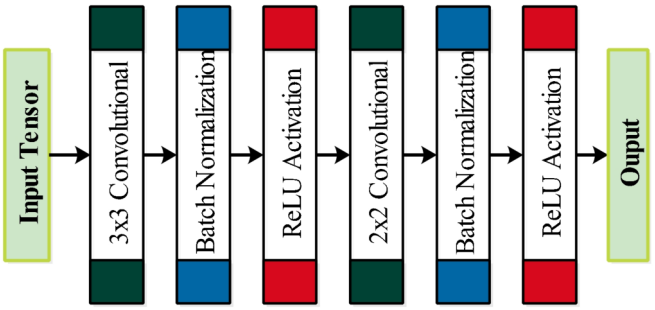


Fig. 2. Structure of an individual expert block used in the proposed Multiscale Mixture of Experts (MoE) framework. The block receives the input tensor and processes it through a sequence of layers: a 3×3 convolutional layer for local feature extraction, followed by batch normalization to stabilize training and ReLU activation to introduce non-linearity. A second convolutional layer with a 2×2 kernel is then applied for further feature refinement, again followed by batch normalization and ReLU activation. The output of this block provides specialized feature representations that contribute to the MoE's adaptive expert routing mechanism.

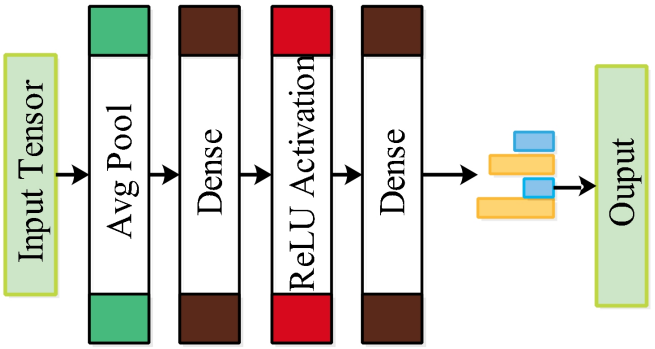


Fig. 3. Architecture of the dynamic router module in the proposed Multiscale Mixture of Experts (MoE) framework. The router receives the input tensor and applies an average pooling operation to capture global spatial information. The pooled features are passed through a fully connected (dense) layer, followed by a ReLU activation to introduce non-linearity, and then through a second dense layer to generate expert routing scores. These scores are normalized into a probability distribution, which determines the relative contribution of each expert in the MoE to the final feature aggregation.

Channel and Spatial Attention: The output of the Mixture of Experts block is passed to the channel attention module to calculate the channel-wise importance and emphasize the important features. For this purpose, the feature map is passed through two parallel blocks simultaneously. Each block is composed of a pooling layer, two convolutional layers, and a ReLU activation function. The only difference is that one block contains an average pooling layer while the other contains a max pooling layer to extract both local and global channel information. The outputs of both blocks are added, and a sigmoid activation function is applied to them to produce attention weights, which can be defined as:

$$A_c = \lambda\left[F_c\left(A_p\left(\hat{Y}\right)\right) + F_c\left(M_p\left(\hat{Y}\right)\right)\right] \tag{5}$$

Where

$$F_c = W_b * \sigma\left(W_a * Y + b_a\right) + b_b \tag{6}$$

Here, λ represents the sigmoid activation function, A_p and M_p represent average and max pooling, W_a and W_b represent the weight matrices of the convolutional layers, and b_a and b_b are their respective bias terms.

These attention weights are then multiplied with the original feature map to enhance the important features while suppressing the less important ones. Mathematically, it can be defined as:

$$Y_c = A_c \times \hat{Y} \tag{7}$$

These channel-wise refined features are then passed to the spatial attention module to emphasize spatially important locations. Here, the feature map is passed through average and max pooling layers simultaneously to capture both average and max features. The outputs of both layers are then concatenated and passed through a 7×7 convolutional layer and a sigmoid activation function to produce a spatial feature map, which is then used to weigh the spatial features according to their importance. Mathematically, it can be represented as:

$$A_s = \lambda\left(W_s * \left[A_p(Y_c) \parallel M_p(Y_c)\right] + b_s\right) \tag{8}$$

Here, $\parallel$ represents concatenation. This weighted feature map is then multiplied with the channel-wise recalibrated feature map to generate a channel and spatial-wise emphasized feature representation:

$$Y_s = A_s \times Y_c \tag{9}$$

Here, Y_c and Y_s represent the outputs of the channel and spatial attention modules, respectively.

Multi-scale Feature Learning: The recalibrated feature map is then passed to the multi-scale feature extraction block, which consists of four parallel layers, each with a different kernel size to extract information at various scales. Each branch is composed of a convolutional layer, batch normalization, and a ReLU activation function to extract and normalize features.

In Branch 1, the convolutional layer has a kernel size of 1×1 to capture pointwise features while reducing the dimensionality of the feature map. In Branch 2, the kernel size is increased to 3×3 to capture small and mid-sized tumors, while in Branch 3, a convolutional layer with kernel size 5×5 is used to extract contextual information.

The last branch, Branch 4, consists of a max pooling layer to summarize spatial information, followed by a 1×1 convolutional layer, batch normalization, and a ReLU activation to enhance spatial robustness in the model. The outputs of all the branches are concatenated before being passed to the next block. Mathematically, this operation can be represented as:

$$Y_M = \text{concat}[Y_1, Y_2, Y_3, Y_4] \tag{10}$$

Where the output of each branch is calculated as:

$$Y_i = \sigma\left(\beta\left(W_i * Y_s\right) + b_i\right), \quad \text{for } i \leq 3 \tag{11}$$

$$Y_i = Y_i(M_p), \quad \text{for } i = 4 \tag{12}$$

Here, W_i and b_i denote the weights and bias of the i-th convolutional layer, β is the batch normalization function, σ is the ReLU activation, M_p is max pooling, and Y_s is the input from the spatial attention module.

These multi-scale extracted features are then passed through another Mixture of Expert block with same architecture as before, however the depth of network is increased to 256 channels. This block ensures that features are extracted at various level of abstractions and model can learn to handle diverse range and types of tumors. Another channel and spatial attention module is incorporated after this Moe block to remove the noise from feature map and to pay attention on more tumor relevant features. It is then followed by another set of Moe block and attention modules incorporated in the very same architecture but with higher depth to extract progressively complex features. So, the input tensor after passing through three Moe blocks followed by attention modules, finally headed towards the main classifier for final classification.

Final and Auxiliary Classifier:

In the main classifier, the feature map is first passed through an adaptive average pooling layer to summarize the spatial information, and the output is passed to a dense layer that projects this information from a higher-dimensional feature space to a lower-dimensional feature space. A ReLU activation function is applied afterward to introduce non-linearity, followed by a dropout layer with a dropout rate of 0.5, which randomly deactivates 50% of the neurons to prevent the model from overfitting. Finally, another dense layer is used to convert the feature maps into class logits. Mathematically, this can be defined as:

$$C = W_2^{'} \cdot \partial\left(\sigma\left(W_1^{'} \cdot A_p(Y)\right)\right) \tag{13}$$

Here, ∂ represents the dropout layer, and $W_1^{'}$ and $W_2^{'}$ represent the weights of the dense layers.

Apart from the main classifier, two auxiliary classifiers are also introduced in the model—one after the first spatial attention module and another after the second spatial attention module. These auxiliary classifiers predict class labels from intermediate features, helping the model improve feature learning during the early stages of training.

Each auxiliary classifier consists of an adaptive average pooling layer followed by a dense layer, a ReLU activation, a dropout layer, and another dense layer—configured in the same way as the main classifier. The auxiliary losses are added to the model's overall loss only during training. During model evaluation (testing), only the loss from the main classifier is considered.

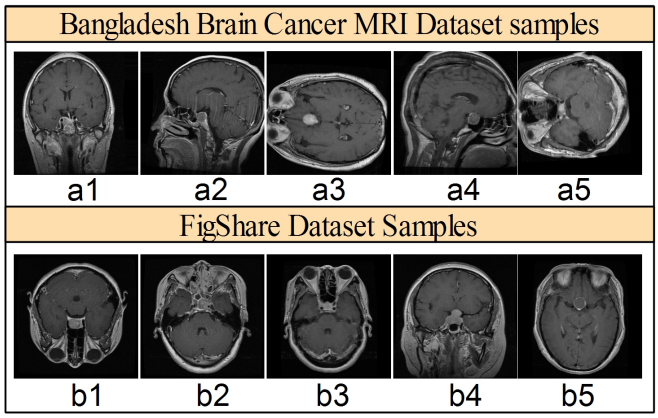


Fig. 4. a1-a5 presented the samples of Bangladesh brain cancer MRI and b1-a5 presented the sample images of Figshare dataset.

IV. Dataset Collection and Augmentation

The experimental design of this work uses two publicly available datasets of brain tumor image classification: https://data.mendeley.com/datasets/mk56jw9rns/1 and https://figshare.com/articles/dataset/brain_tumor_dataset/1512427. In both the Brain Cancer raw MRI data and the Figshare brain tumor dataset, which are displayed in Figure 4, there are three classes in the image classification data for brain tumors. The first dataset contains *Brain_Glioma*, *Brain_menin*, and *Brain Tumor*, while the second dataset includes *Meningioma*, *Glioma*, and *Pituitary Tumor*.

Consequently, Figshare dataset has limited number of images, which are insufficient to train deep learning models effectively. To solve this problem, the original images are splitted into 55% for training, 15% for validation, and 30% for testing. After that,we employed the data augmentation technique on training portion. In this method, three image translations are performed such as flip left, flip right, and rotation to increase the diversity of the dataset. The whole description of the datasets are described in Table I.

TABLE I
Brain tumor datasets with number of classes, images, and splits

No.of Class	Original Images	Train:Test: Validation
Bangladesh Brain Cancer - MRI dataset		
Brain_Glioma	2004	1102 / 602 / 300
Brain_Menin	2004	1102 / 602 / 300
Brain_tumor	2048	11026 / 615 / 307
Figshare brain Tumor dataset		
Glioma	1426	784 / 429 / 213
Meningioma	708	784 / 213 / 106
Pituitary tumor	930	784 / 280 / 139

V. Experimental Setup and Evaluation

A. Training and Parameters

For the training of proposed Mutliscale MoE model, 55% of data is used for training, 15% for validation during the learning phase, and the remaining 30% is allocated for testing. The training process is configured with a mini-batch size of 16, a learning rate of 0.00001, epochs is 50, number of experts is 5, auxiliary weights is 0.03, and the Adam optimizer. To prevent overfitting, early stopping is applied with a learning rate decay factor of 0.2, a patience value of 5 epochs, and a minimum learning rate threshold of 0.00001. All experiments are conducted using Python in a PyTorch environment, executed on a desktop system equipped with an NVIDIA RTX 3060 GPU 12 GB and 24 GB of system RAM.

B. Evaluation Metrics

To rigorously assess the performance of the proposed multiscale MoE model, we based on a comprehensive range of standard and complex evaluation metrics. The standard metrics are Precision, Recall, F1 score, and Support, which together provide a solid assessment of the models discriminative ability, sensitivity, and assessment of classification accuracy. In addition, we incorporated complex metrics such as Balanced Accuracy, F-beta score, Expected Calibration Error (ECE), Brier Score, and an Overall Testing. These metrics provide a fairly complete review of a models efficacy, including dealing with class imbalance, evaluating the tradeoff between precision versus recall when tailoring importance weights, and exploring the calibration of the predicted probabilities. Balanced Accuracy is important in providing equal performance for all classes, no matter how prevalent they are, while the F-beta score provides adjustable weighting so inferences can emphasize recall according to the requirements of any particular application. ECE and Brier Score are important measures of probabilistic reliability, measuring the covariance between predicted certainty and actual outcomes. Calibration-aware metrics are very important in clinical decision making, especially when it comes to high-stakes environments, when both predictive certainty and balanced performance for all classes are important to achieve accurate diagnosis, optimal treatment planning, and ultimately improve overall patient care. the complex metrices are mathematically defined as:

$$\text{BA} = \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{14}$$

$$F_\beta = (1+\beta^2)\cdot\frac{\text{Precision}\cdot\text{Recall}}{(\beta^2\cdot\text{Precision})+\text{Recall}} \tag{15}$$

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{16}$$

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\,|\text{acc}(B_m) - \text{conf}(B_m)| \tag{17}$$

Where C is the number of classes, β calculates the weighting between precision and recall, and N is the number of samples.

VI. Results of Proposed Multiscale MoE

The classification performance of the proposed Multiscale MoE model on the Figshare dataset shows excellent discriminative performance for all three brain tumors, a shown in Table II. Class-wise results indicate that pituitary tumor is classified with the highest precision which is 0.99, demonstrating that while the model correctly classified almost every tumor type, it had an excellent ability to correctly avoid false positives for pituitary tumor while maintaining a high sensitivity to pituitary tumor classification. Glioma classification similarly showed robust performance across both precision and recall with a score of over 0.97 and culminating in an F1-score of 0.98., Meningioma classification while still showing strength precision is0.92, recall is 0.93, and F1-score is 0.93 showed a greater degree of incorrect classification then the other tumors. Examination of the confusion matrix in Figure 5 shows that of the 76 meningioma cases, 12 were found to be glioma and 3 were classified as pituitary tumors. Moreover, of the 59 gliomas, 8 were misclassified as meningiomas. These errors are likely a result of the similarity of radiological appearance for these tumor subtypes, particularly when tumors have overlapping anatomy, as well as more subtle similarities in texture and intensity data that could challenge the MoE model's ability to create distinct class separation boundaries from the MRI scans of these tumor types.

Meanwhile, compared to the Bangladesh dataset in Table II, the proposed model provides greater consistency and balance in performance for each tumor class overall, with each tumor class exhibiting precision, recall and F1-score values exceeding 0.95. Glioma detection remains highly superior precision is 0.98, recall is 0.98, and F1-score is 0.98, meningioma detection achieves balanced precision which is 0.96 and recall is 0.95, pituitary tumor showed exceptional recall is 0.99 with precision of 0.96. In Figure 6, the confusion matrix shows few false predicted classifications with meningioma 10 samples predicted as glioma and 22 samples predicted as pituitary tumor. In addition, a few pituitary tumor such as 9 cases were predicted as meningioma. Generally, the confusion matrix shows fewer cross-class prediction errors when compared to the Figshare results which could be attributed to differing greater inter-class separability. Nevertheless, as can be concluded from both results, the few predicted cases can be easily explained by the overlapping tumor morphology, similar distributions of partial volume effects related to the MRI acquisition.

The extensive performance evaluation of the proposed Multiscale MoE, on the Figshare and Bangladesh datasets, supports the models apparent classification performance, robustness and likely generalization from several sources of data (Figshare and Bangladesh). For the Balance Accuracy measure which is useful to evaluate overall performance across all classes independent of sample distribution was shown to achieve 0.958 for Figshare and 0.9691 in Bangladesh. Both values highlights that the model consistently provides high recognition accuracy independent, or with major class imbalances. Similarly, the F-beta scores that coincidentally achieved the same values (0.9591 for Figshare and 0.9691 for Bangladesh) supports the model's performance aspect of maintaining a fair balance between positive predictive value

TABLE II
Classification results of proposed multiscale MoE on both selected datasets

Class	Precision	Recall	F1-score	Support
Figshare Results				
Glioma	0.97	0.98	0.98	429
Meningioma	0.92	0.93	0.93	213
Pituitary Tumor	0.99	0.96	0.98	280
Bangladesh Results				
Glioma	0.98	0.98	0.98	602
Meningioma	0.96	0.95	0.96	602
Pituitary Tumor	0.96	0.99	0.97	615

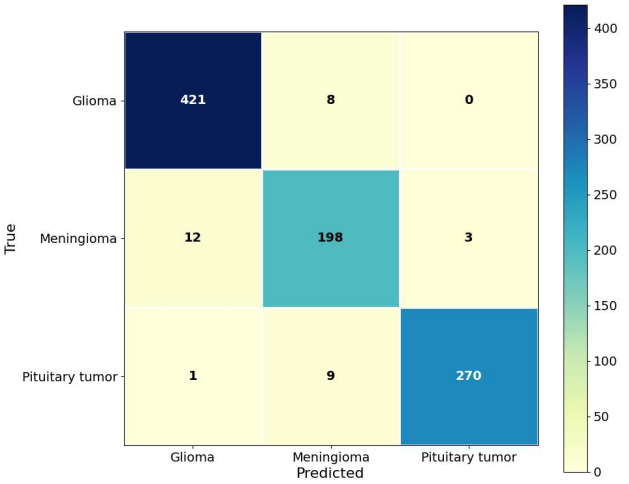


Fig. 5. Confusion matrix of proposed multiscale MoE on Bangladesh brain cancer dataset.

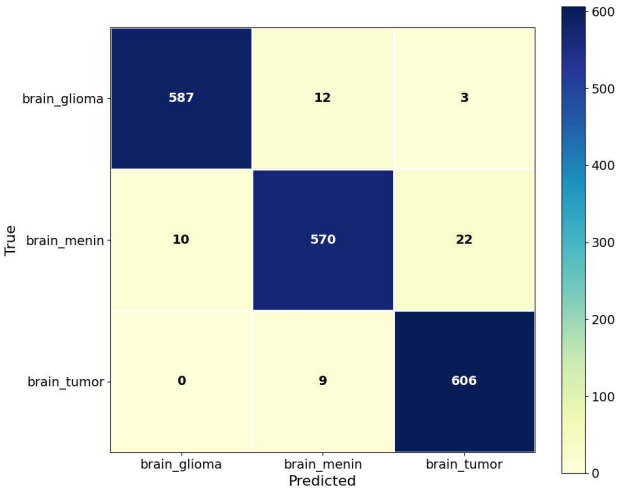


Fig. 6. Confusion matrix of proposed multiscale MoE on Figshare dataset.

(PPV) and sensitivity on both datasets. With the clinically important aim in not creating potential false negatives in brain tumours, we believe the model's F-beta performance being high shows that the model was successfully able to prioritize sensitivity for precision at some gotten trade-off.

Not surprisingly on reliability predictors, both the Expected Calibration Error (ECE) were quite low at 0.0229 for Figshare (and quite low at 0.0086 for Bangladesh) suggestively showing that close the predicted probability provided by model is to the actual observed outcome (the Bangladesh suggesting near perfect calibration). The Brier Score which is a metric of the mean squared difference between predicted probabilities (prediction) and actual outcomes (ground truth), were again shown to be lower in the Bangladesh dataset (0.0456) than Figshare dataset (0.0617), suggesting that model was making more confident and accurate probability assessments. The Overall Testing accuracy achieved over 96% for both datasets (96.42% for Figshare dataset and 96.92% for Bangladesh dataset) and demonstrated good fit and discriminative ability of the model supports its adaptable nature across datasets.

In sum, the performance results all imply that the proposed MoE model is able to provide excellent accuracy in classification, together with probability calibration which is very important in clinical environments where diagnostic accuracy remains paramount while confidence in predicted probabilities play a large role to reliable decision making and effective planning of the treatment decision.

TABLE III
COMPLEX METRICES PERFORMANCE OF PROPOSED MULTISCALE MOE ON SELECTED DATASETS

Metrics	Figshare Results	Bangladesh Results
Balanced Accuracy	0.958	0.9691
F-beta Score	0.9591	0.9691
ECE	0.0229	0.0086
Brier Score	0.0617	0.0456
Overall Testing	96.42	96.92

VII. ABLATION STUDIES

A. Model Computational Complexity

In the first experiment, an ablation study has been conducted aomng the proposed multiscale MoE model against state-of-the-art architectures such as Swin-T, SMViTv2-H and V-MoE based on parameters, GFLOPs and inference time, as shown in Table IV. The comparative results show that although SMViTv2-H [22] has the largest number of parameters 667M and GFLOPs 120.6, it has the highest inference time which is 24.50 seconds. Next, V-MoE [23] has the largest number of parameters which is 14.7B but a moderate GFLOPs value 12.35 and a reasonable inference time which is 11.71 sec, showcasing the efficiency of the expert-based architecture. Given the size of the model, while [24] had the least number of parameters which is 28M and GFLOPs 4.5, its inference time of 12.08 seconds may indicate a smaller representational capacity based on the reduced computational load. Our proposed multiscale MoE model returns a decent trade-off, remaining a lightweight model 25.27 million parameters with competitive computation complexity which is 11.95 GFLOPs. Furthermore, the proposed model exhibits the fastest inference time which is 10.85 sec of the entire study which we believe indicates that the proposed model is able to use multiscale feature processing and expert routing effectively, enabling the delivery of insights quickly and efficiently in a real-time enviroment.

TABLE IV
COMPARISON OF SOTA AND PROPOSED MODEL BASED ON PARAMETERS, GFLOPS, AND INFERENCE TIME.

Models	No. of Parameters	GFLOPS	Inference Time (s)
Swin T [24]	28	4.5	12.08
SMViTv2-H [22]	667M	120.6	24.50
V-MoE [23]	14.7B	12.35	11.71
Proposed	25.27	11.95	10.85

B. Comparison with state of the Art Models

The second ablation study described compares proposed multiscale MoE with leading state-of-the-art models, as shown in Table V. In performance measure F-beta, the proposed multiscale MoE model had the highest score for F-beta which is 0.951 indicating that proposed model provided the most focused balance between precision and recall score, and was best suited for classification tasks where the cost of both false positives and false negatives are critical. The ECE score of proposed model 0.0223 was slightly worse than the SMViTv2-H [22] model, which is 0.018, indicating that while our model was simultaneously reliable, SMViTv2-H [22] had a very slightly better measure of confidence with the prediction probability. Importantly, For the Brier Score the MoE had totally lowest value calculated for the Brier Score which is 0.0603 indicating that our proposed model had the highest accuracy of probability prediction. Finally, in testing accuracy score, our proposed MoE model had the highest score of 96.14% compared to all SOTA methods. While each individual performance measure demonstrates that the proposed MoE outperformed some of the competing methods, the collective measures prove that our proposed model not only had predictive performance better than state-of-the-art models, but also more reliable prediction probability and estimates.

TABLE V
ABLATION STUDY BASED ON F-BETA, ECE, BRIER SCORE, AND ACCURACY AMONG THE PROPOSED AND SOTA METHODS

Methods	F-beta	ECE	Brier Score	Accuracy (%)
Swin T [24]	0.925	0.030	0.075	94.80
SMViTv2-H [22]	0.945	0.018	0.062	95.90
V-MoE [23]	0.935	0.025	0.067	95.40
MViTv2-B [25]	0.940	0.020	0.065	95.60
Proposed MoE	0.951	0.0223	0.0603	96.14

C. Expert Utilization Across Selected Datasets

In this experiment, the utilization of each experts of proposed mutilscale MoE has been measured across the bangladesh brain cancer dataset, as shown in Figure 7. this

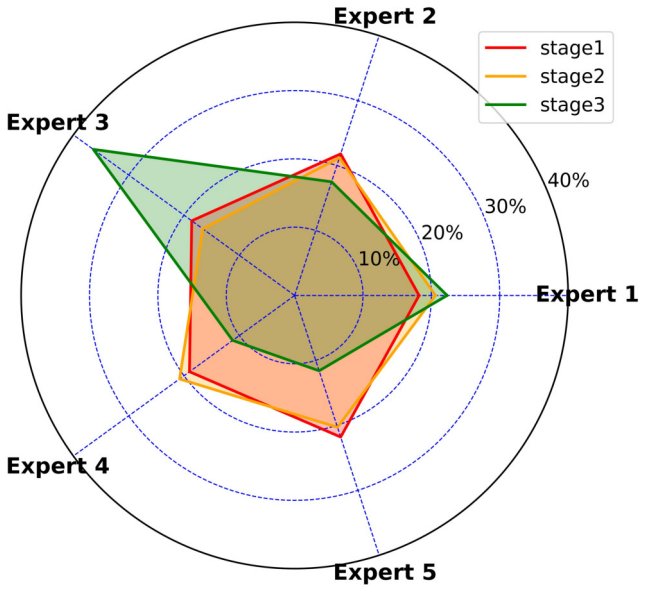


Fig. 7. Expert utilization patterns across the three stages of the proposed Multiscale Mixture of Experts (MoE) model on the Bangladesh Brain Cancer MRI dataset. The radar plot shows the percentage of routing weight assigned to each expert (Expert 1–Expert 5) by the dynamic router during different processing stages. In stages 1 and 2 (red and orange lines), the routing distribution is relatively balanced across experts, with moderate emphasis on Experts 1, 2, 4, and 5. In stage 3 (green line), Expert 3 dominates with nearly 40% routing weight, indicating its specialization in extracting high-level, fine-grained features critical for final classification, while other experts contribute in supporting roles.

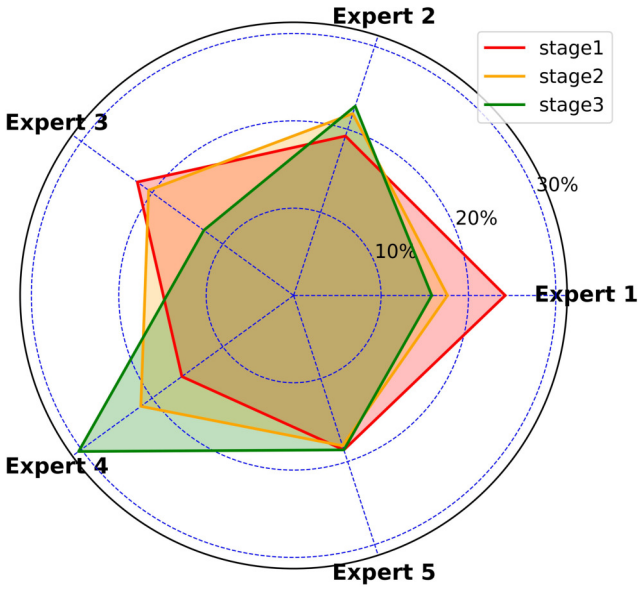


Fig. 8. Expert utilization patterns across the three stages of the proposed Multiscale Mixture of Experts (MoE) model on the Figshare Brain Tumor dataset. The radar plot displays the percentage of routing weight assigned to each expert (Expert 1–Expert 5) by the dynamic router during stage 1 (red), stage 2 (orange), and stage 3 (green). While stages 1 and 2 show a relatively balanced distribution across experts with moderate emphasis on Experts 2, 3, and 4, stage 3 exhibits a strong shift toward Expert 4, which receives the highest routing weight (over 30%), indicating its specialization in extracting discriminative features for the Figshare dataset's unique imaging characteristics.

plot clearly illustrates a distinct usage pattern of the proposed multiscale MoE experts in three stages. During the early stages, Experts 1, 2, 4, and 5 have relatively balanced and moderate routing weights, demonstrating the use of a distributed feature extraction approach in the beginning stages. A major shift occurred in stage 3 when the routing weight of Expert 3 increased substantially, to nearly 40%. This routing weight reflects that Expert 3 was the most prominent expert contributing to the eventual prediction decision. The increased weight of Expert 3 suggests that it was especially proficient at producing the fine-grained, high-level features required to classify thesubtle cases within the Bangladesh dataset. The other experts clearly contributed during this stage but their routing weights had decreased considerably, suggesting they were only being used in a supportive role of general feature representation rather than specific discrimination. The reliance on Expert 3 at the final stage represents the most low-level decision making and should provide a perspective about the versatility of the proposed model where the routing mechanism initiated relevant extremely relevant experts depending on the specific feature characteristics of the dataset.

In Figshare dataset, the routing behavior of experts shows an opposing pattern in specialization. The most apparent observation in this context was the increase in Expert 4 usage in stage three, overtaking Expert 3 whose share dropped significantly, as shown in Figure 8. Expert 4 became the expert with the highest routing weight during stage 3, with over 30%. This suggests that Expert 4 has learned feature extraction that is more appropriate given the unique imaging features of the Figshare dataset, which could include variations in tumor boundaries, texture distributions, and contrasts specific to different imaging modalities. All other experts 1, 2 and 5 showed relatively stable routing weights across the staging, indicating that the same experts participated consistently and contributed relatively equally in the decision making process when comparing with the Bangladesh dataset. Expert 3 lost its influence by later stages; it is clear from the data that what it had learned were less informative features for high level refinement processes in Figshare brain tumor model. Overall, the Figshare results exhibit more evenly shared expert utilization, where it appears the model was able to shift the use of experts on-the-fly as required; dynamically shifting emphasis under the final stage to Expert 4 when classifying images based on this dataset's unique properties.

VIII. Conclusion

Brain tumors are a significant global health challenge because of their inherent aggressiveness, poor patient prognosis, and difficulty in diagnosis. Notably, MRI scanning is acknowledged as the gold-standard diagnostic modality and ongoing treatment tool for brain tumors; however, automated classification techniques often struggle with the highly variable

morphologies of brain tumors, overlap in features, and different conditions on images. In this work, the study proposes multiscale MoE, a Mixture of Experts framework that unifies a CNN and attention-based modules in a single model that support multiscale approaches. The framework builds into the architecture a way to efficiently capture tumor-specific features using multiscale feature extraction, channel-spatial attention, and dynamic expert routing mechanism.

This study used two public datasets: the Bangladesh Brain Cancer MRI database and Figshare Brain Tumor dataset. the presented framework achieved precision, recall, and F1-scores greater than 0.95 for all classes of tumors including a balanced accuracy of 0.9691, a F-beta score of 0.9691, an ECE of 0.0086, a Brier score of 0.0456, and a general overall accuracy of 96.92%. With the Figshare dataset, precision scores approached 0.99, recalls approached 0.98, the balanced accuracy score was 0.958, the F-beta score approached 0.9591, the ECE was 0.0229, the Brier score was 0.0617, and a general overall accuracy of 96.42%. Compared to other state of the art models and approaches, multiscale MoE produced the highest testing accuracy 96.14% and the lowest Brier score 0.0603, and was ultimately the fastest model demonstrating inference times of 10.85 s without tuning.

Despite the promising results, there are limitations with the proposed multiscale model. it is reliant on MRI data because it was only trained and validated with data from two publicly accessible datasets and may not sufficiently reflect the vast variability routinely encountered in imaging protocols, tumor subtypes, and demographics when treating patients globally. This may have implications on the model's generalizability when implemented and applied in various medical environments.

Future work will focus on more data from more than one institute and add additional imaging modalities such as with PET and CT so that the model may learn from more various characteristics in tumours. Domain adaptation and transfer learning methods can be used to improve robustness across scanners and imaging acquisition protocols.

REFERENCES

[1] S. Lapointe, A. Perry, and N. A. Butowski, "Primary brain tumours in adults," *The Lancet*, vol. 392, no. 10145, pp. 432–446, 2018.

[2] D. Sipos, B. L. Raposa, O. Freihat, M. Simon, N. Mekis, P. Cornacchione, and Á. Kovács, "Glioblastoma: clinical presentation, multidisciplinary management, and long-term outcomes," *Cancers*, vol. 17, no. 1, p. 146, 2025.

[3] J. Huang, H. Li, H. Yan, F.-X. Li, M. Tang, and D.-L. Lu, "The comparative burden of brain and central nervous system cancers from 1990 to 2019 between china and the united states and predicting the future burden," *Frontiers in Public Health*, vol. 10, Oct. 2022.

[4] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," *Ca*, vol. 75, no. 1, p. 10, 2025.

[5] A. Moiyadi, V. Singh, R. Tonse, and R. Jalali, "Central nervous system (cns) tumors," in *Tata Memorial Centre Textbook of Oncology*, pp. 379–404, Springer, 2024.

[6] M. Ijaz, I. Hasan, B. Aslam, Y. Yan, W. Zeng, J. Gu, J. Jin, Y. Zhang, S. Wang, L. Xing, *et al.*, "Diagnostics of brain tumor in the early stage: current status and future perspectives," *Biomaterials Science*, 2025.

[7] S. Gunasekaran, P. S. Mercy Bai, S. K. Mathivanan, H. Rajadurai, B. D. Shivahare, and M. A. Shah, "Automated brain tumor diagnostics: Empowering neuro-oncology with deep learning-based mri image analysis," *Plos one*, vol. 19, no. 8, p. e0306493, 2024.

[8] A. Nazir, A. Hussain, M. Singh, and A. Assad, "Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis," *Multimedia Tools and Applications*, vol. 84, no. 17, pp. 17677–17740, 2025.

[9] R. Disci, F. Gurcan, and A. Soylu, "Advanced brain tumor classification in mr images using transfer learning and pre-trained deep cnn models," *Cancers*, vol. 17, no. 1, p. 121, 2025.

[10] V. Hassija, B. Palanisamy, A. Chatterjee, A. Mandal, D. Chakraborty, A. Pandey, G. Chalapathi, and D. Kumar, "Transformers for vision: A survey on innovative methods for computer vision," *IEEE Access*, 2025.

[11] D. Hussain, M. A. Al-Masni, M. Aslam, A. Sadeghi-Niaraki, J. Hussain, Y. H. Gu, and R. A. Naqvi, "Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: Methods, applications and limitations," *Journal of X-Ray Science and Technology*, vol. 32, no. 4, pp. 857–911, 2024.

[12] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, p. 32, 2025.

[13] S. Tehsin, I. M. Nasir, R. Damaševičius, and R. Maskeliūnas, "Dasam: Disease and spatial attention module-based explainable model for brain tumor detection," *Big Data and Cognitive Computing*, vol. 8, p. 97, Aug. 2024.

[14] S. Tehsin, I. M. Nasir, and R. Damaševičius, "Gatransformer: A graph attention network-based transformer model to generate explainable attentions for brain tumor detection," *Algorithms*, vol. 18, p. 89, Feb. 2025.

[15] C. Srinivas, N. P. KS, M. Zakariah, Y. A. Alothaibi, K. Shaukat, B. Partibane, and H. Awal, "Deep transfer learning approaches in performance analysis of brain tumor classification using mri images," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 3264367, 2022.

[16] H. A. Khan, W. Jue, M. Mushtaq, and M. U. Mushtaq, "Brain tumor classification in mri image using convolutional neural network," *Mathematical Biosciences and Engineering*, 2021.

[17] A. Pashaei, H. Sajedi, and N. Jazayeri, "Brain tumor classification via convolutional neural network and extreme learning machines," in *2018 8th International conference on computer and knowledge engineering (ICCKE)*, pp. 314–319, IEEE, 2018.

[18] K. Muhammad, S. Khan, J. Del Ser, and V. H. C. De Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 507–522, 2020.

[19] C. Narmatha, S. M. Eljack, A. A. R. M. Tuka, S. Manimurugan, and M. Mustafa, "A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumor mri images," *Journal of ambient intelligence and humanized computing*, pp. 1–9, 2020.

[20] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, "Multi-grade brain tumor classification using deep cnn with extensive data augmentation," *Journal of computational science*, vol. 30, pp. 174–182, 2019.

[21] H. Mzoughi, I. Njeh, A. Wali, M. B. Slima, A. BenHamida, C. Mhiri, and K. B. Mahfoudhe, "Deep multi-scale 3d convolutional neural network (cnn) for mri gliomas brain tumor classification," *Journal of Digital Imaging*, vol. 33, no. 4, pp. 903–915, 2020.

[22] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

[23] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[25] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 13–24
DOI: 10.15439/2025F6250
ISSN 2300-5963 ACSIS, Vol. 43

SHARE: An Interactive Learning System for Improving Legal Consultation Services and Training

Stephanie Tan
0000-0002-5755-2426
Delft University of Technology
Email: s.tan-1@tudelft.nl

Wendy Aartsen
hands4grants
Email: w.aartsen@hands4grants.eu

Dicky van Hamersveld
hands4grants
Email: d.vanhamersveld@hands4grants.eu

Maarten van der Sanden
Ontwerpatelier LUS
Email: maarten@ontwerpatelierlus.nl

Catholijn M. Jonker
0000-0003-4780-7461
Delft University of Technology
Email: c.m.jonker@tudelft.nl

***Abstract*—Providing citizens with transparent, efficient, and consistent legal advice is a critical challenge to uphold governmental fairness in societies. However, this remains a tall order for already overburdened organizations. This paper proposes an interactive learning system that seeks to address these challenges through an organizational learning lens, in collaboration with the Legal Desk in the Netherlands. This system, called SHARE, serves the dual purpose of supporting legal consultations and enhancing the training of legal advisors. SHARE captures and stores strategic decision-making knowledge from experienced legal advisors and facilitates its transfer to less experienced advisors. We explain in this case study the system's approach, prototype design, and evaluation through exploratory user tests. Our findings suggest that this prototype has the potential to contribute towards improved knowledge management for conducting legal consultations within the Legal Desk. To conclude, we discuss the outlook and broader implications of this approach in knowledge-driven work environments.**

I. Introduction

Today's citizens expect government services to be more transparent, accessible, and responsive, and these expectations continue to rise. While there has been many efforts to improve service delivery, they often struggle to fully meet the public's growing demands. Despite advancements such as self-service online portals and more recent solutions like large language model-based chatbot government services [1], uptake remains limited. The success of digitization and automation often hinges on factors like capacity, location, country, cultural practices, and situational contexts. As a result, in-person and telephone channels still dominate most citizen interactions with their government [2]. This makes staff members, who can deliver professional, clear, consistent, and courteous services, crucial to citizen satisfaction. Such direct service places staff at the forefront of fostering a citizen-oriented approach, emphasizing the importance of their training and education. In this work, we present a case focused on enhancing legal knowledge exchange and training for legal advisors in the delivery of legal consultation services, in collaboration with the Legal Desk in the Netherlands.

Learning from experience is a cornerstone of legal education and training [3]. Drawing on past experiences helps inform future practice decisions, a form of learning that is essential in fields such as medicine, law, education, and business, where hands-on experience is critical for developing professional competencies. However, this form of competence requires on-the-job learning that takes years. At the Legal Desk, it takes new hires at least 1.5 years to become experienced advisors. This training time is essential but time-consuming for the Legal Desk that is overworked. Compared to 2022, 2023 saw a 24% increase in in-person counter visits (from 57,900 to 71,700) and 148% increase in consultation appointments (from 15,500 to 38,500) [4]. However, the number of legal advisors increased by 2 (from 256 to 258), and the number of more experienced legal advisors increased by 1 (from 31 to 32). The workload of the legal advisors has significantly increased with the rising demand and limited growth in personnel.

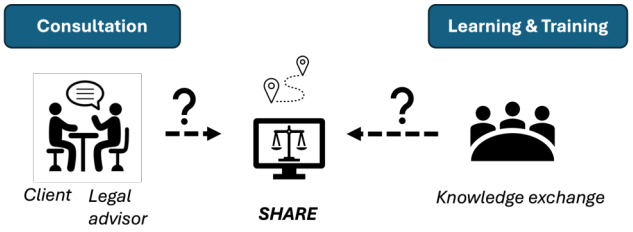


Fig. 1. SHARE aims to establish a knowledge connection between consultation and learning sessions within the existing workflow.

Our contribution is a novel interactive learning system called SHARE designed to improve legal consultation services and training of the legal advisors (see Figure 1). Our approach focuses on aligning structured and strategic knowledge elicitation, transfer, and management, while ensuring compatibility with existing organizational practices for ease of adoption.

We present this system as a solution to a highly complex and "wicked" problem [5], that is enhancing organizational learning to create more citizen-oriented legal services. Within the organization, there are many task and information interdependencies that are challenging to navigate and hence there is an intrinsic resistance for the solution. To this end,

we emphasize an open-ended, design-thinking approach from shaping the problem statement to co-creating solutions.

The SHARE system consists of physical tabletop components to support interaction and collaborative learning, next to a digital component (i.e., a web application) to capture and store structured legal information and the advisors' ways of working. This paper aims to explain a case study for the co-creation process, implementation, and the preliminary evaluation of this interactive learning system.

II. Related Work

A. Knowledge management system

Organizations rely on systematic approaches to retain knowledge. Despite of theoretical frameworks in organizational learning (e.g., single-loop, double-loop learning concepts [6], [7], organizational knowledge creation theory [8], and the five building blocks theory [9]), these approaches lack practical guidance and prescriptions and practitioners also match practical approaches to theory. A notable approach is Communities of Practice (CoP), which focuses on providing an informal learning environment where novices and experienced members of the community interact with each other [10], [11]. Through these interactions, CoP act as a forum for members to learn from each other and past experiences and to trigger joint systematic problem solving. CoP can also be formed by members outside of the organizational borders to help and enable learning within the organization (i.e., the scenario in this case study) [12].

Knowledge management is closely linked to organizational learning. It involves harvesting information and converting it into stored knowledge while ensuring that the knowledge is accurate and up-to-date. It may result in concrete and tangible outputs from the learning process, and facilitates in fortifying organizational memory. Organizational learning is a means to achieve continuous knowledge creation, retention, and utilization. Our physical and digital prototype solution which focuses on the knowledge elicitation, representation, retention, and the transfer process itself is situated at the intersection of complementary knowledge management and organization learning disciplines.

B. Legal learning and education

Legal education has traditionally relied on case-based methods, lectures, and moot court exercises to teach students legal reasoning and professional skills [13], [14]. A number of modern approaches focusing on experiential learning and gamification and interactive platforms have been proposed for legal education. Experiential learning leverages clinical programs (e.g., shadowing) and simulations, that aims to bridge the gap between theory and practice [15]. Gamification platforms also provide different scenarios for the legal learners to explore, providing a risk-free environment for practicing decision-making and experimentation [16].

Our proposed system shares some of the design goals as previous approaches (i.e., interactive and collaborative platform to access legal content [17]). However, our system is oriented towards legal advisors specifically via co-creation approaches so that we maximize likelihood of adoption of this new learning tool. We provide both low-tech (i.e., paper) and digitally innovative form-factors to create a more flexible learning framework.

C. Participatory design

Participatory approaches give substantive agency to stakeholders in product design and development. These approaches follow various traditions drawn from multiple disciplines, such as technology design, policy-making, and social sciences [18]. Belonging to the general approach in user-centered design, participatory design aims to incorporate a diverse range of perspectives, including for non-experts [19], [20]. It encourages continuous exploration and monitoring of the power dynamics between multiple stakeholders [21], [22]. Co-design is a type of participatory design that aims to foster user ownership by actively involving the stakeholders to contribute to the design solutions [23], [24]. The collaborations are more active and sustained, emphasizing on more equal partnerships and shared decision-making.

III. Approach

Our research methodology is principally based on practice-led research where the production, application, and evaluation of creative work leads the process of research [25]. Based on action research strategies that systematically generate and validate new information, procedures, and understandings thereof through practice [26], we emphasize on problem discovery and solving. In this section, we describe the objectives and the co-creation process leading to the design and implementation of the interactive learning system.

A. Conception and rationale

The main interdependent objectives of the SHARE interactive learning system are: (1) Ensuring consistency & quality in customized legal advice; (2) Improving knowledge retention & transfer within the Legal Desk; (3) Finding strategic and possible routes/solutions to cases; (4) Contributing to a citizen-oriented government.

Achieving consistency in customized legal advice (Objective 1) relies on effectively addressing the challenge of improving knowledge retention and transfer within the Legal Desk (Objective 2). Similarly, jointly exploring alternative strategic routes to appropriate advice in cases and problem-solving approaches (Objective 3) can only have a meaningful impact if Objectives 1 and 2 are also met. By addressing Objectives 1 through 3, the project makes a step towards fostering a citizen-oriented government (Objective 4). Citizens could be helped by an organization with streamlined learning and service delivery process. In this context, finding answers to our research questions (RQ) leads to meeting Objectives 1, 2 and 3. While ensuring and measuring consistency in custom legal advise depends on actively using the SHARE system, Objectives 2 and 3 form the foundation for developing the SHARE interactive learning system.

B. Co-creation process

The process started with observation sessions with three legal advisors at the counter and subsequently observing four consultation hours with clients. These exploratory observation sessions helped the core team to understand the workflow of the Legal Desk, the complexity of the cases, and expertise required to handle the cases.

Over a period of 12-months (May 2023-June 2024), the core team conducted a series of co-creations, design sessions, one-on-one appointments, and evaluation sessions with the Legal Desk advisors at the immigration law department. The first three sessions were scoping and brainstorming sessions. Over this period, we co-created various diagrams, mind maps, schematics, and sketches as a necessary endeavor to envision and materialize the SHARE prototype. Preliminary depictions were milestones of progress toward fulfilling the intended problem discovery and definition.

1) Shape Language: Along with post-its, pens, and papers for brainstorming, we also used the Shape Language [1] to support communication between the participants. Shape Language is designed as an intuitive communication instrument to address complex problems. It consists of 3D geometric shapes inspired by architecture (e.g., pyramids, spheres, cube, etc.). These geometric shapes are manipulated without instructions and free of predefined meaning to facilitate the abstraction of complex problems through a joint sense-making process. It has been shown that geometric shapes are related to emotions, metaphoric clues, and are associated with affective value that may assist in interaction and learning [27], [28] . The Shape Language was augmented with a 2D chip variant including triangles, circles, squares, bars/sticks, upon suggestions during the co-creations. They are incorporated in our proposed learning system to provide some tactile interactions that could support learning [29] (see Figure 2).



(a) (b)

Fig. 2. Shape Language in use during co-creation (a) and client test (b).

2) Transitioning into solution-building: The initial co-creation sessions were exploratory. Topics ranged from introducing the use of the Shape Language, identifying regularly occurring complex types of client situations (we here call these canonical situations) and design challenges. In line with the design thinking approach of continuous solution-finding [30], we had different goals per co-creation session, while maintaining an overarching approach to create a divergent thinking mindset, before eventually moving into solution design. The co-creation sessions were insightful to share knowledge and experiences, followed by the participants thinking *together* about feasible solutions to perceived obstacles and organically arriving at the desire to implement those solutions.

In December 2024, the team converged on the concept of using *canvases* as a type of laying field for case problem-solving. The idea is that these canvases physically represents the core topics, questions and answers essential for strategic information gathering on canonical situations and are designed to support conversations with clients. As canvases capture the diagnostic strategic knowledge of experienced legal advisors, jointly preparing the canvas content results in knowledge exchange with less experienced colleagues.

3) Elucidating strategic decision-making: We further define what the structure of strategic decision-making is in the context of using the canvases. The procedure starts with a so-called working thought and ends with an associated advice. The process resembles hypothetical reasoning. Based on an initial assessment of the situation and the desire or expectation to be able to provide specific advice to the client, the advisor forms a working thought (hypothesis). This working thought will be tested by asking targeted relevant questions searching for the key information. The key information they look for is two-fold: essential information that should always be satisfied for giving that advice, and also checking contradictions, i.e., information that invalidates the working thought and advice strategy. Depending on the answers and the evidence obtained, the associated advice is tailored to the client situation, or the working thought is reformulated and the process repeated until an advice is given. This procedure helps advisors to quickly capture the essential information about a case by choosing a route across the canvas(es). As the canvas contains the strategic knowledge of the more experienced advisors, it unlocks this knowledge for the less experienced advisors by following the same working thought and route for problem-solving.

4) Canvas co-design: Being able to formulate the above-mentioned working thoughts and questions to seek key information is essential for the success of the canvas approach. As a first approach we worked with a collection of questionnaires via Google forms to identify the key information and the most relevant questions asked by the advisors to obtain the essential answers from the client, drawing on their experience gathered from canonical case scenarios [2]. We asked the advisors to create a list of topics or the so-called knowledge card. For each knowledge card, we asked for the underlying reason why the knowledge card belongs to a canvas (with the assumption that there are some topics that are common between the canonical cases). We then asked the advisors to formulate several example questions for each knowledge card that the advisor

[1]See VormTaal: https://www.vormtaal.com/

[2]The elicited working thoughts and questions are not client-specific but based on canonical case scenarios. No personally identifiable information were included.

should ask or consider to obtain the key information. This process was intuitive, as the experienced advisor, after some initial questions, completed the Google form questionnaires for some of the earlier identified canonical situations, i.e., a specific *legal framework* (i.e., the combination of a legal topic $\{\text{family migration}, \text{integration}, \text{end of relation}\}$ with a geographical region pertaining to the legal case in question $\{\text{NL}, \text{EU}, \text{3rd countries}\}$).

In the following co-creation design session, the advisors placed the collected knowledge cards per canonical situation on a canvas. They exchanged knowledge and together arrived at the order of the questions and knowledge on the canvas or adjusted where necessary. Here, working thoughts resemble the concept of hypothesis based on the given facts from the situation and serves as the starting point of further investigation using the questions on the knowledge cards. The participants decided to create a so-called "situation canvas" to support the collection of information that is key to any case in legal field of immigration law. With this approach, the ideation of the prototype, i.e., the use of canvases to capture strategic decision-making to facilitate advice-giving during consultation, is finalized. While the physical versions of the canvases (i.e., paper versions) were straightforward and intuitive to implement given the defined workflow, we also agreed on a scalable *and* flexible implementation. This motivated a web application for building and interacting with the canvases to easily store and access canvas-derived information.

IV. The SHARE Prototype

The prototype consists of a *physical canvas* (in paper form) and a *digital canvas* (in web application form). The deliberate choice for having both is that each enhances different modes of communication. We chose to design both forms of the canvas, which showed to be of added value during the client test sessions (see Sec. VI-E) as advisors and clients have different preference. The differing properties of the physical and digital versions of the canvas system are related to explainability and accessibility of information and discussed in this section.

A. Situation canvas

The situation canvas provides the first impression of the key case details for physical paper and digital web version, respectively (see Fig. 3). In the situation sketch box, the Legal Desk advisors can use the 2D Shape Language chips (see Sec. III-B1) to visually represent the case situation. The situation information depicts the people involved and their relation to the client. For the physical version, these chips are available in different colors (red, yellow, green). In contrast, the web version supports "drag and drop" operation of the shapes, which we will extend in future work to include different colors and more flexible drawing options. After sketching the situation, the system provides a notes block to record additional key information on all the people involved in the client's case, their age, nationality, residence history, and dependency status. The web application version mirrors the paper version. Leveraging the situation sketch and the basic information gathered about the legal case, the advisors make an informed decision on the legal framework that is most relevant to the case (e.g., integration for 3rd country individuals).

B. Knowledge cards and conversation canvas

Each knowledge card includes a collection of questions related to a topic relevant to the chosen legal framework. Answer types ($\{\text{free text}, \text{single choice}, \text{multiple choices}\}$) are coupled to each question. Each knowledge card has two explanation fields for: *(i)* a generic explanation/information of how and why this knowledge card is important to *all* legal frameworks, and *(ii)* a canvas-specific explanation that contains information of why this knowledge card is important to the *particular* canvas. For example, suppose the users decide to create a knowledge card for "nationality". Explanation type *(i)* contains generic information such as a list of countries in EU, EEA, and Schengen area, which are facts regardless of the legal framework. Explanation type *(ii)* may contain information about Suriname, a former colony of the Netherlands. This information would be relevant to "Integration", compared to other legal frameworks, as the integration requirements could be different for individual of Surinamese nationality. Fig. 4 shows the paper and web version of the knowledge cards on a conversation canvas.

While the physical paper version and the digital web application are designed to present the same information, there are few key differences that emphasize the strengths of paper and digital versions respectively. First, the aforementioned explanation fields are available in the digital version as we store the information in a database. This is impractical to replicate for the paper canvas version because of the number of documents and the amount of both structured and unstructured information associated with each topic. Secondly, the paper canvas has a limited footprint (i.e., only a handful of questions would fit on paper). This is a deliberate design choice that encourages the advisors to choose the *most* relevant content on topical- and answer-level. The paper canvas limits the possibility of getting sidetracked by case-specific details, which would be of particular added value when dealing with complex scenarios while finding a strategic solution. The web application, however, is designed to be flexible and store unstructured information into structured templates. To strike a balance between generality and specificity, the web application enables the users to select which knowledge cards, questions, and answers should be shown on a specific canvas. The principle of putting information that is necessary and sufficient to cover the canonical scenarios when building canvases was explicitly communicated to the users.

C. Advice cards

Advice cards are unique features of the web application (see Fig. 5). Each advice card contains the advice name, supporting information to the advice, an explanation of the advice at Common European Framework Level - B1 Dutch language level (i.e., at a common every-day Dutch level rather than using legal jargon), and an action list pertaining to the advice.

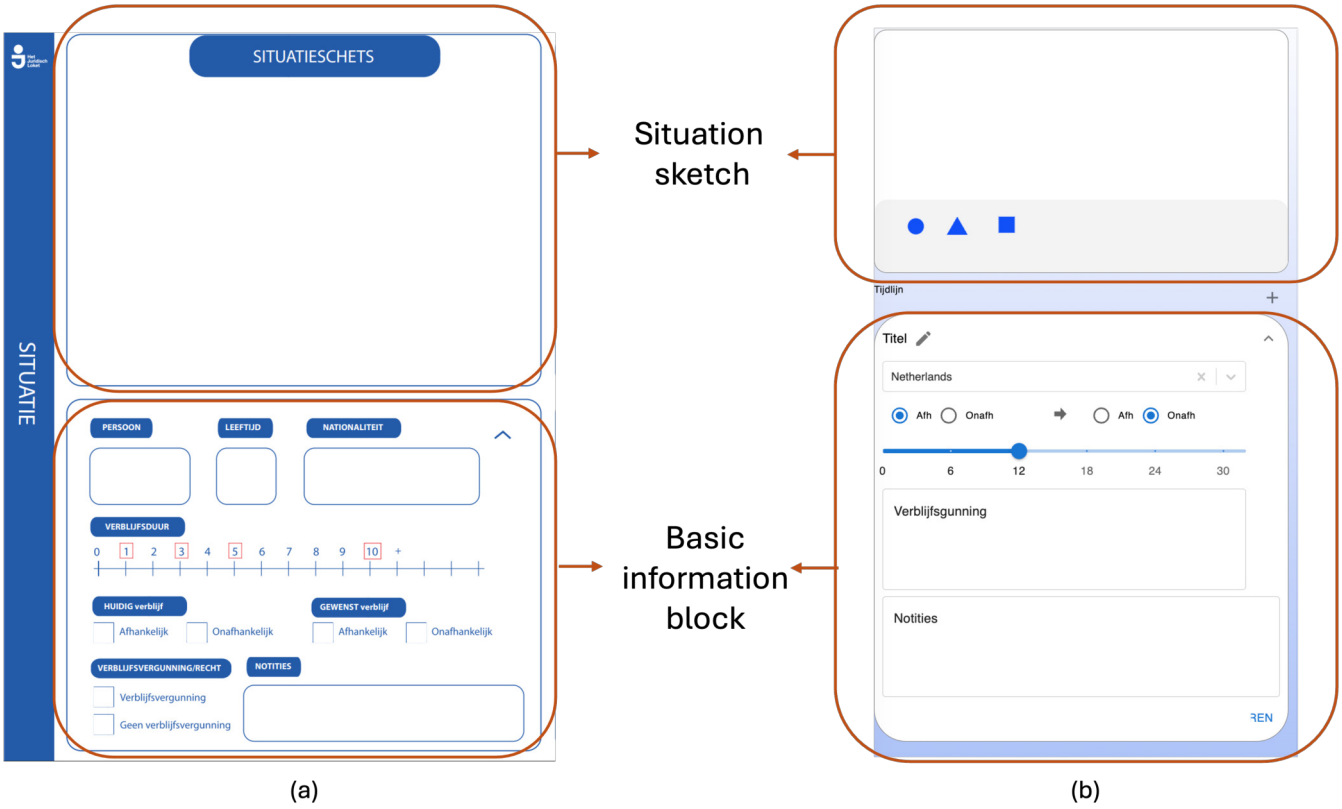


Fig. 3. Situation canvas in (a) paper and (b) web application (in Dutch, with labeled legends).

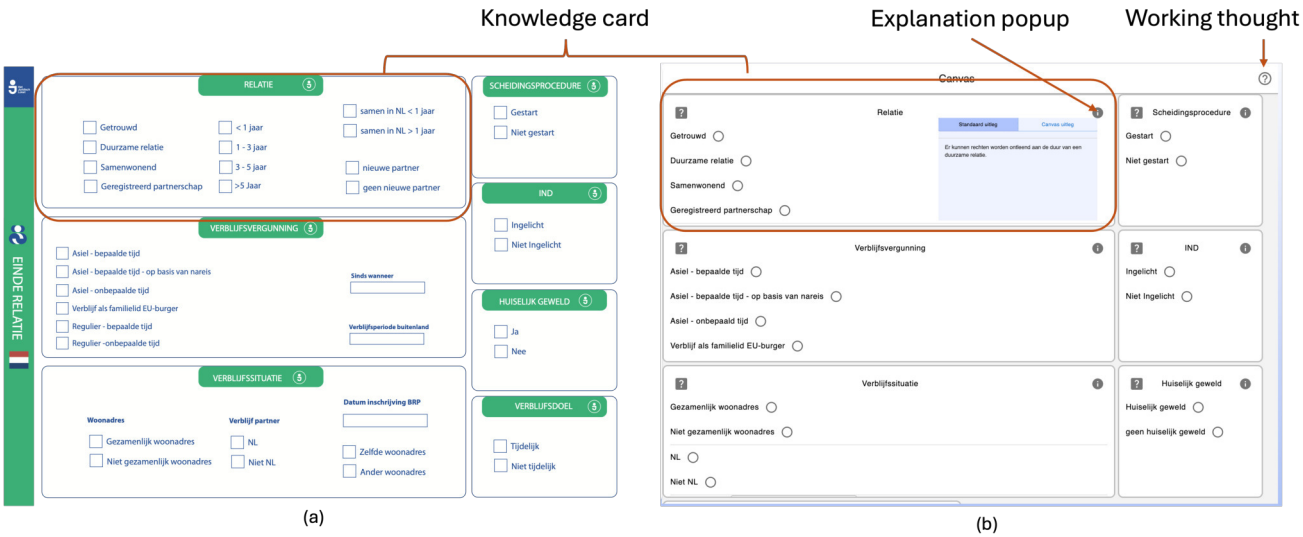


Fig. 4. Conversation canvas in (a) paper and (b) web application (in Dutch, with labeled legends).

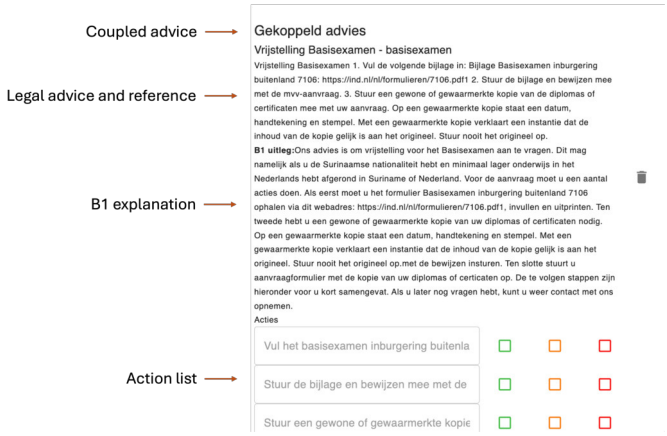


Fig. 5. Example of an advice with the actions list (in Dutch, with labeled legends).

Concretely, the action list should contain tasks to complete, for example, bringing proof of marriage certificate. Advice cards can be associated to the answer combinations on the canvas. When the desired advice cards are picked for the canvas during a conversation, the advisor can indicate the status of the actions in the action list, where "red" means the action cannot be completed, "orange" as action in progress, and "green" as action completed. This way, both the advisors and clients, and also the subsequent advisors taking over this case have a clear overview of the status of the case. The B1 Dutch explanation and the action list are automatically generated into a PDF document available for printing.

D. Automated advice suggestion

The conversation canvas is stored after creation and becomes non-editable during use (i.e., during consultation hours). The conversation canvas now contains strategic questions that need to be answered before giving advice to clients (the need depends on the strategic path of posing questions chosen by the advisor). Based on the single-choice and multiple-choice answers (excluding free text responses for this iteration), the web application automatically suggests predefined or previously given advices for which the answer choices match. This is achieved by storing the answer combinations along with the advice when the conversation canvas is saved. As the user selects the same answer combination in a new conversation canvas, the application back-end runs the match check. The matched advice is shown at the bottom of the conversation canvas. Advisors can review all matched advices and select applicable ones. After selection, the advices are modifiable for the specific instance without changing the original advices serving as more general templates.

V. Use Case

A. Give advice with the use of the canvas

In the current way of working at the Legal Desk, advisors take legal consultation hours (approximately 60-minute sessions for immigration law) with clients after intake at the desk. The consultation hours are oriented towards problem-solving for the client case. Case details are registered in Webtop in the form of text. Advice to the client can be emailed to clients.

The consultation process largely relies on strategic decision-making based on strong expert knowledge. The SHARE system aims to facilitate this process in improving the working experience of advisors and client engagement during consultation. Through the use of the situation and existing conversation canvases categorized by legal frameworks, the SHARE system enables a more streamlined process of asking questions and gathering evidence. We posit that using the system would benefit less experienced legal advisors in particular as they could rely on previous cases and their advice, and structured references of legal information. We also hypothesize that the SHARE system would be welcomed by clients as they can visualize their situation and work through the conversation canvas with the advisor during consultation. Results from the user test designed for this use case and our findings are presented in Sec. VI-A.

B. Create a canvas with a group

To complement "online" real-time support to advisors during client consultations, the SHARE system also facilitates learning "offline", especially during collaborative reflection sessions. We envision that advisors would discuss patterns from cases and create new conversation canvases, knowledge cards, and advice cards to suit their needs. Advisors could synthesize structured knowledge, supplemented by legal references, and store them in the web application.

Knowledge management and efficient knowledge transfer are critical for retaining institutional knowledge (to mitigate knowledge erosion), promote innovation (inspired by group discussions), and ensure continuity (to minimize the impact of employees turnover and retirement). Crossan et al. take these distinct stages together in the four pillars of a learning organization: intuition, interpretation, integration and institutionalization [31]. The SHARE system not only supports documenting explicit knowledge (e.g., links to documents, websites, etc.) but also captures tacit knowledge by enabling advisors to document each consultation in a structured way.

We believe that such a system would promote Communities-of-Practice in introducing a way of working that anchors group learning. By collaboratively reflecting on representative and challenging cases using the SHARE system, advisors can actively engage with one another. This interaction encourages joint brainstorming of creative solutions, benefiting both experienced and less experienced advisors. While experienced advisors contribute their insights, less experienced advisors gain valuable knowledge to better prepare for consultation hours. Results from the user test designed for this use case and our findings are presented in Sec. VI-B.

VI. Evaluation

To evaluate the SHARE system prototype, we have conducted a series of usability tests focusing on the aforementioned two use cases. The usability tests aimed to assess the

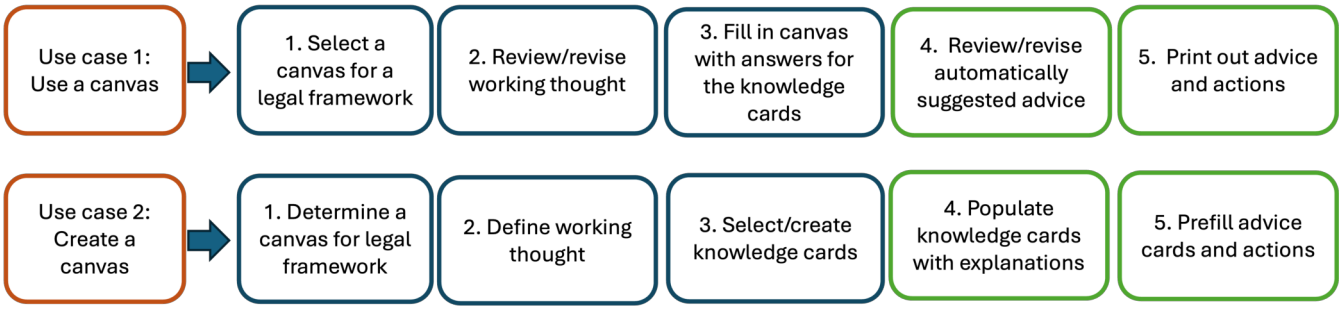


Fig. 6. Workflow of the two use cases of SHARE. Green-outlined boxes show the functionality that is only available in the web application, not on paper. Note that a canvas needs to be created before it can be used. SHARE provides a growing number of starter template canvases.

perceived usefulness and ease of use of the system. Post-task questionnaire evaluations were collected, with questions based on the questionnaire from [32]. The questionnaire is on the 7-point Likert scale, where 1 represents "extremely unlikely", 7 represents "extremely likely", and 4 represents "neutral" (see list of questions in Table I). Verbal feedback was recorded as notes by research members. Two additional observation sessions and a client test session were conducted, which provide triangulation to validate our findings and gain a more comprehensive understanding. While these pilot tests are preliminary due to a lean development process, they provide positive indicators for deployment of the system in the organization after the pilot and shed light on some limitations and future work.

A. *User test: giving advice*

This study focuses on using the SHARE system to advising clients. The experiment was conducted in two sessions at the Legal Desk headquarter office. The first session was attended by three more experienced legal advisors and the second session was attended by one experienced advisor, one strategic development manager, and four less experienced legal advisors. The categorization of experience level is made based on a combination of years of experience, self-assessment, and perception from the strategic development contact involved with the research team. The experimental setting is a low-fidelity role-play simulation. The participants were divided into groups of two with a member of the research team joining the experiment as a facilitator and primarily acted as the client seeking legal help.

In both sessions, the groups were presented with four example legal cases from the legal domain of integration and immigration law. These four cases are fictional but have been reviewed by an experienced advisor prior to the session to ensure that they are representative of actual cases. They were first asked to experience the paper print-out copy of the conversation canvas for the role-play consultation of one case. During this experiment the advisors filled out the situation sketch (i.e., basic information about the case and clients) and the interactive canvas (i.e., the canvas containing the strategic questions and their answer choices). Subsequently, they filled out a post-task questionnaire about this task [3].

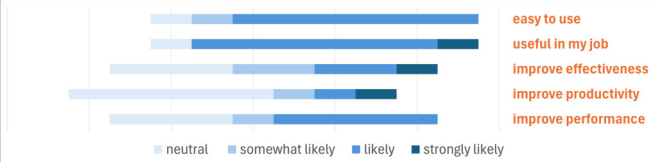


Fig. 7. Overview of ratings distribution of the web application for giving advice.

A tutorial walk-through of the web application was shown to the experiment participants before they started the experiment. The tutorial demonstrates basic operations in the application while working with an example case. It was played once at the beginning at the session. The participants then completed the tasks of working through four legal cases on their own work laptop. Advisors in the same group alternated the control of the application between cases. After completing the cases, the advisors were asked to fill the post-task questionnaire. The results are shown in Fig. 7 for the ratings distribution related to ease of use and long-term effect for job performance. The session concluded with a round-table discussion and feedback.

The results indicate a positive impression for perceived usefulness and ease of use. Despite the SHARE system being a proof-of-concept, the results show that there is potential to improve effective, productivity, and performance upon adoption and system iteration when giving advice to clients. In the context of this user test, results from the post-task questionnaire and after-scenario test from the paper and web version are comparable. Sample size of this comparison is small with three users that participated in evaluation of both versions. Therefore, our assessment remains qualitative. One user gave the feedback that "except of some points in the document (referring to the paper canvas), it is easy to fill in", and that "at first it take some getting used to [do the task]".

[3]This evaluation of the physical version was only conducted in the second session with the less experienced advisors, as the more experienced advisors were already familiar with the physical canvases from previous co-creation sessions.

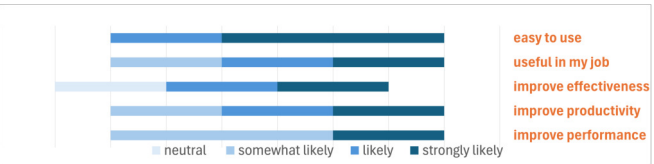


Fig. 8. Overview of ratings distribution of the web application for creating a new canvas.

B. User test: create a new canvas

This study aims to explore the usability of the SHARE system for creating a new *digital* interactive canvas, also as a means of learning and knowledge transfer between legal advisors. This study consists of one session attended by two more experienced and one less experienced legal advisors ($N = 3$). In this session, the advisors worked through a series of tasks in the web application, with increasing complexity [4]. A video tutorial of this creation process was shared with the advisors before the test session. Note that the tasks in the exercise were different from the ones shown in the tutorial video. The tasks mainly covered the following topics: (1) edit and create knowledge cards, (2) edit existing canvases, (3) create new canvases, and (4) edit and create advice cards.

We designed these tasks such that upon completion, the users should be able to make adjustments to existing components in the web application and create new content (e.g., a new canvas with complete knowledge cards and advice cards). The results of the post-task questionnaire are shown in Fig. 8 for a subset of the questionnaire questions related to ease of use and long-term effect for job performance. It shows a generally positive indication for perceived usefulness and ease of use in this *Create* mode of the web application.

C. Questionnaire results analysis

In Table I, we report the average of the ratings across subjects and all questions for perceived usefulness and perceived ease of use, for both user tests. After segmenting results from less experienced and experienced legal advisors, we found slightly higher averaged perceived usefulness scores (4.88 out of 7) from less experienced advisors than experienced advisors (4.25 out of 7). This result is in line with the intuition that experienced advisors may not need the SHARE system to facilitate them during advice consultation in the first place, therefore finding it less useful. Note that these opinions from the advisor were formed before learning about the usefulness according to their clients, see Sec. VI-E. We also note that the perceived short-term usefulness is lower than long-term usefulness for both use cases. To supplement these results and demonstrate the extent of inter-rater agreement, we calculate the Krippendorff's alpha values for over all the ratings on both user tests. We found that there is a fair agreement [33] in the advisors' evaluation for both use cases, 0.27 and 0.20, respectively. However, we also note the disparity in ratings among the advisors. With these results, we aim to demonstrate the collective perception of the SHARE system among advisors who agreed to our tests, and we emphasize that they are not implicative of generalization capability. These post-task questionnaires are self-reports and, though anonymized, taken in the presence of the research team, which leads to subjectivity and bias of the measure. While the results indicate a positive change with the use of the SHARE system, the lack of a control group and objective limits the ability to determine whether the perceived usefulness and ease of use can be attributed to the system itself.

Due to the small sample size, we supplemented the questionnaire with a round of post-task discussions where the advisors gave verbal feedback. We highlight the comments related to lack of usefulness on a short-term basis (translated from Dutch). These comments corroborate the average ratings that we see in the questionnaire.

> "I think that the SHARE web app is now designed with a structure that is flexible. The frame that has been set up is clear and easy to use. But it will take a lot of time in the short term to fill in all the legal information and to ensure that all components are complete. It is a product that will be very useful in the long term, it will still be quite a task to work everything out." - an advisor's feedback after the giving advice user test

> "I think this system has an incredible amount of potential. It will take a lot of work to process all the relevant legal information, but the app is user-friendly and offers the possibility to do so." - an advisor's feedback after the creating canvas user test

Even for the use case in giving advice (where the advisors use a fixed template canvas), one advisor points out that there is still much information to be filled in, potentially referring to the need to revise the advise card and action lists. Surprisingly, creating a canvas is perceived to be easier than using a canvas when giving advice, although creating a canvas is of higher complexity than using an existing canvas. It could be because some advisors already had been exposed to the web application during the "giving advice" user test which predates the "creating a canvas" user test (Sec. VI-B). These effects and observations need to be further confirmed with larger sample size in the future.

D. Observation sessions

In addition to our findings from the post-task questionnaires, observations were conducted in two separate accounts with a focus on observing aspects related to organizational learning (i.e., evidence of knowledge sharing within the organization) The first observation session aimed to inform the design of the physical canvas, which occurred before the "Giving Advice" test (see Sec. VI-A) and was attended by two experienced advisors and one less experienced advisor. The participants agreed that the overview and the visual aspects of

[4]Here we loosely determine the notion of complexity as the transition between different modules and the number of clicks to achieve the desired results.

TABLE I
PUEU RATINGS FOR "GIVING ADVICE" AND "CREATING A CANVAS" (MIN: 1, EXTREMELY UNLIKELY; MAX: 7, EXTREMELY LIKELY).

	Giving advice ($N = 9$)			**Creating a canvas** ($N = 3$)		
Perceived usefulness	Avg(Std)	Min	Max	Avg(Std)	Min	Max
Using the system in my job would enable me to accomplish tasks more quickly	4.6(1.4)	2	6	5.3(0.6)	5	6
Using the system would improve my job performance (short-term)	3.7(1.5)	2	6	4.7(1.5)	3	6
Using the system would improve my job performance (long-term)	5.2(1.0)	4	6	5.7(1.1)	5	7
Using the system in my job would increase my productivity(short-term)	3.3(1.1)	2	5	4.7(1.5)	3	6
Using the system in my job would increase my productivity (long-term)	4.9(1.2)	4	7	6.0(1.0)	5	7
Using the system would enhance my effectiveness on the job (short-term)	3.8(1.5)	2	6	4.3(1.5)	3	6
Using the system would enhance my effectiveness on the job (long-term)	5.2(1.1)	4	7	5.7(1.5)	4	7
Using the system would make it easier to do my job	5.2(0.8)	4	6	5.3(1.2)	4	6
I would find the system useful in my job	5.9(0.8)	4	7	6.0(1.0)	5	7
Perceived ease of use						
Learning to operate the system would be easy for me	5.8(1.0)	4	7	6.7(0.6)	6	7
I would find it easy to get the system to do what I want it to do	5.4(0.9)	4	7	6.7(0.6)	6	7
My interaction with the system would be clear and understandable	4.9(0.6)	4	6	6.3(0.6)	6	7
I would find the system to be flexible to interact with	4.8(0.8)	4	6	6.0(1.0)	5	7
It would be easy for me to become skillful at using the system	5.8(1.0)	4	7	6.7(0.6)	6	7
I would find the system easy to use	5.6(0.7)	4	6	6.7(0.6)	6	7

the physical canvas and the knowledge cards would be most helpful during collaborative reflection and client consultation hours. It also offers support to see the difference between exceptions or business-as-usual cases. The less experienced advisor especially appreciated the structure and the guidance of the canvas and the cards. The more experienced advisors specifically refer to the Shape Language and the value of using these in explaining complex issues. It was observable that the participants are slightly pushed out of their comfort zone, as the canvas is not intuitive and needs explanation. However, we interpret this as an opportunity as being out of comfort zone could encourage learning behavior and deliberations and interactions among each other [34].

The second observation session focused on observing advisors using the web application for knowledge transfer and sharing. In the session, the participants were supported in using the SHARE instrument while they discussed a case. It seems as if they adjust their way of thinking towards the way the elements work together. The way that the instrument is designed and its intentional generalizable layout provoked the participants to discuss nuances. This aspect energizes the meeting, whereas the explanation working through the exercises in creating a new digital canvas required effort from the participants. Strong form of expert knowledge gained from experiential learning consisting of both canonical and noncanonical work practices and their discussions was observable from the sessions.

E. Client test sessions

We observed the SHARE system being used in two client advice sessions with an experienced legal advisor. Both clients signed the informed consent to participate in the observation session and agreed to a short interview after the session. One case concerns social benefits with part-time work; another concerns right to stay in the Netherlands after end of relationship. To integrate the SHARE system in the consultation, the procedure is divided into 3 stages: (1) preparation of the interactive canvas and working thoughts for the case context acquired from the counter visit intake, (2) using the physical and digital canvases interchangeably during consultation, and (3) review the outcome of the consultation and integrate information into Webtop (which is the current system for registering case information). The research team did not give any other instructions on how to use the SHARE system.

Through observations of two client cases, we found that both the system and workflow have room for improvement, with some highlights from client feedback. The ad-hoc update of the canvases, given the case scenario, could be cumbersome yet necessary. Modifying the canvas template in the web application in preparation for a specific client takes approximately 5 minutes, as it involves reviewing the existing knowledge cards and choosing additional applicable others. The need to modify a template arose because of the currently limited number of templates. Over time, templates have to be made for all canonical cases to ensure broader case coverage. Modifying physical printed canvases requires manually manipulating the layout of knowledge cards on the PDF document. Another point of improvement is that the current advice print-outs lack proper formatting. The advisor resolved this by first creating the advice in SHARE, then copying it to Webtop where they could manipulate the layout, and then printing the advice from Webtop. This made this part less efficient in real-time consultations. For these reasons, the SHARE system still needs to be more integrated with the existing tools that the advisors have to use due to standard practices in the organization currently, both in terms of procedure and infrastructure.

However, both clients provided positive feedback on the utility of the canvas, particularly the paper version, and the shape chips (as part of the Shape Language kit, see Sec. III-B1). In one case, the chips played a significant role in guiding discussions around work and social benefits. After several attempts at verbal explanations, the advisor used the chips to demonstrate and clarify the different components of income, which the client understood. These tools were deemed

helpful in providing a clear and comprehensive overview during the consultations. One client indicated that she saw great potential in using the shape chips as they could symbolize certain concepts and facilitate understanding (translated from Dutch and paraphrased from original verbal feedback).

F. Limitations

1) Bias in results: Despite the general positive trend in how SHARE is perceived by the users at the Legal Desk, there are several sources of bias. In our practice-led approach, we are often deeply embedded in the research process both as a practitioner and an investigator. This dual role blurs boundaries between personal preferences and professional priorities and objective research goals. This likely has an impact on our evaluation of the prototype. While we mitigated these effects through the use of existing questionnaires to avoid leading questions, anonymous submission of the questionnaire, and leveraging the only non-Dutch and non-Dutch speaking researcher to conduct the user tests, these biases would still exist in our results. The use of the only post-task questionnaire is also less ideal due to the lack of objective measures of the consultation and learning process and the shortage of advisor participants (which is a niche group) who are already overloaded with work. More objective assessment of the prototype could be achieved through a scalable roll-out and experimentation based on procedural measures such as task completion time, and qualitative and quantitative metrics of legal advises. The management of the Legal Desk would have to agree with the time investment of their advisors at the cost of offering support to their clients. The evaluation of the system may need to become a continuous effort if and when they decide to roll out the system in the organization.

2) Mode of interaction: Because the SHARE prototype exists in two formats, a physical printout paper version and a web application version, one experienced advisor has reported difficulty in deciding which format to use, which suggests an ambiguity in the purpose of use of either of the two formats.

In the post-session interview, the advisor said that

> "Well, [I prefer] the web version but only if it is complete. The online version, I cannot work as fast as that, ok I'm going to create something else right now. During the meeting with a colleague, it is possible. But with a client, it is difficult and it takes time..."

This implies that the paper version has a clear advantage compared to the web application version during the two consultation hours with real clients. During our observations (without providing instructions to the advisor), it was clear that the advisor made an effort to use both paper and web version. She also reported that there was a lot of juggling, bringing too much cognitive load to keep up with using both formats while trying to hold a conversation with the client. Another possible reason is the insufficient familiarity and training, particularly for adjusting an existing canvas or creating an new canvas with the web application in the little time they have for preparing for a consultation hour.

This limitation could be addressed by further optimizing the dual-format usage, and by providing clear guidance on when to use which format. With further testing for both the paper version, digital version, and their combination with control groups in the experimental setup and clearly defined metrics for efficiency and effectiveness (e.g., time to complete the same task), we could make a more informed decision based on the (dis)advantages of both formats.

In the meantime, the advisor also indicated the potential of using the system for training new colleagues.

> "I think it will be useful to a new colleague...She will start advising for immigration law. It [the SHARE system] will be useful for learning and training and see how to do things. It'll be easier than me telling them this is how it works."

This statement corroborates the second observation session on assessing how SHARE contributes to enhancing organizational learning capacity, where group discussion and information exchange were observed. Even at this prototype stage, the advisor thinks that it will already be helpful in training with the existing form-factor and partial information (focusing on immigration law) in SHARE.

3) Filling in the appropriate legal content: In the current SHARE approach, an important challenge to address is to ensure sufficient coverage of legal frameworks and information. So far we have solicited information from the advisors and built the canvases manually with the advisors' contributions. The question is to which extent we can extend the SHARE web application to streamline and (partially) automatically generate the sets of relevant questions, after which the advisors only need to review, revise, and order the knowledge cards on the new canvas and supplement it with their expert knowledge. Given the flexibility of the SHARE web application, the advisors could review and revise the automatically generated information and supplement it with their expert knowledge. Using this approach, the process could be more time-efficient than starting from scratch using google forms and manually setting up the knowledge cards, references, and explanations.

Using a toy example case provided by a legal advisor regarding the right to invoke legal presumption for labor, we found that automatically generated responses from large language models to the expert-provided working thought led to a largely overlapping set of knowledge cards compared to the ones provided by the legal advisor. We note that while the automatically generated responses are a feasible starting point for automating the creation of canvases, some pivotal questions are only present in the legal advisor's approach. However, this provides us with an approach to scale the system beyond its prototype phase.

4) Technology integration, maintenance, and ownership: Paper and digital version of the SHARE learning system and their instructions are accessible and integrated into the Immigration Law department's Microsoft Teams channel. There is an agreement with the research team for light maintenance of the current version of the web application, and possibilities to answer any questions related to use.

However, the web application is currently not integrated into the technology stack at the Legal Desk. The integration is challenging due to the existing third-party software company already maintaining the technical infrastructure for the organization. The research team has limited resource to interface with the software company together with the Legal Desk and comply with existing regulations. For this reason, we have explicitly instructed the advisors not to put any personally identifiable information of the clients into the SHARE system. To allow for scalable deployment and to ensure that any potential information of the clients remains within the Legal Desk for privacy reasons, the web application needs to be integrated within the existing infrastructure (i.e. Microsoft Azure services) in the future.

VII. Discussion

A. Preserving the qualities of paper version of SHARE

As explained in the system's evaluation, both the physical paper and the interactive web versions of canvases have their advantages and disadvantages. The question naturally arises if we can combine the advantages of both versions. The idea was formulated to use tablets running the application on the table between advisor and client as an option that comes close. This approach maintains the benefits of physical interaction while incorporating the advantages of the interactive web version, providing the client with a clear overview and structure during the consultation. However, the advantage of touching the shapes for enhancing the feeling of co-creating and ownership for both the advisors and clients would not be captured by the use of the tablet. Further research is necessary to enable situated real-time and multimodal understanding of the advisor-client interaction.

B. Generalization to other departments and organizations

Currently SHARE focuses on the immigration law department at the Legal Desk but can be generalized and adapted to other topics and organizations by leveraging its core principles of flexibility and knowledge-sharing. On a content-level, it would be straightforward to augment the existing immigration law with other legal topics (e.g., labor relations) at the Legal Desk, as we could query the Knowledge Bank. On the approach-level, many disciplines show similar ways of working, that focus on experiences (i.e., the transition from inductive reasoning to deductive reasoning as novices become experts). Novices gain from inductive reasoning to build experience through trial-and-error. Experts rely on deductive reasoning for efficient problem-solving, which starts from general principles or rules (in our case, the working thoughts) to specific cases. The SHARE system explicates the problem-solving approach and we envision it to cut down spent and wasted effort in learning from trial-and-error. SHARE elucidates and trains users in acquiring a type of strategic and structured knowledge that is crowd-sourced by the experts and built collaboratively and organically within the organization. Taking these guiding principles of the SHARE system, organizations could adapt the content and approach to fit teams that wish to learn and train more efficiently.

C. Continuous engagement with the learning system: a mindset for organizational learning

Engaging continuously with learning systems requires consistent and regular interaction with both formal and informal channels of knowledge, including structured training programs, peer collaboration and feedback, and personal development initiatives. Organizations like the Legal Desk must integrate multiple components to support systematic learning, such as a technology platform for storing strategic expert knowledge (e.g., SHARE), a strong learning culture that encourages exploration and idea-sharing, and leadership that prioritizes continuous learning as a core value.

From a technical perspective, we envision the SHARE learning system evolving towards being more automated through the integration of richer datasets and more advanced underlying algorithms. This trajectory aligns with a broader goal of embedding SHARE within a hybrid human–AI learning organization for future work. Building on the framework introduced by Tan et al. [35], the human–AI learning loop is designed to support a dynamic spectrum of knowledge exchange modalities. These range from traditional human–human interactions to collaborative human–AI learning, and even fully autonomous multi-agent AI–AI knowledge transfer. Each mode and learning pathway contributes uniquely to the hybrid human-AI system's adaptability and capacity for continuous improvement, and foreshadows the potential of SHARE as a scalable learning platform.

From an organizational point of view, the full implementation of SHARE and the aforementioned human-AI learning framework involves aligning learning activities with strategic goals and existing work activities. Challenges such as time constraints, as seen in the Legal Desk's immigration law department, can hinder consistent engagement. This should be taken seriously since these time constraints may also lead to the lack of integration of organization learning into existing organizational systems [31] and eventually missing the opportunity to develop a sustainable learning ecosystem between the various groups of professionals [29].

Overburdened advisors often have to prioritize immediate tasks over dedicated learning, making it difficult to balance work responsibilities with continuous education. Flexible learning opportunities, such as micro-learning sessions or workflow-integrated learning, can help mitigate these challenges. To this end, we believe the SHARE system has taken a step in demonstrating how learning can be integrated into the way of working for an organization with limited resources. SHARE enables a minimum setup of the hybrid human-AI organizational learning framework where employees become part of a learning organization and actively contribute to joint knowledge transfer and development without requiring significant change to the way of working.

VIII. Conclusion

In this paper, we introduced a novel interactive learning system designed to enhance legal consultation services and training at the Legal Desk. Using a co-creation approach, we collaboratively defined the problem of the organization's high turnover rate among legal advisors and developed a solution motivated by this critical challenge. Our proposed system, SHARE, integrates physical tabletop components with a digital web application. SHARE supports the explication of strategic expert knowledge on conducting efficient and effective consultation hours to provide their clients consistently with appropriate and customized advice. SHARE stores and presents that strategic knowledge to all advisors (experienced and less experienced) in such a way that less experienced advisors benefit from this strategic knowledge. Each component serves distinct purposes and leverages its unique strengths, working in synergy to achieve the desired outcomes for improved learning and consultation experiences for clients. We evaluated the prototype using a combination of user tests, observations, and interviews. We found a positive perception towards the prototype (further confirmed by the deployment of the current prototype in the organization) both from the advisors and from the clients point of view. We discussed the limitations of the prototype and suggested potential future work. These include designing and measuring more objective measures and scaling up experimentation with more users. Overall, our findings highlight the potential of SHARE as an effective instrument for addressing organizational learning challenges by enhancing legal consultation experiences and training. Future iterations of the system can be further refined for impact and scalability, paving the way for broader adoption and sustained benefits within the Legal Desk. The system can also be extended and adapted for other organizations that experience similar knowledge erosion problems and users that do similar types of knowledge-intensive work as legal advisors.

Acknowledgment

This research was (partially) funded by an NL Ministry of Justice and Safety subsidy under the project title " SHARE — Vind Samen het Alternatief". We thank het Juridisch Loket for their contribution to this project.

References

[1] J. Han, J. kang Lu, Y. Xu, J. You, and B. Wu, "Intelligent practices of large language models in digital government services," *IEEE Access*, 2024.

[2] W. J. Pieterson and W. E. Ebbers, "Channel choice evolution: An empirical analysis of shifting channel behavior across demographics and tasks," *Government Information Quarterly*, vol. 37, no. 3, p. 101478, 2020.

[3] P. C. Brayer, "A law clinic systems theory and the pedagogy of interaction: creating a legal learning system," *Conn. Pub. Int. LJ*, vol. 12, p. 49, 2012.

[4] Het juridisch loket jaarverslag 2023. Stichting Het Juridisch Loket. [Online]. Available: https://media.production.juridischloket-dev.nl/JL_JAARVERSLAG_2023_ONLINE_6a4ed0a716.pdf

[5] H. W. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy sciences*, vol. 4, no. 2, pp. 155–169, 1973.

[6] M. Easterby-Smith, M. Crossan, and D. Nicolini, "Organizational learning: debates past, present and future," *Journal of management studies*, vol. 37, no. 6, pp. 783–796, 2000.

[7] C. Argyris, "Double loop learning in organizations," *Harvard business review*, vol. 55, no. 5, pp. 115–125, 1977.

[8] I. Nonaka and H. Takeuchi, "The knowledge-creating company," *Harvard business review*, vol. 85, no. 7/8, p. 162, 2007.

[9] S. C. Goh, "Toward a learning organization: The strategic building blocks," *SAM Advanced Management Journal*, vol. 63, pp. 15–22, 1998.

[10] J. Lave and E. Wenger, "Communities of practice," 1996.

[11] E. Wenger, "Communities of practice: The key to knowledge strategy," in *Knowledge and communities*. Routledge, 2009, pp. 3–20.

[12] S. Borzillo, S. Aznar, and A. Schmitt, "A journey through communities of practice: How and why members move from the periphery to the core," *European Management Journal*, vol. 29, no. 1, pp. 25–42, 2011.

[13] S. Kurtz, M. Wylie, and N. Gold, "Problem-based learning: An alternative approach to legal education," *Dalhousie LJ*, vol. 13, p. 797, 1990.

[14] A. Lynch, "Why do we moot-exploring the role of mooting in legal education," *Legal Educ. Rev.*, vol. 7, p. 67, 1996.

[15] D. I. Thomson, "Defining experiential legal education," *J. Experiential Learning*, vol. 1, p. i, 2014.

[16] V. Bouki, D. Economou, and P. Kathrani, ""gamification" and legal education: A game based application for teaching university law students," in *2014 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL2014)*. IEEE, 2014, pp. 213–216.

[17] N. J. Knauer, "Learning communities: A new model for legal education," *Elon L. Rev.*, vol. 7, p. 193, 2015.

[18] F. Delgado, S. Yang, M. Madaio, and Q. Yang, "The participatory turn in ai design: Theoretical foundations and the current state of practice," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–23.

[19] J. Gregory, "Scandinavian approaches to participatory design," *International Journal of Engineering Education*, vol. 19, no. 1, pp. 62–74, 2003.

[20] J. Simonsen and T. Robertson, *Routledge international handbook of participatory design*. Routledge, 2012.

[21] C. A. Le Dantec and S. Fox, "Strangers at the gate: Gaining access, building rapport, and co-constructing community-based research," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 1348–1358.

[22] B. Cooke and U. Kothari, *Participation: The new tyranny?* Zed books, 2001.

[23] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *Co-design*, vol. 4, no. 1, pp. 5–18, 2008.

[24] M. Steen, "Co-design as a process of joint inquiry and imagination," *Design issues*, vol. 29, no. 2, pp. 16–28, 2013.

[25] C. Frayling, "Research in art and design (royal college of art research papers, vol 1, no 1, 1993/4)," 1994.

[26] J. McNiff, *Action research: Principles and practice*. Routledge, 2013.

[27] C. L. Larson, J. Aronoff, and E. L. Steuer, "Simple geometric shapes are implicitly associated with affective value," *Motivation and Emotion*, vol. 36, pp. 404–413, 2012.

[28] M. Novak and S. Schwan, "Does touching real objects affect learning?" *Educational Psychology Review*, vol. 33, no. 2, pp. 637–665, 2021.

[29] I. Pyrko, V. Dörfler, and C. Eden, "Thinking together: what makes communities of practice work?" *Human relations*, vol. 70, no. 4, pp. 389–409, 2017.

[30] B. H. Banathy, *Designing social systems in a changing world*. Springer Science & Business Media, 2013.

[31] M. M. Crossan, H. W. Lane, and R. E. White, "An organizational learning framework: From intuition to institution," *Academy of management review*, vol. 24, no. 3, pp. 522–537, 1999.

[32] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[33] J. Hughes, "krippendorffsalpha: An r package for measuring agreement using krippendorff's alpha coefficient," *arXiv preprint arXiv:2103.12170*, 2021.

[34] D. A. Kolb, *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[35] S. Tan, W. Aartsen, D. van Hamersveld, and C. Jonker, "Towards hybrid intelligence in learning organizations," *Hybrid Human-AI conference*, 2025.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 25–34
DOI: 10.15439/2025F4150
ISSN 2300-5963 ACSIS, Vol. 43

New Education Challenges in Profiling Digital Experts for a Digital Economy Era

Ivan Luković
0000-0003-1319-488X
University of Belgrade – Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: ivan.lukovic@fon.bg.ac.rs

Marija Đukić
0000-0002-1136-4278
University of Belgrade – Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: marija.djukic@fon.bg.ac.rs

Dragan Stanojević
Independent Consultant
Griffenfeldts gate 17 F,
0460 Oslo, Norway
Email: dstanoje@gmail.com

***Abstract*—Nowadays, modern business includes acquisition and storing enormous data volumes, larger than ever before. Such data represent a significant value that an organization or a society can utilize to reach created goals and provide sustainable development. Unfortunately, a daily practice still intensively points out to the problem of a serious gap between the identified needs for knowledge, on one hand, and inability of the disciplines of Computer Science, Software Engineering, Informatics, and Data Science (CSSEIDS, for short), combined with the modern software technologies, to address such needs in an effective way, on the other hand. One of the important causes of such phenomenon is in a lack of strongly educated and interdisciplinary oriented experts showing an appropriate level of knowledge both in CSSEIDS, as well as in the disciplines of Business, Management and Economics, for a specific problem domain. In this paper, we address issues on how to come to more flexible and interdisciplinary oriented study models capable of producing various forms of digital managers and digital engineers, as a new profile of experts, ready to cope with digital economy and digital transformation in a modern society. Massive deployment of such experts is a way to significantly raise the level of organization maturity regarding capabilities for: information management, quality management, business processes, and big data analytics.**

***Index Terms*—Digital Transformation, Education of Digital Experts, Digital Engineer, Digital Manager.**

I. Introduction

NOWADAYS, modern business includes acquisition and storing enormous data volumes. Practically, we are faced with exponential growth in the amount of data collected. In [1], the authors discuss about big data as an emerging phenomenon. They state that "computing systems today are generating 15 petabytes of new information every day—eight times more than the combined information in all the libraries in the U.S.; about 80% of the data generated everyday is textual and unstructured data." As such, we can state that we are in the Age of Big Data. Current state of data and knowledge management in modern business, characterized by the big data age, is in that the data represents a significant value that an organization or a society can utilize to reach created goals and provide sustainable development. By [2], Big Data is the next frontier for innovation, competition, and productivity.

By our experience, in well-matured companies, the company management typically shows a clear recognition of needs for generating corporate knowledge from data. To address this need, they are ready to deploy quantitative, analytical methods, with the aim of effective data use in the decision process and various company management activities.

Despite that there is a clear recognition of significant worth ingrained in stored data, unfortunately it is still mostly unexploited. Data is operationally used in a short time frame, and then archived, without further utilization in the process of generating corporate knowledge. Tremendous amounts of data are available, while many institutions fail to make efficient use of the huge amount of data available, or look for patterns. One reason notified in [3] and [4] is that it is because the business appetite for doing so didn't exist, while the survey presented in [5] notifies the main obstacle to use advanced data analytics in the lack of understanding how to apply it.

Hereby we identify a great discrepancy between business needs on one hand side, and Information Technology (IT) capabilities on the other hand side. We notify significant needs for generation of corporate knowledge from data, while we notify inability of modern software products to effectively address these needs, despite that huge amounts of data already exist, and modern IT tools provide excellent technical capabilities. Such phenomenon can be perceived as a new form of Software Crisis introduced even in 1970's. However, it is never ending, present even for decades in various forms. Now, we can call such phenomenon a Big Data Crisis. From the managers' point of view, the phenomenon is perceived as an evident lack of return of investment (ROI) in big data projects in which engineers are unable to satisfy stakeholders' and managers' evident needs. However, from the engineers' point of view it is perceived as managers' and stakeholders' insufficient and unsatisfactory understanding of both IT capabilities, as well as business and decision processes.

However, Big Data Crisis is not just about perceiving factors 'on the surface' of the problem. It is to go deep into its causes and address them in a strategic way. We can identify the following main causes of Big Data Crisis, covering both managers' and engineers' side of the problem:

(A) Unsatisfactory level of organization maturity regarding the capacities for: information management; corporate knowledge management; business processes and institutionalization; and quality management;
(B) Unsatisfactory level of accumulated knowledge in a problem domain; and
(C) Unsatisfactory level of accumulated knowledge in the domain of computer science, software engineering, informatics, and data science (CSSEIDS), necessary for the development and formal specification of models for software products aimed at generation of company knowledge and decision support.

Addressing the Big Data Crisis causes is a strategic and long-life task. It implies addressing all its significant causes simultaneously.

In this paper, we intend to contribute to addressing (B) and (C) causes. It is an important endeavor related to the education process, and mostly formal academic education process. Our goal is to address issues on how to come to more flexible and interdisciplinary oriented study models capable of producing various forms of digital managers and digital engineers, as a new profile of experts, ready to cope with Digital Economy (DE) and Digital Transformation (DT) in a modern society. We believe that massive deployment of such experts is a way to address in some aspects the (A) cause, i.e. to significantly raise the level of organization maturity regarding capabilities for information management, corporate knowledge management, quality management, business processes, and big data analytics. Let us notify that full addressing of the (A) cause is out of the scope of this paper, and it is a matter of our future research work.

Apart from Introduction and Conclusion, the paper is organized through four sections. The analysis of present research works is given in Section II. In Section III we discuss current findings regarding present education models. Perspectives and requirements about the improved education process are discussed in Section IV, while in Section V we outline a more flexible and interdisciplinary oriented academic education model to cope with Big Data Crisis.

II. Related Work

In recent years, labor markets have experienced significant changes, primarily due to the rapid technological progress introducing new skill requirements for DE. This change resulted in a gap between the skills companies seek and the ones available in the job market. As a result, it is essential to align the education curricula with the industry requirements. Underscoring this point, in [6] it is argued that curricula must be regularly updated to prepare students effectively for the digital landscape. Similarly, in [7] the authors proposed a new framework to measure the alignment between higher education digital skills and industry requirements, deriving factors for graduate preparedness from a systematic analysis of the relevant literature.

To address the evolving needs of the labor market, the education sector is increasingly compelled to adopt a systematic approach to transformation enabled by digital innovations. The authors of [8] analyzed existing DT frameworks to identify the critical components that educational leaders must consider when developing strategies for effective digital change. This study reviewed contributions from 50 organizations involved in education policy, including research institutions and technology companies providing hardware, software, and digital consulting services, as well as 15 Ministries of Education. The findings highlighted four principal themes: Leadership (94%), People (67%), Technology (56%), and Experience (33%). In [9], the authors investigated whether education is perceived as an essential factor in the broader digital transformation narrative by analyzing scholarly works from diverse scientific fields. The findings suggest that both education and workforce competencies are crucial for advancing digitalization and that many barriers to technology implementation can be mitigated through targeted educational initiatives. Furthermore, in [10], the authors argue for the establishment of interdisciplinary research and education centers that focus on IT skills development through industry partnerships and experiential learning.

In [11], it is proposed a theoretical framework to guide development of study programs that align with job market requirements, addressing the shortage of DT leaders. The authors note that universities often struggle to adapt swiftly due to lengthy bureaucratic procedures and curriculum development cycles. The authors of [12] reported on an educational project from the University of Birmingham. In collaboration with industry partners, the aim of the project was to strengthen students' personal development in the digital domain and provide exposure to digital systems used across industries. Subsequent evaluations revealed that this initiative enhanced students' digital competencies, resulting in a 30% improvement in their knowledge of the digital skills covered.

The skills required for the digital economy era cannot be acquired through a single discipline; rather, they demand an interdisciplinary approach that combines academic learning with professional training. Supporting this view, the authors in [13] introduce an educational framework with two components: a Competence Model for identifying essential skills for DT and a Digital Transformation Maturity Model for evaluating both individual and organizational competencies. Similarly, in [14] the authors analyzed the profiles of modern, data-driven companies and identified two groups of skills required for Data Scientist roles: (1) professional and attitude skills and (2) critical thinking, communication, collaboration, organizational awareness, and ethics. Focusing on Brazil's emerging economy, in [15] the authors identified a misalignment between the skills required by IT companies and those taught in IT-related education programs. Companies' managers emphasized that non-cognitive or 'soft' skills distinguish exceptional IT

professionals. The study concluded that Computer Engineering programs tend to prioritize technical skills over social and emotional competencies, highlighting the need for a more holistic educational approach to produce well-rounded graduates.

In contrast, in [16] the authors examined students' perceptions, where students' population is one of the main groups entering the workforce, regarding studying in digital environments and their consequent preparedness for employment in DE. In a case study conducted in the city of Osijek in Croatia, the results indicated that 35.8% of students considered the traditional education model to be unsatisfactory. Additionally, 71.6% of respondents believed that acquiring knowledge in DE would likely or certainly benefit their future employment or self-employment. Similarly, in [17] it is investigated the impact of Digital Transformation Initiatives (DTIs) on the Future Job Prospects (FJPs) of senior engineering students, demonstrating a direct, empirically measurable relationship between comprehensive digital educational environments and enhanced perceptions of career readiness.

From almost all related works analyzed it follows that having deep digital skills will be more and more important factor of better employability in the future job market. A review of scholarly work on this topic highlights the need to revise higher education curricula to provide learning opportunities across diverse domains. Universities must foster stronger connections with industry, as these partnerships can co-create authentic learning opportunities and provide insights into evolving labor market demands. In response to this need, we propose fundamentals of a new education model that introduces the digital expert profiles of a Digital Manager and Digital Engineer. Such model emphasizes the importance of interdisciplinary knowledge, integrating technical competencies such as mathematics, statistics, and engineering with business, economics, social sciences, and soft skills, thereby producing graduates who are not only technically proficient, but also adaptable, collaborative, and prepared to lead in the DE era.

III. Initial Findings

One of the hot issues in creating Computing, Computer Science (CS), Informatics (I), or Software Engineering (SE) curricula in academic education is in what extent they are to be characterized as: (i) "self-contained", (ii) interdisciplinary oriented, or (iii) applied, and how to properly adjust a level of overlapping with other disciplines. It is not always easy to constitute high quality study programs that will properly meet the balance between those three characteristics, and to satisfy current and future needs of industry, at the same time.

As discussed in [18], in software industry of well-developed economies, we can notify a strong fitting between the skill and education requirements for specific job positions, and the level of education and experience of software engineers or IT experts being hired at those positions. Also, such software engineers typically show a higher level of specialization in some disciplines or problem domains. On the contrary, in underdeveloped or even developing economies, fitting between the required skills and education level for the job positions, on one hand side, and the level of education and experience of software engineers on the other hand side, is often not appropriate, and we can notice hiring overqualified or underqualified experts at some positions, to a wider extent. The level of specialization depth of software engineers to some disciplines or problem domains is not as strong, as in well-developed economies. However, it seems that a common characteristic of the Human Resource (HR) market in both cases is that interdisciplinary oriented professionals are always welcomed and better positioned, particularly as the requirements for such profiled experts are typically much over the HR market capacities.

In [15] the authors advocate about significant changes in the technology-related job market over recent years, mainly due to technological advances. It pushes industry toward new demands for skilled professionals, and it is a crucial factor to consider the future industry needs for HR experts. The authors notify: (i) a gap between industry needs and professional profiles available in the job market; and (ii) difficulties in finding employees who meet the required profile. All that results in financial loss and extra training expenses.

Absence of almost any academic education strategy leads to total leaving the academy to the operational market laws. Consequently, it comes to the two paradoxes [18]:

(P1) More interdisciplinary oriented experts, capable of covering a wide range of tasks, knowledge and skills are always significantly better positioned in the software industry HR market, while academic institutions offer study programs that are rather self-contained, i.e. oriented to a narrower knowledge scope; and

(P2) Students or young software engineers believe that they will be better positioned in software industry HR market just as they are good IT experts, i.e. programmers, while employers rather expect experts capable of recognizing and resolving their interdisciplinary oriented and complex requirements.

In [19], it is stated that the culture of interdisciplinary orientation of education is usually poorly developed, while interdisciplinary education is most likely in experiment phase. Even more, there is no consensus on how to evaluate the outcome of interdisciplinary education.

If we say here that academic institutions motivate, often 'in silence', education of 'more specialized experts', that means experts that are pure software or informatics engineers, mathematicians, business administration managers, or various domain experts.

Traditional taxonomy of education and research areas that is commonly applied in the academic accreditation process in some countries (e.g. in Serbia) is very rigid and comes to the disciplines of (a) science and mathematics; (b) engineering and technology; (c) social science and humanities; and (d) medicine. We identify study programs of the three categories, covering in some extent disciplines of CSSEIDS. Those are: (1) Specific study programs in Computer Science, Software

Engineering, or Informatics (CSSEI); (2) Study programs in (Applied) Mathematics (AM); and (3) Study programs in Economics, Business Administration and Management (EBAM). Many universities traditionally provide study programs in all three categories, for decades. A common belief among university education staff is that such traditional disciplines are quite enough to satisfy the needs of future HR market in the DE era.

By our long-term education experience, we identify the typical students' and even teachers' behavioral patterns of all three study program categories, as follows [18].

(1) Students from specific CSSEI study programs are predominantly technology oriented. Often, they express their animosity to the mathematical, and even more organizational, managerial or economics disciplines, with a belief that this knowledge is not necessary to them, and that someone else is to possess it. Study programs of this category often provide just a modest level of knowledge from mathematics and business administration. On the other hand, such students express their strong interest in learning a typical technology knowledge in IT.

(2) Students from AM study programs are predominantly formally oriented. They believe that technology knowledge is of a lower-level value. Also, they are not aware of the necessity of having a knowledge from business administration, management, or economics. Development of skills aimed at practical application of adopted knowledge in various application domains is often underestimated or even neglected. Students from this category believe that complexity of things is just of a logical nature – the things are more complex, just if they are logically complex, while other forms of complexity that evidently exist in engineering and science are rather neglected. Study programs of this category often provide a modest level of CSSEI knowledge, as well as business administration knowledge.

(3) Students from EBAM study programs show a strong awareness about the importance of having the CSSEI and AM knowledge in resolving the complex problems in organization systems. However, in a lack of formal knowledge from these disciplines, they believe that someone else is to resolve such problems, while their task is just to rent high quality CSSEI and AM experts to resolve the problems. Study programs from this category motivate learning a highly formalized knowledge from CSSEI and AM rarely.

Literally, we may say that the three identified behavioral patterns form "a universe of not joinable worlds" [18].

As such, a question arises: who will maintain the 'interfaces' between various disciplines? We need to systematically educate interdisciplinary oriented professionals capable of creating interconnections between various disciplines with a satisfactory deep level of knowledge of details. Our personal view is that academic education in CSSEIDS shows best chances to address this issue successfully, both in the areas of business and organization systems and scientific computations.

To perform some form of a test of our assumption, we consulted ChatGPT (June 2024). The findings are that "education in CSSEI can be closely related to digital manager skills in several ways. Strong technical foundation and problem-solving mindset are crucial for effective digital management. By aligning the technical expertise from CSSEI education with the strategic and operational skills required for digital management, professionals can effectively lead DT initiatives and drive organizational success in the digital age. By evolving the classical education approaches in CSSEI, we come to ability to (i) better prepare graduates for the multifaceted role of digital managers in modern companies; and (ii) equip them with business acumen, and leadership skills required to drive DT and innovation."

Despite that we identified the three behavioral patterns while developing data science curricula as discussed in [18], the patterns are general and should be strongly considered in creating new study programs in support of DE. DT as a movement to DE is driven by changes in: (a) business models; (b) culture; (c) behavior of individuals; and (d) how technology is driving and is driven by these changes [20]. It is much more driven by the first three factors, supported by IT, than has been driven by advancements in technologies themselves. As stated in [20], strategy, but not a technology drives DT. We can say that it should be strongly considered as an important assumption in creating any education program in support of DE and DT.

IV. New Education Perspectives

To provide a movement to DE by deployment of DT principles, an organization must create power for organization transformations that are predominantly disruptive by its nature. It means that the incremental model of improvements is considered too slow and inefficient in most cases. DT strongly implies a faster model of discontinuity with a full reconsideration of the company business model and acceleration of company management in better understanding of the transformation process. DT must provide scaling from digital experiments and pilots to digital best practices [21]. An enterprise ecosystem in DE is organized through five layers, i.e. digital platforms (given here from bottom to top): 1. IT Infrastructure Platform; 2. Delivery Platform; 3. Talent Platform; 4. Leadership Platform; and 5. Business Model Platform. The main characteristics of the digital platform layers are that they should be designed as multidisciplinary oriented, dynamically coupled with a possibility of easy adaptation, with evident capabilities for learning and reconfiguration, and easily extensible with porous borders. A strong requirement to apply DT principles successfully and transform organization into a DE enterprise ecosystem is to develop those new digital platforms. The development process of digital platforms is considered as a unique venture in each organization. There is no silver bullet, no way to buy it anywhere – it must be designed and implemented in the organization. Common requirements for such process are integration of business and IT needs and requirements; and establishing a clear collective vision of the

company management, with the main purpose to provide company growth, as the main business priority [21].

A strong prerequisite for initiating a DT endeavor is to provide enough DE experts with a new profile, capable of high-quality support of the DT process. By [22], workforce skills are the most difficult barrier for building organizational capabilities for DT. We are faced with an evident lack of well-educated digital experts ready to support DT process. This phenomenon is called Digital Talent Gap [22], [23].

Academic education is not an exception. It should be subjected to the DT process, to come to the strategic goals necessary to support movement to DE and effective production of digital experts, and by this contribute to overcoming the Digital Talent Gap problem. As such, academic education is also to be transformed in a way to reach the new digital 5-layered platform. In [24], the authors propose adjusted priorities of academic education and training of teachers with a goal to come to well-educated specialists capable of managing and using digital technologies.

By our viewpoint, academic education is always the first-class entity for profiling DE experts – Digital Managers and Digital Engineers. It should be combined with additional approaches to education, such as short courses, training of various forms, and company schools or universities. In [25], the author proposes several priorities of the academic education transformation. One of the priorities is integration of university and corporate education. While corporate universities provide ultra-modern, but not systematic enough and fundamental education, classical universities are criticized as being still far from modern society education and new teaching technologies. However, they offer very important fundamental knowledge necessary for an expert to last on the HR market and to be able for long-life self-learning and moving to emerging technologies easily. Another important priority identified in [25] is that universities should become drivers of DT of society and economy. Nowadays, a new mission of universities, their technology parks, and business incubators together with teachers is to contribute to the development of (youth) startups.

To address Big Data Crisis Causes (B) and (C) and provide a transformation of academic education to face with DE, we start from the hypothesis that formal academic education is to strongly motivate raising the level of: (i) interdisciplinarity and multidisciplinarity in profiling students' knowledge: (ii) abstraction and formalization skills: (iii) quantification, metrics, and analytical skills; (iv) specification, implementation, and integration skills; (v) business, management, communication, and soft skills; and (vi) problem domain skills. There are three general approaches to initiate modernization of a study program to face DE: a) starting with a CSSEI curriculum; b) starting with an EBAM curriculum; or c) hybrid or Integrated approach. The first approach motivates producing highly technically proficient digital engineers and, in some extent, digital managers, who can bridge the gap between technology and business. The second approach motivates producing digital managers with strong strategic and operational skills, who can understand and leverage technology. The third approach means creating a curriculum mostly from scratch. Integration in this approach means unifying existing CSSEI, EBAM, and possibly AM curricula. The outputs in all cases can be interdisciplinary programs, joint, double, or dual degree programs, or even just collaborative courses.

To initiate the process of study program transformation, it is important to identify and formulate the general requirements to the transformed curriculum. In this case, we predominantly address such requirements for a CSSEI program. A proposal of the general requirements includes the following:

(1) Interdisciplinary Curriculum;
(2) Project-Based Learning;
(3) Soft (Non-Cognitive, Metaskills) Skills Development;
(4) Industry Partnerships and Internships;
(5) Continuous Learning and Adaptation;
(6) Entrepreneurship and Innovation;
(7) Analytical Thinking and Data-Driven Decision Making;
(8) User Experience (UX) and Design Thinking;
(9) Cybersecurity and Compliance; and
(10) Mentorship and Networking.

Identification of those ten requirements was inspired by the authors' experience, as well as the related work research mostly presented in [15], [26], [27], [28], [29], [30], and [31].

(1) Interdisciplinary Curriculum is an important advantage, as traditional CSSEI programs focus heavily on engineering and technology-oriented subjects, while it is a requirement to integrate it with business, management, digital marketing, and leadership courses. Integration with mathematics (algebra, discrete mathematics, probability theory) and statistics is also a must. It is to strive for embedding more holistic educational experience and combine general engineering and CSSEI disciplines. By this, one can create a field where interdisciplinary and multidisciplinary education has room to grow.

(2) Project-Based Learning is an achievement where not only theoretical knowledge with isolated coding projects are practiced. It means real-world project-based learning, working in multidisciplinary teams, simulating the collaborative environment. It is even more effective if it is applied in coordination with the selected companies that are ready to provide their real word problems or case studies to be adapted to students' projects with the company experts ready to be involved in the mentorship program of project-based learning.

(3) Soft Skills, Non-Cognitive, or Metaskills Development is an achievement where focus is not on technical skills only. The program should incorporate soft skills training, such as communication, teamwork, leadership, emphasizing effective communication, conflict resolution, and team management. Required skills for DT era are entrepreneurial vision, creative ability, holistic vision, with a focus on collaboration, ethics and social responsibility, and codes and programming.

Majority of executives consider soft skills as increasingly important, while there are studies suggesting that both soft and hard skills are needed. In [32] and [33], the authors identify the following key leadership skills: communication skills, language knowledge, organizational skills, subject-specific

knowledge, digital literacy, and self-reflection. Executives are to be emphatic, open-minded, flexible, motivated, and stress tolerant. In [34] and [35] the authors identify behavioral leadership skills as core abilities of managers. They include motivational skills, team-building skills, and emotional intelligence. Also, in [35], the authors identify the Altro-centric leadership, as an other-centered leadership style, or adaptive leadership that is mandatory to support DT as a disruptive process. To provide such leadership style, increasingly important soft skills are empathy, humility, integrity and compassion, integrity, creativity, and human–ethical sense of judgement to AI's data-driven information and judgments. Managers' focus of such style is on coaching, motivating and empowering employees, while engineers' focus is on algorithmic management practice, particularly supported by Artificial Intelligence (AI) and Generative AI methods, and on handling quantifiable managerial tasks and quantitative performance evaluation.

(4) Industry Partnerships and Internships require strengthening partnerships with technology-oriented companies and startups. It includes internships, co-op programs, collaborative projects, experience and networking opportunities.

(5) Continuous Learning and Adaptation means that the curriculum is not just fixed with a static set of courses. It nurtures the culture of continuous learning and adaptability to teamwork. It insists to flexibility in offering elective courses on emerging technologies, providing access to online resources, and encouraging participation in workshops and conferences.

(6) Entrepreneurship and Innovation is a requirement for having courses on entrepreneurship, innovation management, and startup ecosystems. It should inspire students to think creatively and rationally, and take risks, allowing staying ahead of digital trends and implementing cutting-edge solutions.

(7) Analytical Thinking and Data-Driven Decision Making is an achievement in which not only coverage of data structures and algorithms with limited application to business contexts is enough. Instead, it requires having advanced courses in Data Analytics (DA), Machine Learning (ML), Business Intelligence (BI), decision theory, operational research, optimization methods, and stochastic processes. Emphasizing applications in strategic decision-making is expected.

(8) User Experience (UX) and Design Thinking requires integration of UX design, Human Computer Interaction (HCI), visualization methods, and design thinking courses. It provides emphasizing user-centric product development with full consideration of the Requirements Engineering (RE) discipline.

(9) Cybersecurity and Compliance requires fostering cybersecurity education and culture and including topics on regulatory compliance, risk management, and ethical considerations in digital management.

(10) Mentorship and Networking requires creation of various mentorship programs to connect students with industry professionals and alumni. It should provide guidance and career advice to the students.

In [22], critical skills gaps for DT are identified: Analytics, Big Data, Digital Strategy, Financial Modeling, Digital Marketing, Security, Privacy, Risk, Compliance, Smart Product Development, Software Development, and Web/mobile Development. We see those recognized skills for Digital Managers and Digital Engineers.

In [21], the author discusses new roles of Chief Information Officers (CIOs) in the DT process, by comparison to the past and well recognized CIOs' roles, as shown in TABLE I. It is evident that new roles of CIOs require much more proactivity, well understanding of both business and technology and applying strategic thinking effectively. As such, a new profile of a Digital Manager or even Digital Engineer should provide high-quality knowledge to cope with new roles of CIOs in DT, where CIOs take positions very close to strategic company management.

TABLE I.
CIO ROLES IN THE PAST AND IN DT FUTURE

Past CIO Roles	Future CIO Roles
Focus on IT outcomes	Focus on business outcomes
Sequencing of operations	Creating collaborative agenda
Support	Overcoming the unconquerable
Costs control	Developing the income plan
Process Reengineering	Data utilization
Outsourcing	Design of business
Focus on functions	Focus on (DT) platforms
Searching for parity	Searching for differentiation
In the scope of IT sector	Everywhere in the organization
Focus on IT risks	Focus on business risks

Finally, it worths to mention that a new notion of Digital Intelligence (DI) is proposed in [36] and [35]. By this, all employees, including managers, should develop DI as an ability to learn digital technologies, deal with digital technologies appropriately, read, decode and manipulate digital information, and acquire and apply new knowledge and skills connected to digital technologies to address insights and openness, and improve operational efficiency and outcomes.

V. A New Education Model

To cope with Big Data Crisis Causes (B) and (C) by applying DT in academic education, we propose introducing the profiles of Digital Manager and Digital Engineer. A body of knowledge of the study programs supporting the two profiles is to be based on four widely set-up scientific and engineering general disciplines (Fig. 1):

(1) Computer Science, Informatics, (Software) Engineering, and Data Science (CSIEDS);
(2) (Applied) Mathematics, Statistics, and Science (MSS);
(3) Business, Social Science, and Economics (BSSE); and
(4) Humanities and Soft Skills (HSS).

As such, a body of knowledge includes numerous disciplines. We give here a list of the disciplines grouped by the four general disciplines, not pretending to perceive it as a complete list of the knowledge being offered.

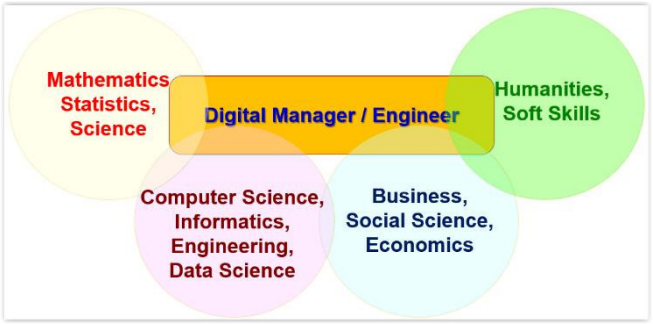


Fig. 1. General disciplines for digital manager and digital engineer

(1) Computer Science, Informatics, (Software) Engineering, and Data Science as a general discipline includes all core CSSEI disciplines, such as Programming, Computer Architecture, Operating Systems, Algorithms, and Compilers. Also, it covers Computational Intelligence and ML, HCI, Software Engineering, Databases, Information Systems, DevOps, and Cyber Security. Then it covers UX and Requirements Engineering, System Engineering, Decision Engineering, Formal Methods, Domain Specific Modeling and Languages, Model-Driven Engineering, Conceptual Modeling, and System Thinking, where general engineering disciplines are included.

(2) (Applied) Mathematics, Statistics, and Science should cover Algebra and Linear Algebra, Calculus, Discrete Mathematics, Graph Theory, Combinatorics, Logic, Probability and Statistics, Stochastic Processes, Operational Research, Optimization Methods, and Data Series Processing. Also, Physics or relevant science disciplines for a problem domain can be included.

(3) Business, Social Science, and Economics should cover Domain knowledge and Customer Experience (CX) with some selected focus areas across physical and social sciences of interest, Finance and Quantitative Finance, Financial Engineering, Marketing, Fundamentals of Economics, as well as Organization Design, Management, Privacy, Entrepreneurship, and Strategic Thinking.

(4) Humanities and Soft Skills should cover Communication, Critical Thinking, Adaptability, Problem-solving, Leadership, Creativity, and Innovation topics [22], [23].

Looking at the pretty long list of disciplines included in education of Digital Engineers and Digital Managers, it is obvious that one student cannot select and pass all those disciplines through B.Sc. and M.Sc. levels of education (five years in total). As such, following student's affinities and vision, some selection of all those disciplines must be done. We propose a flexible selection model, in which students can profile themselves, literally speaking in a 'Lego bricks' style. As such, some possible business roles of interest that can be derived from our general body of knowledge model can be:

a) The Digital Manager roles:
   - Digital Product Manager,
   - Digital Business & Finance Manager,
   - Digital Sales & Marketing Manager,
   - Digital UX / CX Engineer,
   - Digital Operations Engineer, and
   - Business Analyst,

b) The Digital Engineer roles:
   - ML Engineer,
   - Data Scientist,
   - Data Analyst,
   - System (Design) Engineer,
   - Data Engineer,
   - Information System Engineer,
   - Cyber Security Engineer,
   - DevOps Engineer, and
   - Software Engineer.

The list is not complete. By further practice of application of such flexible model, new roles may appear, or some of the roles being listed here can become obsolete.

Following the proposed body of knowledge we further propose a highly flexible education model and curricula structures that are based on the following principles.

(P1) Coverage of all four general curricula disciplines: MSS, CSIEDS, BSSE, and HSS, as mandatory disciplines. Students are forced to select courses from each of the mandatory disciplines.

(P2) The program structure is designed in a way that each general discipline is covered with a defined percentage of share in the whole structure. Thus, students are forced to select courses from the mandatory disciplines to a designed extent.

(P3) The coverage of all general disciplines is mandatory at all study levels (including B.Sc., M.Sc., Ph.D. studies), while percentage of share may vary slightly from one to the other level.

Regarding the percentages of share of the general curricula disciplines, we differentiate between the two model interpretations: (a) basic and (b) refined. In a basic interpretation, the share percentages of general disciplines are defined just for the common roles of Digital Manager and Digital Engineer. As such, the following principle should hold.

(P4a) In the basic model representation, the percentage of share of all four general disciplines is customized to the roles of Digital Manager and Digital Engineer. It is expected a relatively higher share of CSIEDS for Digital Engineer, while a share of BSSE is higher for Digital Manager.

As an illustration, for Digital Engineer we can define percentages of share as: (MSS: 20%; CSIEDS: 50%; BSSE: 15%; HSS: 15%), while for Digital Manager we can propose percentages of share as: (MSS: 20%; CSIEDS: 30%; BSSE: 35%; HSS: 15%).

In the refined interpretations, the share percentages of general disciplines are defined at the finer level of granularity, for specific Digital Manager and Digital Engineer roles. Even, study program designers have a freedom not just to refine a model interpretation to the specific roles, but also to the sub-disciplines of the four general disciplines. As such, the following principle should hold.

(P4b) In the refined model representation, the percentage of share of all four general disciplines is customized to the specific roles defined under the Digital Manager and Digital Engineer, as general roles. Further refinements are possible by decomposing the four general disciplines to the subdisciplines.

As an illustration, for Data Scientist, we can decompose the CSIEDS discipline to Computer Science, Informatics, and Engineering (CSIE) and Data Science (DS). Also, MSS can be decomposed to Mathematics and Science (MS) and Statistics (S). Then we can define percentages of shares as: (MS: 10%; S: 15%; CSIE: 30%; DS: 15%; BSSE: 15%; HSS: 15%). For Data Engineer, the distribution of percentages will be different in favor to CSIE: (MS: 10%; S: 10%; CSIE: 50%; DS: 10%; BSSE: 10%; HSS: 10%). On the contrary, for Business Analyst, the distribution of shares will be in favor to BSSE: (MS: 10%; S: 15%; CSIE: 15%; DS: 15%; BSSE: 30%; HSS: 15%).

(P5) MSS, CSIEDS, BSSE, and HSS disciplines are mandatory rather than sole courses. A proposal is to have very few mandatory courses, e.g. max. 1 per each semester. In this way, students are rather free about what courses to select from each mandatory discipline.

(P6) A study program structure should provide a selection of a great number of courses in each of the four mandatory general disciplines.

(P7) Each course is to be classified according to the level of rigor and defined prerequisites, as follows: (i) Basic; (ii) Intermediate; and (iii) Advanced. There are no prerequisites for basic courses. Prerequisites for each intermediate and advanced course are clearly defined, however not always with sharp requirements.

(P8) Students are advised to select courses from all three levels (basic, intermediate, advanced) to a designed extent, while there is a minimum percentage share of the courses of all three levels. The percentage share of the course levels is different for various study levels (B.Sc., M.Sc., and Ph.D. level).

As an example, for B.Sc. study level one can define a minimum of 20% of basic courses, 30% of intermediate courses, and 20% of advanced courses. For M.Sc. study level, a percentage of share may significantly vary with respect to the conformance of the previous level of studies completed by a student. As a rule, at the M.Sc. study level a study program designer can expect a predominant selection of advanced courses. However, if a conformance of the previously completed B.Sc. level of study is not very strong with the M.Sc. level being enrolled, some basic or intermediate courses are important to be included in the student's study plan.

(P9) Internship is a mandatory slot in any study program of the B.Sc. and M.Sc. levels. Study visit of at least one month to some other foreign university is a mandatory requirement in any study program of the Ph.D. level.

(P10) The study framework should provide a formal possibility of including distinguished industry experts in the lecturing process, at the selected advanced courses and to the designed extent.

Following all the principles, and particularly (P4a), (P4b), and (P8), it is possible to come to a 'default' profiling of students, by creating series of matrices of the Business Roles / Study Program Disciplines type. The matrix cells can be organized in various ways:

– recommended percentages (or even numbers) of basic, intermediate, and advanced courses;
– recommended percentages of share (or numbers of courses) of each of the four general disciplines;
– recommended percentages of share (or numbers of courses) of individual subdisciplines derived by the decomposition of the four general disciplines; and
– recommended individual courses that are appropriate for a business role.

A study model which is in line with the principles presented in this paper can be perceived as highly flexible. Such flexible model requires a possibility of running some massive courses both in Winter and Summer semesters and working with students in smaller groups, to motivate collaborative work and project-based learning. Thus, such model requires engagement of larger number of education staff. Both education staff and students are to be highly motivated for their involvement in such study model. By this, significant organizational and social efforts are to be invested in raising the level of motivation and awareness of the students and teachers about the values coming from the implementation of such model.

A significant difference between the proposed model and 'conventional' models is that the students take very strong responsibility for their future profiling in this model. In 'conventional' models, students take a more passive role just having freedom to select and enroll in a study program and then selecting some elective courses. Here, students must be quite proactive and strongly motivated to think about the details of their professional future early in advance. Therefore, proactive mentorship of students in their profiling in this model is of crucial importance, bearing in mind that mentor will just give advice, while students will make final decisions, as they should be the most responsible for themselves. We see such approach as an effective way to create proactive, satisfied, motivated, brave, and highly innovative experts ready to cope with the issues of modern business in DE. One evident psychological issue in the implementation of such approach is that the teachers predominantly believe that they know best what an appropriate selection of topics for students is, as the students "are mostly not enough mature to know it in advance". Instead, we insist here that the students of a higher level of responsibility must know what the best selection of topics for them is.

In the proposed model, students have much wider freedom in a selection of individual courses than in 'conventional' models. On one hand, it creates a high degree of freedom, and the university management may complain about realistic execution of such study programs. On the other hand, we believe that in the first two school years of execution, typical clusters

of selected courses will be created naturally, and it will reduce the degree of freedom to a rational level, without forcing students to select courses they do not want to select and thus reducing their level of freedom.

In general, development, implementation, and execution of such study models is more expensive than of 'conventional' models. It is always a question of how to involve relevant stakeholders who are ready to support such models in practice, and it could be a matter of further research.

VI. Conclusion

Addressing Big Data Crisis Causes and challenges of DT is a strategic and long-life task. To address Digital Talent Gap phenomenon, a DT approach is necessary to open new perspectives for academic education. Facing the era of DE requires new profiles of Digital Manager and Digital Engineer and producing a huge number of digital experts ready to cope with DE challenges. In some segments, formal academic education for digital experts is to be significantly transformed to the level of a disruptive modification. A new and more effective form of academic education is expected to be in line with the DT principles, with clearly defined education and business strategy. It should be designed as a flexible framework providing reasonable combination with high quality alternative forms of education. Such study programs must present a clear multidisciplinary and interdisciplinary orientation. They are expected to be implemented through highly flexible education models and curricula structures to guarantee freedom of choice to students, raising their level of responsibility, vision, motivation and understanding in their profiling, and adopting both fundamental and cutting-edge knowledge, necessary for an effective participation in the DE processes in the upcoming decades.

In this paper, we have proposed general principles on how to come to the new model of academic education for DT and DE, as a prerequisite for strengthening human resources and proper addressing Digital Talent Gap phenomenon. Creating new academic educational models requires formulation of a clear vision and strategy. As a rule, such education models are more expensive, first to develop and implement, and then to execute, for several reasons. We believe that the process of creating such new education models is painful, and takes a lot of courage, particularly as academic community is rather conservative and rigid. However, societies and economies that are more proactive and show more bravery in realizing what are the real advantages, as well as tangible and intangible benefits of such transformation, have much better chances to open new strategic perspectives of sustainable and long-term society development. For emerging economies, it could be an excellent opportunity to reach the level of well-developed economies more efficiently. Otherwise, missing such opportunity means further prolonging the development delays.

Acknowledgment

This work was supported by the University of Belgrade – Faculty of Organizational Sciences.

References

[1] Chiang, R.H.L., Goes, P., Stohr, E.A., "Business intelligence and analytics education, and program development: a unique opportunity for the information systems discipline". *ACM Trans. Manag. Inf. Syst*, 2012, 3(3).

[2] McKinsey Global Institute: *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey, 2011.

[3] Tulasi, B., "Significance of big data and analytics in higher education", *Int. J. Comput. Appl.*, 2013, 68(14), 21–23.

[4] Gartner (Analysts: Genovese, Y., Prentice, S.), *Pattern-Based Strategy: Getting Value From Big Data*, ID: G00214032, 2011.

[5] *Survey on Stakeholders in Serbia – A report in the Advanced Data Analytics in Business* (ADA) ERASMUS+ project No. 598829-EPP-1-2018-1-RS-EPPKA2-CBHE-JP.

[6] Halim, E., Sie, D. S., Hansel, G., Gui, A., Syamsuar, D., Teoh, A. P., "Revamping Curriculum Designs in Digital Talent Accelerators for the Modern Technological Landscape". *2024 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics* (RI2C), 2024, pp. 1–6. IEEE.

[7] Civilcharran, S., Maharaj, M. S., "A framework to determine the digital skills preparedness of graduates for industry", In: *Proceedings of 2018 International Conference on Intelligent and Innovative Computing Applications* (ICONIC), 2018, pp. 1–6. IEEE.

[8] McCarthy, A. M., Maor, D., McConney, A., Cavanaugh, C., "Digital transformation in education: Critical components for leaders of system change". *Social Sciences & Humanities Open*, 2023, 8(1), 100479.

[9] Uzule, K., Gobniece, J. Z., "Digital transformation of economies through technology, education and competences", *Problems of Education in the 21st Century*, 2024, 82(5), 758.

[10] Gulyamov, S. S., Ruziev, R., Suyunova, D., "Establishing Specialized Research and Education Centers as an Imperative for Training Next-Generation IT Specialists", In: *Proceedings of 2024 4th International Conference on Technology Enhanced Learning in Higher Education* (TELE), 2024, pp. 354–356. IEEE.

[11] Kopackova, H., Simonova, S., Reimannova, I., "Digital transformation leaders wanted: How to prepare students for the ever-changing demands of the labor market", *The International Journal of Management Education*, 2024, 22(1), 100943.

[12] Torbaghan, M. E., Sasidharan, M., Jefferson, I., Watkins, J., "Preparing students for a digitized future", *IEEE Transactions on Education*, 2022, 66(1), 20–29.

[13] Wolff, C., Omar, A., Shildibekov, Y., "How will we build competences for managing the digital transformation?", 2019, In: *Proceedings of 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (IDAACS), 2019, Vol. 2, pp. 1122–1129. IEEE.

[14] Demchenko, Y., Degeler, V., Oprescu, A., Brewer, S., "Professional and 21st century skills for data driven digital economy", In: *Proceedings of 2023 IEEE Global Engineering Education Conference* (EDUCON), 2023, pp. 1–8. IEEE.

[15] Goulart, V. G., Liboni, L. B., Cezarino, L. O., "Balancing skills in the digital transformation era: The future of jobs and the role of higher education", *Industry and Higher Education*, 2022, 36(2), 118–127.

[16] Glavaš, J., Uroda, I., Mandić, B., "Students' perception of studying in digital environment and preparedness for workplaces in digital economy-current state and perspectives", 2019, In: *Proceedings of 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics* (MIPRO), 2019, pp. 1391–1395. IEEE.

[17] Panxhi, M., Thanasi-Boçe, M., Hoxha, J., "Digital transformation, entrepreneurial competence, and future job prospects in engineering education", *Cogent Education*, 2025, 12(1), 2526429.

[18] Luković, I., "Issues and Lessons Learned in the Development of Academic Study Programs in Data Science", *21st International Conference DAMDID 2019*, In: *Proceedings*, *Springer Nature, CCIS 1223*, 2020, DOI: 10.1007/978-3-030-51913-1_15.

[19] Yao, Y., Qi, P., Zhu, Y., "Research on Interdisciplinary Education in Digital Economy", *Advances in Social Science, Education and Humanities Research*, 2019, Vol. 376, DOI: 10.2991/ssch d-19.2019.32.

[20] Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D., Buckley, N., *Strategy, not Technology, Drives Digital Transformation*, Research Report by MIT Sloan Management Review and Deloitte University Press, 2015.

[21] Kočović, P., *Serbian CIO Agenda*. Union University "Nikola Tesla", Faculty for Information Technologies and Engineering, 2019, ISBN 978-86-81400-16-6.

[22] Synergy Consulting, Malaki M., *1245-1330 Session 04: The Digital Talent Gap – Looking at leaders and laggards based on a global survey of enterprise workforce digital skills readiness*, 2019, URL [available in August 2025]: https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/Pages/Events/2019/jul-iran-dtx/Workshop-on-%E2%80%9CDigital-Transformation-in-Digital-Economy%E2%80%9D.aspx
[23] Capgemini Consulting, *The Digital Talent Gap Developing Skills for Today's Digital Organizations*, 2013, URL [available in August 2025]: https://www.scribd.com/document/340823943/The-digital-talent-gap-Capgemini-pdf.
[24] Azoev, G., Aleshnikova, V.I., Sumarokova, E.V., "Benchmarking of the educational process in the field of digital economy in Russia", *In Proceedings of the 1st International Scientific Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth"* (MTDE 2019), 2019,.DOI: 10.2991/mtde-19.2019.144.
[25] Kovalev M. M., "Education for the Digital Economy (Образование для цифровой экономики)", *Cifrovaja transformacija* [*Digital transformation*], 2018, 1 (2), pp. 37–42.
[26] Nambisan, S., Lyytinen, K., Majchrzak, A., Song, M., "Digital Innovation Management: Reinventing Innovation Management Research in a Digital World". *MIS Quarterly*, 2017, 41(1), 223-238.
[27] Heikkinen, K., Räisänen, T., "Role of Multidisciplinary and Interdisciplinary Education in Computer Science: A Literature Review", *Managing Global Transitions*, 2018, DOI: 10.26493/1854-6935.16.159-172, 16(2), 159-172.
[28] Helle, L., Tynjälä, P., Olkinuora, E., "Project-based learning in post-secondary education – theory, practice and rubber sling shots", *Higher Education*, 2006, 51(2), 287-314.
[29] Robles, M. M., "Executive Perceptions of the Top 10 Soft Skills Needed in Today's Workplace", *Business Communication Quarterly*, 2012, 75(4), 453-465, DOI: 10.1177/1080569912460400.
[30] Keegan, A., Turner, J. R., "Quantity versus quality in project-based learning practices", *International Journal of Project Management*, 2001, 19(4), 255-264, DOI:10.1177/1350507601321006.
[31] McCrohan, K. F., Engel, K., Harvey, J. W., "Influence of Awareness and Training on Cybersecurity", *Journal of Internet Commerce*, 2010, 9(1), 23-41, DOI: 10.1080/15332861.2010.487415.
[32]] Klus, M. F., Müller, J., "Identifying Leadership Skills Required in the Digital Age", *CESifo* Working Paper No. 8180*,* 2020, DOI: 10.2139/ssrn.3564861.
[33] Klus, M. F., Müller, J., "The digital leader: what one needs to master today's organisational challenges", *Journal of Business Economics*, 2021, 91, 1189 - 1223.
[34] Kluz, A., Firley, M, *How To Be a Leader in The Digital Age*, URL [available in August 2025]: https://www.weforum.org/agenda/2016/05/how-to-be-a-leader-in-the-digital-age/, 2016.
[35] Henderikx M., Stoffers J., "An Exploratory Literature Study into Digital Transformation and Leadership: Toward Future-Proof Middle Managers", *Sustainability*, 2022, 14(2):687, DOI: 10.3390/su14020687.
[36] Boughzala, I., Garmaki, M., Chourabi, O., "Understanding How Digital Intelligence Contributes to Digital Creativity and Digital Transformation: A Systematic Literature Review", In: *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Maui, HI, USA, 7–10 January 2020, pp. 320–329, URL: https://hdl.handle.net/10125/63779.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 35–42
DOI: 10.15439/2025F1820
ISSN 2300-5963 ACSIS, Vol. 43

Q-ID: A Reinforcement Learning Framework for Adaptive Intrusion Detection

Maisha Maliha and Mohammed Atiquzzaman
School of Computer Science
University of Oklahoma
Norman, Oklahoma, USA
Email: maisha.maliha-1@ou.edu, atiq@ou.edu

***Abstract*—The growing sophistication and frequency of cyber threats in communication networks demand Intrusion Detection Systems (IDS) that adapt to evolving attack patterns. Traditional approaches, based on static rules or purely supervised models, often fail to recognize novel attacks, leaving critical infrastructures exposed. Reinforcement Learning (RL) provides a dynamic alternative by enabling agents to refine detection policies through continuous feedback. In this work, we propose a Q-learning–based Intrusion Detection (Q-ID) system and train it on the CICIDS2017 dataset. The RL formulation defines the state as the flow's feature vector, the action as the classification decision, and the reward as $+1$ for correct predictions and -1 otherwise. To ensure stable convergence, the reward is integrated with cross-entropy loss in a hybrid objective, allowing continued improvement even after the supervised component has plateaued. Unlike prior IDS methods that rely solely on offline supervised training, our approach fuses reinforcement feedback with supervised optimization to support adaptive and robust detection. Experimental results, conducted under class imbalance and realistic evaluation splits, show that the proposed system achieves 99.3% accuracy, outperforming strong baselines including deep neural networks and traditional classifiers. Moreover, the RL agent demonstrates robustness under skewed traffic distributions and adaptability to previously unseen attack types. These results highlight reinforcement learning as a promising paradigm for building resilient IDS in critical communication environments.**

***Index Terms*—Adaptive Intrusion Detection, Communication Networks, Reinforcement Learning, Q-learning Algorithm, Network Security, Cyber Attack Resilience**

I. Introduction

THE rapid growth in the complexity and frequency of cyber threats poses significant risks to modern communication networks. These infrastructures form the backbone of critical services, enterprise systems, and national defense, making them attractive targets for adversaries. Successful intrusions can disrupt operations, compromise sensitive information, and trigger cascading failures across interconnected systems [1]. Traditional Intrusion Detection Systems (IDS), which depend heavily on static rules or signature-based techniques, are increasingly inadequate for identifying sophisticated or previously unseen attacks [2]. Their inability to adapt to dynamic traffic patterns and evolving adversarial strategies underscores the need for intelligent and flexible detection mechanisms.

Recent advances in Machine Learning (ML) and Artificial Intelligence (AI) have been applied to strengthen IDS. Supervised learning approaches, for example, can classify traffic as benign or malicious by learning from labeled datasets. While such models outperform static rule-based systems, they suffer from critical limitations, including reliance on large annotated datasets, assumption of stationary data distributions, and vulnerability to novel or evolving threats. These weaknesses restrict their utility in dynamic and high-stakes communication environments.

Reinforcement Learning (RL), a branch of ML, offers a promising solution. Unlike supervised learning methods that passively rely on historical labels, RL enables an agent to interact with its environment, making sequential decisions that maximize cumulative rewards [3]. This dynamic learning paradigm equips IDS with the ability to self-correct and adapt as new threats emerge, making RL particularly well-suited to adversarial and continuously changing domains.

This work investigates the application of RL for adaptive intrusion detection using the widely adopted CICIDS2017 [4] dataset. Specifically, we propose a Q-learning–based Intrusion Detection (Q-ID) agent that classifies network flows as normal or malicious. The formulation defines the state, action, and reward explicitly, and employs a hybrid training objective that combines reward feedback with cross-entropy loss to ensure stable convergence. Through continuous interaction with the environment, the agent adapts its detection strategies over time, enhancing its ability to recognize both known and previously unseen attacks. The contributions of this paper are threefold.

1) An explicit RL formulation for intrusion detection, including clear definitions of states, actions, and rewards.
2) A hybrid training strategy that integrates supervised and reinforcement signals for stable and continued learning.
3) An empirical evaluation against strong machine learning baselines that demonstrates superior accuracy, robustness under class imbalance, and practical feasibility for deployment in critical communication networks.

The remainder of this paper is structured as follows: we first discuss related work in the field of intrusion detection and RL in Section II. Section III describes the CICIDS2017 dataset and the preprocessing pipeline used for our experiments. We then describe our methodology, data preprocessing and the design of the RL agent in Section IV. Experimental results are presented to demonstrate the effectiveness of our approach in Section V. Section VI provides an ablation study to quantify the contribution of key design choices. Finally, we conclude with a discussion on the implications of our findings and potential future directions for research in Section VII.

II. Related Work

Intrusion detection systems (IDS) have evolved from static signature-based and rule-driven techniques toward more adaptive approaches. Classical IDS methods depend on predefined patterns of known attacks, which limits their ability to recognize novel or sophisticated intrusions. As cyber threats have become increasingly dynamic, research has shifted toward machine learning (ML) and artificial intelligence (AI) techniques capable of adapting to changing attack landscapes.

Reinforcement learning (RL) has emerged as a promising direction for IDS. Otoum et al. [5] proposed a big data–driven RL-based IDS for wireless sensor networks, achieving near-perfect detection accuracy and outperforming prior hybrid approaches. Their work highlighted RL's ability to adapt to evolving attack scenarios. However, existing RL-based IDS approaches often do not clearly define the state, action, and reward components, and few explicitly consider integrating RL signals with supervised learning objectives to stabilize optimization.

Several studies have applied supervised ML models to intrusion detection. Maliha et al. [6] investigated IoT security using algorithms such as K-Nearest Neighbor (KNN), Naive Bayes, Support Vector Machine (SVM), Random Forest, and Decision Tree on the CICIDS2017 dataset, with feature selection based on Random Forest Regressor and Extra Trees Classifier. Their analysis showed that KNN achieved the highest accuracy and F1-score among the evaluated models. Similarly, Choudhary et al. [7] developed a deep neural network (DNN) framework and evaluated it across KDDCUP'99, NSL-KDD, and UNSW-NB15, reporting an average accuracy of 91.5%. Norwahidayah et al. [8] combined particle swarm optimization (PSO) [9] for feature selection with an artificial neural network (ANN), achieving 98% accuracy on the KDDCUP'99 dataset. While these studies demonstrate the value of supervised learning, their effectiveness remains tied to the availability of labeled data and known attack types, reducing generalization to unseen threats.

Neural networks have also been widely applied to anomaly detection tasks such as user profiling [10], command sequence prediction [11], and traffic pattern recognition [12]. More advanced models, including recurrent neural networks and self-organizing maps [13], offer flexibility but often lack transparency in their decision-making, limiting trust and interpretability in security-critical environments.

More recent efforts have explored reinforcement learning-based intrusion detection systems in dynamic and adversarial settings. Ghubaish et al. [14] presented HDRL-IDS, an actor–critic hybrid RL model designed for medical IoT security in 5G environments. Although tailored for low-latency MEC applications, their approach does not benchmark against modern tabular learners or operate on standard datasets such as CICIDS2017. Finally, Mahjoub et al. [15] proposed an RL-driven IDS that adapts both the policy and environment for IoT attack detection using the Bot-IoT dataset. However, the method lacks architectural transparency and does not address class imbalance. In contrast, our proposed Q-ID system introduces a well-defined Q-learning architecture with a hybrid reward–supervised loss, applies to a general-purpose IDS dataset (CICIDS2017), and demonstrates robustness under data skew and previously unseen attacks.

Compared to prior research, the present study contributes by explicitly formulating intrusion detection as a reinforcement learning problem. States are represented by network flow features, actions correspond to classification decisions, and rewards provide direct feedback on detection outcomes. In addition, a hybrid training objective that combines cross-entropy loss with reward-based feedback is introduced to enhance stability and adaptability. This positions our approach at the intersection of supervised learning and RL, aiming to provide robust detection performance while maintaining the adaptability needed for evolving cyber threats.

III. Dataset and Preprocessing

In this study, we utilize the CICIDS 2017 dataset, which is a comprehensive benchmark for IDS. This dataset encompasses a wide range of network traffic data, including both normal and various attack types, making it suitable for training and evaluating IDSs.

A. Dataset

The CICIDS 2017 dataset contains 2,830,683 records stored in a single CSV file. Among these, 2,359,289 instances correspond to benign (Normal) traffic, while 231,073 records represent DoS Hulk attacks. Additional attack categories include 158,930 PortScan, 41,835 DDoS, 10,293 DoS GoldenEye, 7,938 FTP-Patator, 5,897 SSH-Patator, 5,796 DoS Slowloris, 5,499 DoS Slowhttptest, 1,966 Bot, 1,507 Web Attack–Brute Force, 652 Web Attack–Cross-Site Scripting, 36 Infiltration, 21 Web Attack–SQL Injection, and 11 Heartbleed instances, as shown in Figure 1.

Overall, 83% of the dataset is classified as benign traffic, while the remaining 17% corresponds to various attack types, indicating a significant class imbalance. To ensure reliable model training, preprocessing is performed to eliminate errors and redundancies. During this step, inconsistencies are detected among the 2,830,751 data streams and are subsequently removed. Additional preprocessing involves eliminating redundant features and converting categorical variables into numerical representations using the `LabelEncoder` function. Specific corrections include replacing "infinity" values with

-1, handling "NaN" values by replacing them with 0, and correcting inconsistent labels (e.g., converting "FTP–Patator" to "FTP-Patator").

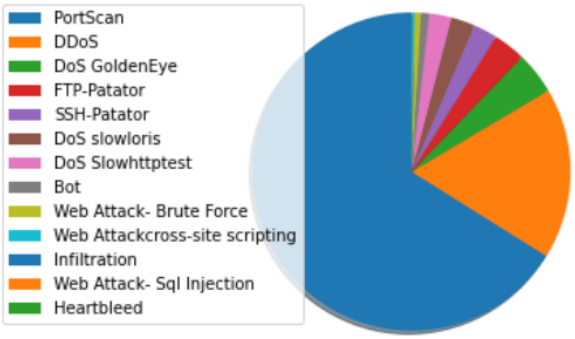


Fig. 1. Distribution of classes in the CICIDS2017 dataset.

B. Feature Information

The dataset comprises 85 attributes that describe network flows. These include identifiers such as flow ID, source IP, destination IP, destination port, protocol, and timestamp. Several attributes capture flow-level statistics, such as flow duration (in microseconds), total forward packets, and total backward packets. Additional attributes summarize packet size distributions, including maximum, minimum, mean, and standard deviation values for both forward and backward packets.

Throughput-related features include flow bytes and flow packets, which measure the total bytes and packets transmitted per second. Inter-arrival time (IAT) features further describe temporal dynamics: Flow IAT Max, Flow IAT Min, Flow IAT Std, and Flow IAT Mean summarize inter-arrival time distributions, while Fwd IAT Total, Fwd IAT Mean, Fwd IAT Std, Fwd IAT Max, and Fwd IAT Min provide corresponding statistics for forward packets. Similar sets of features capture backward packet timing.

Protocol-specific flags are also represented. For instance, the Fwd PSH Flags feature counts packets with the PUSH flag set in the forward direction, indicating that data should be processed immediately. Likewise, the Fwd URG Flags and Bwd URG Flags capture instances where the URG flag is set, signaling urgent packet handling. Other features include Packet Length Mean, Packet Length Std, and Packet Length Variance, which characterize the distribution of packet sizes. Congestion-related flags are also tracked, including CWR Flag Count (Congestion Window Reduced) and ECE Flag Count (Explicit Congestion Notification Echo).

Additional attributes capture traffic behavior such as the download/upload ratio per packet, the average packet size, the label, and the external IP address. Several TCP-specific features are also included, such as the initial window sizes (forward and backward), the minimum TCP segment size, and the count of forward packets containing at least one byte of data. Temporal activity features describe active and idle periods of flows, with mean, standard deviation, maximum, and minimum statistics characterizing time before a flow becomes idle and transitions back to active states.

C. Feature Selection

Given the large number of attributes, feature selection is performed to identify the most informative features and reduce redundancy. This step is essential for building accurate and computationally efficient models. A Random Forest Regressor is employed to estimate feature importance by constructing multiple decision trees on subsampled data and evaluating classification performance across them. Features with higher importance scores are retained for model training, while low-importance features are discarded.

The ranking reveals that the most influential features included Bwd Packet Length Std, Flow Bytes/s, and Fwd Packet Length Std, with importance scores of 0.247, 0.178, and 0.112, respectively. By focusing on these high-weight attributes, the preprocessing pipeline enhances both model accuracy and efficiency, ensuring that the intrusion detection system learns from the most discriminative aspects of network traffic.

IV. Proposed Method

In this work, we develop Q-ID (Q-learning–based Intrusion Detection), an adaptive intrusion detection system (IDS) for communication networks that leverages RL to strengthen cyber defense. The objective is real-time identification of malicious traffic and timely decision support for mitigation, with resilient performance in harsh, contested, and resource-constrained environments. We train and evaluate the approach on the CICIDS2017 dataset, which offers diverse, labeled benign and attack traffic. Our proposed Q-ID employs an RL-augmented training objective alongside supervised learning to improve robustness under class imbalance and enhance generalization to evolving threats.

A. Concept of Deep Q-Networks

In the Deep Q-Network (DQN) framework [16], at each discrete time step t, an agent observes the current state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to a (possibly stochastic) policy $\pi(a \mid s)$, receives a scalar reward r_t, and transitions to a next state s_{t+1} in a fully observable, single-agent reinforcement-learning environment. The (discounted) return [17] from time t is

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \tag{1}$$

where $\gamma \in [0, 1]$ is the discount factor. The objective is to maximize the expected return. The action-value function (Q-function) under policy π is

$$Q^{\pi}(s, a) = \mathbb{E}[R_t \,|\, s_t = s,\, a_t = a]\,, \tag{2}$$

i.e., the expected return obtained by taking action a in state s and thereafter following π.

The optimal Q-function $Q^*(s,a) = \max_\pi Q^\pi(s,a)$ satisfies the Bellman optimality equation [18]:

$$Q^*(s,a) \;=\; \mathbb{E}\Big[r_t + \gamma \max_{a' \in \mathcal{A}} Q^*(s_{t+1}, a') \;\Big|\; s_t = s,\; a_t = a\Big]. \tag{3}$$

Tabulating Q^* over all state–action pairs is generally intractable, so DQN approximates Q with a neural network $Q(s,a;\theta)$. The parameters θ are learned by minimizing the mean-squared temporal-difference (TD) error over transitions (s,a,r,s') drawn from a behavior distribution ρ (e.g., a replay buffer):

$$L_i(\theta_i) \;=\; \mathbb{E}_{(s,a,r,s')\sim\rho}\Big[\big(y_i - Q(s,a;\theta_i)\big)^2\Big], \tag{4}$$

where the TD target is

$$y_i \;=\; r \;+\; \gamma \max_{a' \in \mathcal{A}} Q\big(s', a';\, \theta_i^-\big)\,. \tag{5}$$

Here, θ_i^- denotes the (lagged) target-network parameters, which are updated periodically to stabilize training. Because Q-learning is off-policy, the agent can learn the greedy policy $\pi_{\text{greedy}}(s) = \arg\max_{a\in\mathcal{A}} Q(s,a;\theta)$ while collecting data with a separate behavior policy, commonly ϵ-greedy: with probability $1-\epsilon$ it selects the greedy action and with probability ϵ it selects a random action to ensure sufficient exploration of the state–action space.

B. Implementation Using a Deep Reinforcement Learning Network

The training data are severely imbalanced across classes, which can bias purely supervised learners toward majority classes and degrade detection of rare attacks. To counteract this, we augment standard cross-entropy training with a reinforcement-learning (RL) signal that directly rewards correct decisions. This hybrid objective encourages both calibrated probabilities and value estimates that are consistent with decision quality.

a) Problem formulation: Let the network-flow feature vector be the state $s \in \mathbb{R}^d$ and the predicted traffic label be the action $a \in \mathcal{A}$ (including the *benign* class). Given ground-truth $y \in \mathcal{A}$, we define a bandit-style reward

$$r(s,a) \;=\; \mathbb{1}\{a = y\} \in \{0,1\}. \tag{6}$$

A Q-network $Q(s,a;\theta)$ maps a state to real-valued action scores. For calibrated class probabilities used by the supervised term, we apply a softmax to these scores:

$$\pi_\theta(a \mid s) \;=\; \frac{\exp(Q(s,a;\theta))}{\sum_{a'\in\mathcal{A}} \exp(Q(s,a';\theta))}. \tag{7}$$

b) Hybrid learning objective: We combine (i) supervised cross-entropy with (ii) a temporal-difference (TD) loss from Q-learning. The supervised loss is

$$\mathcal{L}_{\text{sup}}(\theta) \;=\; -\log \pi_\theta(y \mid s). \tag{8}$$

Because training proceeds from a static labeled corpus, we adopt an *offline contextual bandit* view and set $\gamma = 0$ for the Bellman target by default; optionally, a small $\gamma \in (0,1)$ can be used by bootstrapping onto the next sample s' in a minibatch. The TD target and TD loss are

$$y^{\text{TD}} \;=\; r(s,a) \;+\; \gamma \max_{a'\in\mathcal{A}} Q(s', a'; \theta^-), \tag{9}$$

$$\mathcal{L}_{\text{TD}}(\theta) \;=\; \big(y^{\text{TD}} - Q(s,a;\theta)\big)^2, \tag{10}$$

where θ^- are lagged target-network parameters updated periodically to stabilize training. The total loss is the weighted sum

$$\mathcal{L}_{\text{total}}(\theta) \;=\; \mathcal{L}_{\text{sup}}(\theta) \;+\; \lambda\, \mathcal{L}_{\text{TD}}(\theta), \qquad \lambda \geq 0. \tag{11}$$

Minimizing Eqn. (11) yields class probabilities (via Eqn. (8)) and value estimates that satisfy the Bellman consistency implied by Eqn. (9). Importantly, we never interpret softmax probabilities as Q-values; raw network outputs serve as Q-values, while their softmax only supports the supervised term.

c) Training procedure: We train with mini-batch stochastic gradient descent over the dataset (serving as an experience buffer). Exploration is emulated during training via an ϵ-greedy behavior policy with respect to Q; this helps decorrelate targets and mitigates overfitting to majority classes.

The algorithm of our proposed Q-ID is provided in Algorithm 1.

Algorithm 1 Hybrid Supervised–RL Training for Intrusion Detection

Input: Labeled dataset $\mathcal{D} = \{(s,y)\}$, action set $\mathcal{A}$, discount $\gamma \in [0,1)$ (default 0), TD weight $\lambda \geq 0$, exploration rate $\epsilon \in [0,1]$, minibatch size B, target-update period K, learning rate η
Output: Trained parameters θ; inference uses $\hat{a}(s) = \arg\max_{a\in\mathcal{A}} Q(s,a;\theta)$
Initialize Q-network parameters θ; set target parameters $\theta^- \leftarrow \theta$
while not converged **do**
  Sample minibatch $\{(s_i, y_i)\}_{i=1}^B \sim \mathcal{D}$
  for $i = 1$ to B **do** ▷ ϵ-greedy behavior and reward
    With prob. $1-\epsilon$: $a_i \leftarrow \arg\max_{a\in\mathcal{A}} Q(s_i, a; \theta)$; else sample $a_i \sim \text{Uniform}(\mathcal{A})$
    $r_i \leftarrow \mathbf{1}\{a_i = y_i\}$ ▷ Bandit reward from label
    (Optional) choose s_i' as a valid successor state; otherwise set $\gamma \leftarrow 0$
    $y_i^{\text{TD}} \leftarrow r_i + \gamma \max_{a'\in\mathcal{A}} Q(s_i', a'; \theta^-)$ ▷ TD target
  $\mathcal{L}_{\text{sup}}(\theta) \leftarrow -\dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \log \frac{\exp(Q(s_i, y_i; \theta))}{\sum_{a\in\mathcal{A}} \exp(Q(s_i, a; \theta))}$ ▷ Cross-entropy on true class
  $\mathcal{L}_{\text{TD}}(\theta) \leftarrow \dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \big(y_i^{\text{TD}} - Q(s_i, a_i; \theta)\big)^2$ ▷ Mean-squared TD error
  $\mathcal{L}_{\text{total}}(\theta) \leftarrow \mathcal{L}_{\text{sup}}(\theta) + \lambda\, \mathcal{L}_{\text{TD}}(\theta)$
  $\theta \leftarrow \theta - \eta\, \nabla_\theta \mathcal{L}_{\text{total}}(\theta)$ ▷ Optimizer step (e.g., Adam)
  if iteration $\bmod K = 0$ **then**
    $\theta^- \leftarrow \theta$ ▷ Target-network update

d) Correctness and stability considerations:

- *Well-posed loss:* The TD loss (Eqn. (10)) is defined for all real-valued Q. We avoid nonstandard terms such as $-\log R$ that become ill-defined when the return is zero.
- *Offline setting:* With static data and no true environment transitions, $\gamma = 0$ yields a principled contextual-bandit reduction; small $\gamma > 0$ can be used cautiously with bootstrapped s'.
- *Off-policy learning:* The ϵ-greedy behavior policy ensures that Q-learning's off-policy assumption holds during training (data collection is synthetic but diversified).
- *Imbalance robustness:* Class weighting or focal variants of Eqn. (8) can be incorporated without altering the TD component. Calibration can be monitored on a validation split.
- *Stopping criterion:* We stop on validation metrics (e.g., macro-F1/ROC-AUC) rather than requiring the empirical return to reach 1, which may be unattainable on challenging splits.

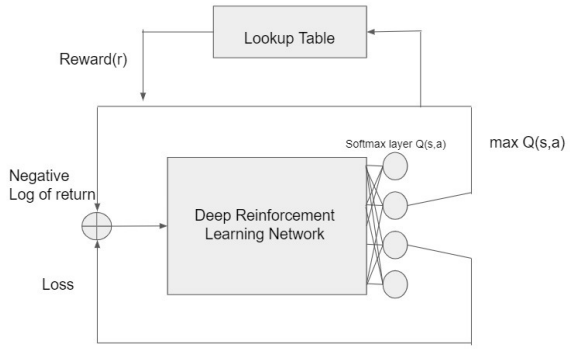


Fig. 2. End-to-end training and evaluation pipeline for the hybrid supervised+RL IDS.

C. Deep Neural Network Used in the RL Module

a) Architecture: The Q-network maps $s \in \mathbb{R}^d$ to $Q(s,\cdot;\theta) \in \mathbb{R}^{|\mathcal{A}|}$ through five fully connected (FC) layers with a lightweight gating-and-residual pathway:

$$h_1 = \mathrm{ReLU}(W_1 s + b_1), \quad W_1 \in \mathbb{R}^{128\times d}, \tag{12}$$

$$h_2 = \mathrm{ReLU}(W_2 h_1 + b_2), \quad W_2 \in \mathbb{R}^{128\times 128}, \tag{13}$$

$$g = W_3 h_2 + b_3, \quad W_3 \in \mathbb{R}^{d\times 128}, \tag{14}$$

$$\tilde{s} = s + (g \odot s), \quad \text{(element-wise gate } \odot\text{)}, \tag{15}$$

$$h_4 = \mathrm{ReLU}(W_4 \tilde{s} + b_4), \quad W_4 \in \mathbb{R}^{128\times d}, \tag{16}$$

$$Q(s,\cdot;\theta) = W_5 h_4 + b_5, \quad W_5 \in \mathbb{R}^{|\mathcal{A}|\times 128}. \tag{17}$$

Here, $\theta = \{W_\ell, b_\ell\}_{\ell=1}^{5}$ are trainable parameters. The gating term $g \odot s$ adaptively re-weights input dimensions before a residual addition $\tilde{s} = s + (g \odot s)$, which empirically improves gradient flow and allows the model to emphasize discriminative flow features. The softmax in Eqn. (7) is applied only for the supervised term Eqn. (8); the raw outputs Eqn. (17) are used as Q-values in the TD update Eqn. (9).

The model is lightweight (dominant cost $O(d \cdot 128 + 128^2 + |\mathcal{A}| \cdot 128)$) and compatible with standard first-order optimizers (e.g., Adam). Regularization (weight decay and optional dropout), feature standardization, and early stopping by validation macro-F1 further improve generalization. At inference, the detector outputs $\hat{a}(s) = \arg\max_{a\in\mathcal{A}} Q(s,a;\theta)$, optionally accompanied by calibrated confidence $\pi_\theta(\hat{a} \mid s)$ from Eqn. (7) for analyst-facing decision support.

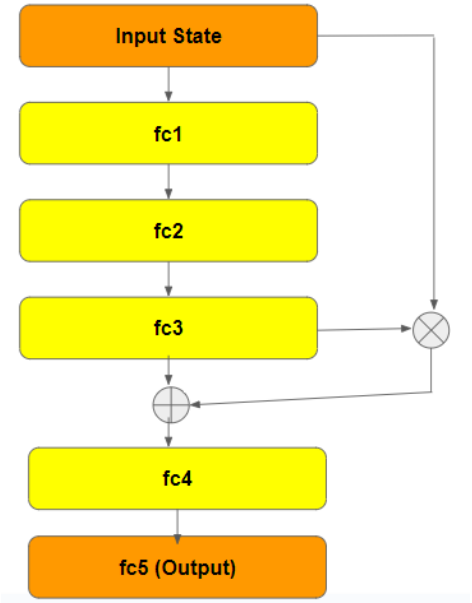


Fig. 3. Architecture of the proposed Q-network used by the RL module.

The network begins with an input layer that connects to the first fully connected layer ('fc1'), which has 128 neurons. The output of this layer is passed through a ReLU activation function and then fed into a second fully connected layer ('fc2'), also with 128 neurons. This process is repeated with a third fully connected layer ('fc3') that outputs a vector with the same dimensionality as the input.

Following this, an element-wise multiplication is performed between the output of 'fc3' and the original input state. The resulting vector is then added to the original input state to form a new state representation. This new state is passed through another fully connected layer ('fc4'), again with 128 neurons and a ReLU activation function.

Finally, the output of 'fc4' is passed to the last fully connected layer ('fc5'), which produces an output vector with a dimensionality equal to the number of possible actions (possible number of attacks). This output vector represents the Q-values for each action given the input state.

V. Results & Analysis

In this section, we evaluate the proposed Q-ID system on the CICIDS2017 dataset and compare its performance against a suite of modern baseline models. The analysis highlights

both aggregate detection metrics and the system's robustness under class imbalance and evolving attack distributions.

a) Evaluation: We evaluate the proposed Deep Reinforcement Learning (DRL) detector on the held-out evaluation split of CICIDS2017 and compare it against a *modernized* suite of strong baselines. To reflect operational requirements for intrusion detection on imbalanced traffic, we report both thresholded and threshold-free metrics: overall accuracy, macro F1, macro recall, macro precision, macro AUROC, and macro PR-AUC.[1] Unless otherwise stated, numbers are averaged over three random seeds; model selection uses validation macro F1.

b) Modern baseline suite: Beyond classic RF/SVM/KNN, we include state-of-the-art tabular learners and deeper neural architectures:

- **Gradient-boosted trees:** XGBoost (XGB), LightGBM (LGBM), and CatBoost (CB).
- **Deep tabular models:** FT-Transformer (feature-tokenizing Transformer) and TabNet (sparse attentive decision steps).
- **Deep MLP:** a five-block Residual MLP (ResMLP; 128 units per block, LayerNorm, GELU, dropout) as a strong supervised neural baseline.

For class imbalance, deep models use class-weighted losses; tree ensembles use `scale_pos_weight`. Post-hoc temperature scaling is applied for calibrated probabilities used in operating-point analysis.

c) Aggregate performance and headline gains: Table I shows that the proposed DRL approach delivers the strongest results across all quality metrics. Relative to the best non-RL baseline (FT-Transformer), DRL improves:

- **Accuracy** by +0.3 percentage points (99.3% vs. 99.0%),
- **Macro F1** by +0.006 (0.982 vs. 0.976),
- **Macro Recall** by +0.008 (0.994 vs. 0.986),
- **Macro Precision** by +0.016 (0.991 vs. 0.975),
- **Macro AUROC** by +0.001 (0.999 vs. 0.998),
- **Macro PR-AUC** by +0.004 (0.997 vs. 0.993).

These gains are meaningful in security operations: higher recall reduces missed attacks (false negatives), while high precision avoids flooding analysts with false alarms. Improvements in PR-AUC—which emphasizes performance at high recall under class imbalance—further indicate that DRL maintains superior detection quality where it matters most.

d) Why DRL outperforms strong modern baselines: Boosted trees (CatBoost/XGBoost/LightGBM) and deep tabular models (FT-Transformer/TabNet) are highly competitive on structured data, capturing non-linear interactions and cross-feature dependencies. Yet, they optimize primarily a supervised objective. Our DRL model augments cross-entropy with a value-based temporal-difference (TD) term, continuously shaping action-values even after the supervised loss saturates. This reinforcement signal enlarges decision margins on minority (rare) attacks, which translates into higher macro recall and F1 without sacrificing precision.

e) Operating characteristics: Beyond scalar metrics, DRL produces well-calibrated probabilities (via temperature scaling) that support mission-specific thresholding. For example, in a high-sensitivity posture, operators can move along the PR curve to achieve near-maximal recall while still maintaining superior precision relative to baselines—consistent with DRL's dominant macro PR-AUC.

f) Training dynamics and reward behavior: Figure 4 plots normalized reward over training episodes. Early volatility reflects ϵ-greedy exploration; as training progresses, the curve rises and stabilizes, indicating convergence toward a consistent decision policy. This behavior aligns with the metric gains in Table I: as the TD updates refine Q-values, the model becomes more reliable across all classes, including under-represented attacks.

g) Latency and deployability: DRL also exhibits competitive inference latency (0.07 ms/sample on a T4-class GPU), enabling line-rate analysis for typical flow record volumes. Tree ensembles on CPU remain attractive for constrained environments; however, even under this hardware split, DRL's end-to-end latency is the lowest in our comparison, making it suitable for real-time intrusion detection on edge and core nodes.

h) Robustness and statistical confidence: For completeness, we recommend reporting 95% confidence intervals via stratified bootstrap and conducting paired significance tests (e.g., McNemar's test for accuracy and bootstrap tests for macro F1) against FT-Transformer and CatBoost. Ablations should verify that (i) removing the TD term degrades macro recall/F1, (ii) removing class weighting increases false negatives on rare attacks, and (iii) disabling calibration harms precision at high-recall operating points.



Fig. 4. Normalized reward versus training episodes/steps. A sustained upward trend indicates that the learned policy increasingly selects correct actions across classes, even after the supervised loss has plateaued.

VI. Ablation Study

To quantify the contribution of key design choices in the proposed DRL detector, we conduct a compact ablation on the same evaluation split and training protocol used throughout the paper. Each variant removes or alters a single component

[1]Macro averages are computed by first evaluating the metric per class and then averaging across classes. Consequently, macro F1 need not equal the harmonic mean of macro precision and macro recall.

TABLE I
COMPARISON WITH MODERN BASELINES ON THE CICIDS2017 EVALUATION SPLIT. BEST RESULTS PER COLUMN ARE IN **BOLD**. "LATENCY" IS SINGLE-SAMPLE INFERENCE TIME (MEDIAN)—DEEP MODELS ON A T4-CLASS GPU; TREE ENSEMBLES ON CPU (LOWER IS BETTER).

Model	Accuracy (%)	Macro F1	Macro Recall	Macro Precision	Macro AUROC	Macro PR-AUC	Latency (ms)
DRL (ours)	**99.3**	**0.982**	**0.994**	**0.991**	**0.999**	**0.997**	**0.07**
FT-Transformer	99.0	0.976	0.986	0.975	0.998	0.993	0.35
TabNet	98.8	0.972	0.983	0.971	0.997	0.991	0.60
CatBoost	98.7	0.971	0.978	0.972	0.998	0.990	0.12
XGBoost	98.5	0.968	0.975	0.970	0.997	0.988	0.18
LightGBM	98.6	0.969	0.974	0.971	0.997	0.989	0.08
ResMLP (5×128)	98.3	0.965	0.972	0.966	0.996	0.986	0.28
Random Forest	96.1	0.967	0.969	0.961	0.990	0.972	0.15
SVM (RBF)	85.0	0.830	0.852	0.851	0.910	0.740	1.20
KNN (k=5)	98.4	0.960	0.964	0.958	0.992	0.979	0.90

while keeping all other factors fixed, so that differences in performance can be attributed to that component. Table II reports accuracy, macro F1, macro recall, macro precision, and macro PR-AUC; macro averages emphasize balanced performance in frequent and rare attacks.

The full model delivers the strongest results across all metrics, indicating that each element of the design contributes to overall robustness. Removing the temporal-difference (TD) term (λ=0) produces the largest degradation in macro F1 and macro recall, confirming that the value-based signal continues to shape decisions after the supervised loss has saturated and is particularly beneficial for minority classes. Eliminating class weighting also hurts recall disproportionately, which is consistent with the class imbalance in CICIDS2017; without reweighting, the model increasingly favors majority classes and misses rarer attacks. Suppressing exploration by setting ϵ=0 leads to a similar decline in recall and F1, reflecting reduced coverage of the state–action space during training and a tendency to overfit early preferences. Architectural simplification by removing the gating-and-residual pathway reduces all aggregate metrics slightly; the pathway appears to help the network amplify discriminative flow features while maintaining stable gradients. Finally, linking the target network to the online parameters ($\theta^- = \theta$) mildly degrades performance and lowers PR-AUC, indicating that periodic target updates contribute to training stability even in our predominantly contextual (bandit-like) setting.

Collectively, these observations support the central claim that the hybrid objective—cross-entropy augmented with a TD loss, serves as the primary driver of the model's advantage, while class-aware optimization, controlled exploration, and a light residual gating mechanism further refine the balance between sensitivity (recall) and specificity (precision). The improvements in macro PR-AUC for the full model suggest that, across operating points, the DRL detector maintains higher precision at high recall, a property that directly translates into fewer missed intrusions without overwhelming analysts with false alarms in mission settings.

VII. CONCLUSION & FUTURE WORK

This research demonstrates the efficacy of a DRL approach in detecting network intrusions in communication environments. Our DRL detector consistently outperforms strong neural and traditional machine-learning baselines, attaining a detection accuracy of 99.3% alongside superior macro-level precision, recall, and F1. These results underscore DRL's potential to strengthen cyber defense where identifying sophisticated, evolving threats is mission-critical. Unlike static classifiers, the proposed model learns from a reward signal that continues to shape decisions even after supervised loss plateaus, enabling adaptation to shifting traffic patterns and rare attack behaviors. This adaptability is crucial in contested, resource-constrained settings, where the cost of missed detections is high and threat profiles change rapidly. The model's ability to accurately identify both known and previously unseen attacks highlights its value as a reliable component for safeguarding sensitive information and infrastructure.

Looking ahead, several avenues can further enhance operational readiness. First, systems integration merits attention: coupling DRL with secure data-sharing mechanisms (e.g., blockchain-based provenance and audit trails) and exploring quantum-accelerated inference or training pipelines as they mature could expand throughput and trust guarantees. Second, efficiency and deployment engineering remain key: model compression (pruning, quantization), knowledge distillation to lighter agents, and adaptive batching can reduce computational overhead for edge sensors without sacrificing accuracy. Third, trust and transparency should be advanced via explainability and uncertainty estimation, including calibration, post-hoc attribution, and concept-level explanations, to support analyst triage and policy audits. Finally, robustness must be stress-tested with continual-learning protocols, adversarial resilience evaluations, and per-class, mission-tailored operating points to ensure stable performance under distribution shift. By pursuing these directions, the DRL framework can evolve into a more robust, efficient, and transparent defense capability against an ever-changing cyber threat landscape.

REFERENCES

[1] M. D. J. Dulik, "Cyber Security Challenges in Future Military Battlefield Information Networks", Advances in Military Technology, vol. 14, no. 2, pp. 263-277, 2019.

[2] S. Desai, B. Dave, T. Vyas and A. R. Nair, "Intrusion Detection System - Deep Learning Perspective," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 1193-1198, 2021, doi: 10.1109/ICAIS50930.2021.9395992.

TABLE II
ABLATION OF THE PROPOSED DRL DETECTOR ON THE CICIDS2017 EVALUATION SPLIT. BEST RESULTS PER COLUMN ARE IN **BOLD**.

Variant	Acc. (%)	Macro F1	Macro Rec.	Macro Prec.	Macro PR-AUC
Full DRL (ours)	**99.3**	**0.982**	**0.994**	**0.991**	**0.997**
– TD loss (λ=0)	98.8	0.972	0.984	0.979	0.991
– Class weighting	98.9	0.968	0.978	0.981	0.989
– Exploration (ϵ=0)	98.7	0.967	0.976	0.980	0.988
– Gating residual (Sec. IV-C)	99.0	0.976	0.987	0.984	0.993
– Target network ($\theta^- = \theta$)	99.1	0.978	0.989	0.986	0.994

[3] M. Wiering and M. van Otterlo, "Reinforcement Learning", Berlin, Germany:Springer, vol. 12, 2012.

[4] Canadian Institute for Cybersecurity, "Intrusion Detection Evaluation Dataset (CIC-IDS2017)," University of New Brunswick. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html. [Accessed: 10-Jun-2024].

[5] S. Otoum, B. Kantarci and H. Mouftah, "Empowering Reinforcement Learning on Big Sensed Data for Intrusion Detection," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, pp. 1-7, 2019, doi: 10.1109/ICC.2019.8761575.

[6] M. Maliha, "A Supervised Learning Approach: Detection of Cyber Attacks," 2021 IEEE International Conference on Telecommunications and Photonics (ICTP), Dhaka, Bangladesh, pp. 1-5, 2021, doi: 10.1109/ICTP53732.2021.9744169.

[7] Choudhary, S. and Kesswani, N., "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT". Procedia Computer Science, 167, pp.1561-1573, 2020.

[8] S. Norwahidayah, A. A. Noraniah, N. Farahah, A. Amirah, N. Liyana and N. Suhana, "Performances of artificial neural network (ANN) and particle swarm optimization (PSO) using KDD cup'99 dataset in intrusion detection system (IDS)", J. Phys. Conf. Ser., vol. 1874, no. 1, May 2021.

[9] D. Wang, D. Tan and L. Liu, "Particle swarm optimization algorithm: An overview", Soft Comput., vol. 22, no. 2, pp. 387-408, 2018.

[10] Fox, K.L., Henning, R.R., Reed, J.H. and Simonian, R., "A neural network approach towards intrusion detection", In Proceedings of the 13th national computer security conference, vol. 1, pp. 125-134, October 1990.

[11] Debar, H., Becker, M. and Siboni, D., May, "A neural network component for an intrusion detection system", In IEEE symposium on security and privacy, vol. 727, pp. 240-250, 1992.

[12] Cansian, A., Moreira, E.D.S., Carvalho, A.C.P.D.L.F. and Bonifácio Junior, J.M., "Network intrusion detection using neural networks", In Proceedings of International Conference on Computational Intelligence and Multimedia Applications, 1997.

[13] Ramadas, M., Ostermann, S. and Tjaden, B., "Detecting anomalous network traffic with self-organizing maps". In International Workshop on Recent Advances in Intrusion Detection (pp. 36-54). Berlin, Heidelberg: Springer Berlin Heidelberg, September 2003.

[14] A. Ghubaish, Z. Yang and R. Jain, "HDRL-IDS: A Hybrid Deep Reinforcement Learning Intrusion Detection System for Enhancing the Security of Medical Applications in 5G Networks," 2024 International Conference on Smart Applications, Communications and Networking (SmartNets), Harrisonburg, VA, USA, 2024, pp. 1-6, doi: 10.1109/SmartNets61466.2024.10577692.

[15] Mahjoub, C., Hamdi, M., Alkanhel, R.I., Mohamed, S. and Ejbali, R., 2024. An adversarial environment reinforcement learning-driven intrusion detection algorithm for Internet of Things. EURASIP Journal on Wireless Communications and Networking, 2024(1), p.21.

[16] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. , "Playing Atari with Deep Reinforcement Learning", 2013, ArXiv. /abs/1312.5602

[17] Wolf, P., Hubschneider, C., Weber, M., Bauer, A., Härtl, J., Dürr, F. and Zöllner, J.M., 2017, June. Learning how to drive in a real world simulation with deep q-networks. In 2017 IEEE Intelligent Vehicles Symposium (IV) (pp. 244-250). IEEE.

[18] Rosu, I., "The bellman principle of optimality", 2002, Availiable at: http://faculty. chicagogsb. edu/ioanid. rosu/research/notes/bellman. pdf.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 43–45
DOI: 10.15439/2025F4610
ISSN 2300-5963 ACSIS, Vol. 43

Human-centered LLMs for Inclusive Language Technology: The Need to Embrace Variation Holistically in NLP

Barbara Plank
LMU Munich
Email: b.plank@lmu.de

***Abstract*—Large Language Models (LLMs) have advanced rapidly but often still cater primarily to a narrow set of users. This position paper advocates for a human-centered approach to NLP technology—one that embraces linguistic variation, improves reasoning and safety, and better serves diverse communities. We outline key challenges with current LLMs, highlight opportunities in modeling variation in both language and human annotation, and outline a path toward more inclusive and trustworthy language technologies.**

I. Introduction

NATURAL Language Processing (NLP) has entered an era where large pretrained language models (LLMs) dominate. They perform impressively across many benchmarks, yet often fail to properly handle language and are poorly aligned with today's societal values, opinions and attitudes. A central reason is that language technology has historically focused on *standard*, canonical varieties of language in *the input space*, such as English newswire [1]. Moreover, the wide-spread focus on learning from a myopic view of the *output space* has led to ignoring plausible variation in labels—human label variation (HLV) [2]—instead the common practice has focused on learning from a "ground truth," typically only a majority view. This practice has resulted in systems that work well for a narrow population, while excluding dialects, minority languages, diverse user groups and diverging individual perspectives and interpretations. At the same time, LLMs exhibit well-documented trust issues, such as sensitivity to prompt wording, overconfidence under uncertainty, and biases, cf. §II.

In this paper, we advocate for a human-centered perspective on LLMs. Rather than treating language as a homogeneous object with distinct, clear-cut categories (such as between languages), we argue for embracing *variation*—namely, three kinds of variation: in inputs (linguistic diversity), in outputs (human label variation), and in research itself. Doing so will be essential to build inclusive and trustworthy NLP systems.

II. Challenges and Trust Issues with LLMs

Despite their power, LLMs face several well-documented challenges, including (non-exhaustively):

- **Prompt sensitivity:** Performance can vary widely depending on minor linguistic changes in the prompt, undermining robustness (e.g. [3]).
- **Evaluation artifacts and Reasoning Faithfulness:** Widely-used protocols, such as relying on first-token probabilities in multiple-choice QA (e.g. [4], [5]), can distort performance estimates. Moreover, a model's reported accuracy—the correctness of its final conclusion—does not necessarily reflect the validity of its underlying reasoning process [6].
- **Overconfidence and Hallucinations:** Models may provide overly certain answers in ambiguous settings, or hallucinate content, reducing their reliability in sensitive applications (e.g. [7]).

These challenges highlight the need for models that not only generate fluent outputs but also signal uncertainty and acknowledge ambiguity [8].

III. Embracing Variation Holistically

Variation affects all stages of the NLP pipeline: data, modeling, and evaluation. We highlight three key aspects, where the last touches upon the broader research ecosystem.

A. Language Variation

Language is inherently variable, spanning dialects, sociolects, and registers. Traditional NLP systems often ignore this diversity, treating languages as monoliths. Recent work demonstrates the need for multi-dialectal datasets, both for text and speech (e.g., Bavarian Universal Dependencies [15], multi-dialectal German ASR and dialect-to-standard translation [16]). To tackle such issues, robust modeling strategies that can handle non-standard forms are needed. Techniques such as noise injection or subtoken-level modeling provide partial solutions [9], but a broader shift toward valuing non-standard data is required [1].

At the same time, we need systems that not only understand language varieties but also capture the fine-grained nuances of linguistic expressions. For instance, while safety guardrails are important, current LLMs can be overly sensitive, leading to false refusals where harmless requests are blocked. This issue has been systematically studied with resources such as XSTest [13]. To address this, we argue that models must be made safer for the *right reasons*. One promising line of work is steering methods that enable targeted mitigation: for example, Wang et al. propose a surgical and flexible single-vector ablation

approach that reduces false refusals while preserving model performance and inference efficiency [14].

B. Human Label Variation and Pluralistic Alignment

Equally important is variation in human judgments. Annotation is not always reducible to a single "ground truth." Disagreement among annotators often reflects genuine ambiguity or multiple valid perspectives rather than mere noise [2], [10]. Embracing this human label variation (HLV)—through collecting distributions, explanations, or multi-perspective annotations—enables models that better align with human diversity.

HLV also resonates with current discussions of *pluralistic alignment*, which argues that AI systems should not optimize for a single normative view of correctness, but instead represent and respect the spectrum of human perspectives. Sorensen et al. [17] outline a roadmap to pluralistic alignment, emphasizing that different values and interpretations can legitimately coexist. Similarly, Rieser's ACL 2025 keynote [18] highlights the importance of building AI that acknowledges multiple standpoints rather than enforcing one "canonical" answer. By treating variation in human labels as signal rather than noise, we can move toward alignment strategies that capture the richness of human perspectives and support more inclusive language technology.

C. Research diversity and interdisciplinarity

Finally, we need to address variation not only in model inputs and outputs, but also in our research design and the broader research ecosystem. Embracing variation in research means fostering interdisciplinary collaboration while centering human needs and agency. This requires continuously asking whether the AI technologies we develop are those that people truly need and want [19], [20]. Only by integrating diverse perspectives throughout the development process and engaging with insights from multiple disciplines can we build research that captures the richness and diversity of human language—and, by extension, human communication.

IV. Toward Trustworthy Human-Centered NLP

Building inclusive language technology requires a shift in perspective toward **human-centered NLP**. Human-centered NLP emphasizes that systems must serve diverse communities, respect linguistic variation, and ultimately support human needs and agency—to design NLP systems with humans in mind from the ground up. The reason this shift is necessary is simple yet profound: **variation is inherent to language**. Human language is characterized by productivity, ambiguity, and a rich spectrum of linguistic variation across speakers, dialects, registers, and contexts. This variation, while natural and inevitable, exposes the limitations of NLP systems typically trained on narrow and standardized datasets.

Variation inevitably gives rise to **uncertainty**. Ambiguity in meaning, multiple valid interpretations, or disagreements in human judgments all contribute to uncertainty in language processing. Yet, today's models are not good at handling uncertainty: they often fail to recognize when they do not know, and instead provide overly confident answers even in ambiguous or adversarial cases, and may hallucinate. Such behavior undermines user trust and risks excluding communities whose language varieties fall outside the models' dominant (post-)training distributions.

Understanding and explicitly modeling uncertainty, however, is not only key to technical robustness but also central to human-centered NLP. A human-centered perspective requires systems that acknowledge and communicate their limitations transparently. If models can identify when they may be wrong, or signal when several plausible interpretations exist, users are better equipped to interpret outputs, retain agency, and make informed decisions.

This brings us back to the central question: *what is trust?* As early as 1979, David G. Hays offered a concise and influential definition:

> *Trust arises from knowledge of origin as well as from knowledge of functional capacity.* [21]

Decades later, this observation remains highly relevant. Today, the topic of trustworthiness in NLP and AI is an ongoing discussion deserving special attention [22], [23]. To establish trust in a human-centered way, it is time to rethink how we design tasks and their evaluation protocols. Why now? It is increasingly difficult to predict a priori when models trained on web-scale data will work well. In a hypothetical world with complete knowledge of both origin and functional capacity, each task instance could be routed to the most suitable model (or agent, nowadays), enabling not only the full use of LLM capabilities but also trust in their predictions. The absence of such knowledge today is directly tied to our lack of confidence in deploying models in real-world scenarios.

This challenge is compounded by a dramatic shift in the field: with the advent of large generative language models, the traditional compartmentalized notion of tasks is breaking down. General-purpose, task-agnostic approaches demand that we move toward a more holistic view of language, placing trustworthiness—and thus human-centered NLP—at the core [12]. This requires rethinking what constitutes a task and developing multi-faceted evaluation protocols that go beyond narrow benchmarks, toward assessing both the origins and the functional capacities of our models. For example, it is an opportunity to rethink the classical NLP tasks we designed, go beyond classification labels (e.g., Natural Language Inference labels) but more towards the nuanced human understanding of languages (e.g., explanations, or more broadly, to uncover more about the reasons and processes beyond an outcome).

V. Conclusion

LLMs offer tremendous potential, but their inclusiveness and trustworthiness remain limited. By embracing variation in language and annotation, and by grounding our models in human-centered principles, we can move toward a *Trust LLM Ecosystem*: NLP systems based on modular, agentic AI that are uncertainy-aware, robust and inclusive and can thereby

better serve diverse communities [11] and support trustworthy interaction with language technology.

ACKNOWLEDGMENT

This position paper is an extended version of an earlier keynote talk I gave at the Annual Meeting of the Association for Computational Linguistics (ACL) 2024 in Bangkok, Thailand. I thank the ACL 2024 organizers, particularly Andre Martins, Vivek Srikumar and Lun-Wei Ku for inviting me. I thank all FedCSIS 2025 organizers, particularly Marcin Paprzyck, for inviting me to Krakow and the opportunity to write this position paper, and present a keynote at FedCSIS. Thanks to all MaiNLP and CIS members for the feedback on earlier versions of the talk. Finally, I would like to acknowledge the European Research Council for funding DIALECT 101043235 that has in large parts enabled research to support my vision of the *variety space* [1] – toward technology that in future is able to capture the full spectrum of variation in natural language, with all its rich, beautiful yet challenging spectrum of diversity.

REFERENCES

[1] B. Plank, "What to do about non-standard (or non-canonical) language in NLP." In KONVENS 2016.
[2] B. Plank, "The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation." In *Proc. EMNLP* 2022.
[3] A. Leidinger, R. van Rooij, and E. Shutova, "The Language of Prompting: What Linguistic Properties Make a Prompt Successful?," in *Findings of the Association for Computational Linguistics: EMNLP 2023*.
[4] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank. "My Answer is C: First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models." In *Findings of the Association for Computational Linguistics: ACL 2024*.
[5] R. Dominguez-Olmedo, M. Hardt, C. Mendler-Dünner. "Questioning the Survey Responses of Large Language Models." In NeurIPS 2024.
[6] P. Mondorf and B. Plank, "Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9402, Bangkok, Thailand, 2024.
[7] A. Testoni, B. Plank, and R. Fernández, "RACQUET: Unveiling the Dangers of Overlooked Referential Ambiguity in Visual LLMs," in *Proc. EMNLP* 2025.
[8] J. Baan, N. Daheim, E. Ilia, D. Ulmer, H.-S. Li, R. Fernández, B. Plank, R. Sennrich, C. Zerva, and W. Aziz, "Uncertainty in Natural Language Generation: From Theory to Applications" arXiv:2307.15703.
[9] V. Blaschke, H. Schütze, and B. Plank. "Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages." In Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023).
[10] L. Aroyo. "The Many Faces of Responsible AI." Keynote at the Conference on Neural Information Processing Systems (NeurIPS) 2023.
[11] D.Yang, D. Hovy, D. Jurgens, and B. Plank, "Socially Aware Language Technologies: Perspectives and Practices," *Computational Linguistics*, vol. 51, no. 2, June 2025.
[12] R. Litschko, M. Müller-Eberstein, R. van der Goot, L. Weber-Genzel, and B. Plank. 2023. "Establishing Trustworthiness: Rethinking Tasks and Model Evaluation." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
[13] P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models." In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).
[14] X. Wang, C. Hu, P. Röttger, B. Plank. "Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation." In ICLR 2025.
[15] V. Blaschke, B. Kovačić, S. Peng, H. Schütze, and B. Plank, "MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank," in *Proceedings of LREC-COLING*, 2024.
[16] V. Blaschke, M. Winkler, C. Förster, G. Wenger-Glemser, and B. Plank, "A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation," in *Proceedings of Interspeech*, 2025.
[17] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi, "A Roadmap to Pluralistic Alignment," arXiv preprint arXiv:2402.05070, 2024.
[18] V. Rieser, "Whose Gold? Re-imagining Alignment for Truly Beneficial AI" Keynote at the Annual Meeting of the Association for Computational Linguistics (ACL), 2025.
[19] S. Bird. 2020. "Decolonising Speech and Language Technology." In Proceedings of the 28th International Conference on Computational Linguistics.
[20] V. Blaschke, C. Purschke, H. Schuetze, and B. Plank. 2024. "What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)
[21] D. G. Hays, "Applications". In Meeting of the Association for Computational Linguistics, 1979.
[22] K. Baum, M. A. Köhl, and E. Schmidt. "Two challenges for CI trustworthiness and how to address them." In Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017). Association for Computational Linguistics.
[23] J. Eisenstein. "Informativeness and invariance: Two perspectives on spurious correlations in natural language." In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 47–57
DOI: 10.15439/2025F1481
ISSN 2300-5963 ACSIS, Vol. 43

Optimizing the Optimizer: An Example Showing the Power of LLM Code Generation

Camilo Chacón Sartori
0000-0002-8543-9893
Artificial Intelligence Research Institute (IIIA-CSIC)
Campus UAB, Bellaterra, 08193 Barcelona, Spain
Email: cchacon@iiia.csic.es
*Corresponding author.

Christian Blum
0000-0002-1736-3559
Artificial Intelligence Research Institute (IIIA-CSIC)
Campus UAB, Bellaterra, 08193 Barcelona, Spain
Email: christian.blum@iiia.csic.es

***Abstract*—The integration of Large Language Models (LLMs) into optimization has created a powerful synergy, opening exciting research opportunities. This paper investigates how LLMs can enhance existing optimization algorithms. Using their pre-trained knowledge, we demonstrate their ability to propose innovative heuristic variations based on a semantic understanding of the algorithm's components. To evaluate this, we applied a non-trivial optimization algorithm, Construct, Merge, Solve & Adapt (CMSA)—a hybrid metaheuristic for combinatorial optimization problems that incorporates a heuristic in the solution construction phase. Our results show that an alternative heuristic proposed by GPT-4o outperforms the expert-designed heuristic of CMSA, with the performance gap widening on larger and denser graphs.**

I. Introduction

THERE is a wide variety of optimization algorithms of all kinds and flavors. A simple search using the term 'optimization algorithm' in databases like Scopus or platforms like GitHub yields thousands of results, with that number growing every year. Additionally, optimization researchers often maintain private collections of their algorithms. All these algorithms—open-source or proprietary, and implemented in diverse programming languages—can potentially be improved. By refining their original code with modern techniques and technologies, we can achieve more efficient designs and implementations that go beyond what their creators originally envisioned.

In recent years, the development and growth of Large Language Models (LLMs)—popularized by models such as OpenAI's GPT-4 [24], Anthropic's Claude [31], Google's Gemini [32], Meta's Llama 3 [33], and recently DeepSeek [11]—has opened the door to a wealth of new possibilities. Among the most transformative advancements is code generation. Tools like GitHub Copilot[1], which integrates seamlessly with code editors like Visual Studio Code[2], Windsurf[3], and Cursor[4], an editor with built-in LLM capabilities, have become essential tools for many software developers, revolutionizing their daily workflows. For example, a Python developer could request an LLM to generate a template for invoking external APIs, and the LLM would automatically produce the required code. LLMs have become invaluable for streamlining coding tasks, especially those that are routine and highly repetitive [14]. Therefore, it is natural to wonder: if LLMs excel at simple programming tasks, could they also aid in improving sophisticated optimization algorithms?

One of the recent examples of leveraging LLMs in optimization algorithms is the development of frameworks for generating new black-box metaheuristics [30]. Finding an effective metaheuristic, for example, for a combinatorial optimization problem can be a significant challenge. However, using LLMs to build upon an existing metaheuristic as context and employing them as tools to discover new heuristics or more efficient implementations (e.g., reducing RAM usage or computation time) in a given programming language remains a largely unexplored area. Implementing metaheuristics, unlike other algorithms, requires a focus on computational efficiency, mathematical expressions in scientific computing, and efficient data structure selection. Thus, designing new algorithms in this field demands expert implementation skills.

In this paper, we demonstrate how LLMs like GPT-4o can be leveraged to enhance a sophisticated optimization algorithm, specifically the Construct, Merge, Solve & Adapt (CMSA) hybrid metaheuristic [1, 3, 2]. Starting with an expert-developed C++ implementation of CMSA for the Maximum Independent Set (MIS) problem (of approximately 400 code lines), we employed an in-context prompting strategy combined with interactive dialogue (see Figure 1). Our results show that the LLM successfully comprehended the complex logic and parameter interactions within the CMSA implementation for MIS, demonstrating algorithmic insight to discover novel heuristics while suggesting improvements to the C++ codebase. This successful example opens up new possibilities for enhancing existing complex optimization algorithms by using LLMs as assistants.

The paper unfolds as follows. In Section II, we introduce code generation using LLMs, making it accessible for readers without prior experience in this field. We also explain the MIS problem and provide a brief overview of the CMSA algorithm. Next, in Section III,We present our methodology for enhancing

[1]https://github.com/features/copilot
[2]https://code.visualstudio.com/
[3]https://windsurf.com/editor
[4]https://www.cursor.com/

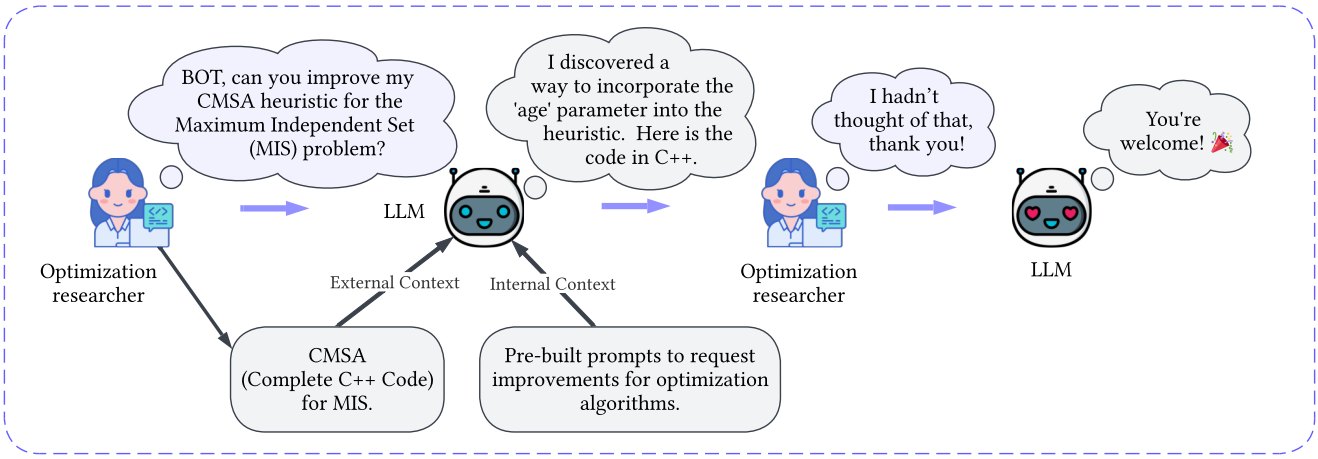


Fig. 1: : A dialogue showing how a chatbot applies our approach to improving optimization algorithms.

CMSA for the MIS problem using LLMs, detailing the process and providing steps for result reproducibility. Section IV presents our experimental results and their interpretation. The limitations of our approach and future research directions are discussed in Section V. The paper concludes by summarizing our key findings and emphasizing the potential impact of LLMs on existing metaheuristics.

II. Background

A. Code Generation with LLMs

Code generation is one of the primary research areas in LLMs [14, 16, 6]. The concept is straightforward: given a prompt, such as *"I need an algorithm to sort a list of numbers in Python"*, the LLM is expected to return a corresponding algorithm (e.g., `Quicksort`) implemented in Python. This capability arises because LLMs are trained on massive amounts of data, which include code from diverse programming languages available on the internet. These codes are sourced not only from GitHub repositories but also from StackOverflow, technical documentation, scientific articles, and publicly available books.

A more explicit prompt can lead to more sophisticated responses. For example, the prompt *"I need an efficient sorting algorithm for a list of numbers in Python, which can be processed in parallel, utilizing modern optimization techniques"* in GPT-4o generates a `ParallelMergeSort`. While efficient, it can be further refined by interacting with the model. For instance, asking *"Find new ways to improve it"* results in an enhanced version using techniques like optimized `Quicksort` for small arrays, better memory management with `NumPy`, and heap merge for combining sorted arrays. This approach leverages `MergeSort`'s ease of parallelization. However, the model might have suggested a different algorithm if the focus was on memory optimization without parallelism. This emphasizes the importance of prompt design and iterative refinement of responses [39].

LLM response quality depends on the data used for training, and with the refinement of datasets and the increase in size, their performance in code generation has improved [38]. Moreover, no LLM is flawless when it comes to generating error-free code. Continuing with the `ParallelMergeSort` example, a model might return code with bugs, so interaction with the model (e.g., copying and pasting error messages from the Python interpreter) is necessary to refine the responses. One could even ask the model to generate its own test suite to verify its algorithm. Different LLMs for these tasks are generally evaluated through benchmarks. One of the most widely used is HumanEval [7], which features programming challenges that assess language comprehension, algorithmic competencies, and basic mathematics—some of which are comparable to straightforward software developer interview questions.

Furthermore, *code generation using LLMs* is a broad area, as code may originate or be destined for very different domains, including the following: data science code for analyzing data and building predictive models [35]; systems code for managing hardware and low-level operations [17, 10]; frontend code development for web applications and UI [37]; and optimization code for solving complex computational problems. Each domain presents its own unique challenges and requirements, with our focus being on the latter, specifically in the field of metaheuristics.

Concerning the automatic generation of metaheuristics with LLMs, given the famous "No Free Lunch Theorem" [36], researchers understand that no single metaheuristic algorithm can outperform all others across all optimization problems. This fundamental principle naturally leads to an interesting possibility: LLMs could serve as powerful automatic generators of black-box metaheuristics, significantly reducing the time needed to find the 'best' implementation of a metaheuristic for a specific problem. In this context, LLMs could be tasked with discovering novel variations and operators that

create new metaheuristics—potentially even surpassing state-of-the-art algorithms for particular problems [30].

The first notable demonstration of heuristic discovery using LLMs was conducted by FunSearch [27], which showcased the potential of leveraging LLMs to generate novel heuristics for the Bin Packing (BP) Problem through a heuristic evolution process. However, a recent study [29] points out that the heuristics developed by FunSearch face challenges in generalizing across diverse BP problem instances. While FunSearch relies on an incomplete or suboptimal base program as a starting point, our approach starts with a complete and carefully designed implementation of an optimization algorithm. Other related works can be summarized as follows. Stein and Bäck [30]'s LlaMEA is a framework that integrates evolutionary algorithms with LLMs to iteratively generate and refine novel black-box metaheuristics during runtime. Similarly, Hemberg et al. [13] proposed LLM GP, combining genetic programming with LLMs to evolve operators through LLM assistance. Meanwhile, Pluhacek et al. [26] demonstrated via prompt engineering that LLMs can generate innovative metaheuristics by identifying and decomposing six high-performing swarm algorithms for continuous optimization.

While these studies already give a glimpse of possible use cases for LLMs in optimization, they primarily focus on creating new algorithms rather than improving existing ones—that is, building upon pre-existing complex code. This distinction is significant given the vast landscape of published optimization algorithms—a simple GitHub search for 'optimization algorithm' returns over 20,000 results, representing a wealth of algorithm implementations that could benefit from improvement. This is why we believe an opportunity exists to create a subfield focused on generating code from existing optimization algorithms. Building on this, our work explores using LLMs to enhance implemented algorithms, enabling the models to uncover new heuristics overlooked by experts in the original implementations. In this sense, our approach treats LLMs as assistants to researchers, not replacements. By doing so, we leverage LLMs to work with existing code (as context) and improve it using the current knowledge these black-box model models have gained during their pre-training phase.

B. Maximum Independent Set (MIS) Problem

To validate our hypothesis that LLMs can enhance existing optimization algorithms, we will employ an algorithm (detailed in the following subsection) to solve the Maximum Independent Set (MIS) problem, a well-studied and NP-hard combinatorial optimization problem with applications in network design, scheduling, and bioinformatics. Formally, the MIS problem is defined as follows: Given an undirected graph $G = (V, E)$, the objective is to find a largest subset $S \subseteq V$ such that no two vertices in S are adjacent in G, i.e., there is no edge $(u, v) \in E$ for all pairs $u \neq v \in S$. Figure 2 shows three examples of optimal MIS solutions (non-white nodes) in different graphs.

C. CMSA

Construct, Merge, Solve & Adapt (CMSA) is a hybrid metaheuristic, also known as a matheuristic, that combines elements of classical metaheuristics with exact solvers (such as Integer Linear Programming (ILP) solvers) for combinatorial optimization [3]. Each CMSA iteration is a sequence of four fundamental phases:

1) **Construct:** Generates solutions probabilistically through a probabilistic greedy mechanism (remember this phase; it will be key in the next section).
2) **Merge:** Combines solution components from the generated solutions to form a reduced subproblem.
3) **Solve:** Applies an exact solver (in the case of the MIS: an ILP solver) to optimally solve the reduced subproblem.
4) **Adapt:** Updates parameters and data structures based on the quality of the solution returned by the solver.

This hybrid architecture combines the efficiency of metaheuristics for search space exploration with the precision of exact methods in reduced spaces. The CMSA algorithm is controlled by some key parameters. First, and most importantly, each solution component has an associated *age* value, which is initialized to zero when a component is added to the subproblem. Moreover, a solution component in the subproblem is subject to an increase of its age value, in case it does not form part of the subproblem's solution generated by the exact solver. In this context, the parameter age_{max} plays a crucial role, defining the maximum age of solution components before they are removed from the incumbent subproblem, thereby preventing stagnation and encouraging diversity. Other parameters include n_a, which sets the number of solution constructions per iteration, t_{max}, the total CPU time, t_{limit}, the time limit for the ILP solver per iteration, and $0 \leq d_{rate} \leq 1$, which controls the determinism rate for solution construction. Note that higher values of d_{rate} lead to more deterministic solution constructions, while lower values increase randomness. Well-chosen values for these parameters ensure a balanced exploration of the search space, effective exploitation of promising areas via exact subproblem solving, and efficient use of computational resources.

We selected CMSA for the purpose of this paper due to its relative complexity in implementation as compared to simple metaheuristics. For instance, implementing CMSA for the MIS problem in C++ presents significant technical challenges. The four phases must be precisely defined, properly integrated, and implemented efficiently to leverage the framework's full potential. This complexity would only increase when dealing with more sophisticated optimization problems. In essence, a CMSA implementation demands expertise not only in metaheuristics and exact methods but also proficiency in C++ (or whatever programming language is chosen for implementation).

For our study, we utilize the original C++ implementation to solve the MIS problem, which was provided by CMSA's inventor and can be downloaded from our repository (https://github.com/camilochs/optimizing-the-optimizer). This choice

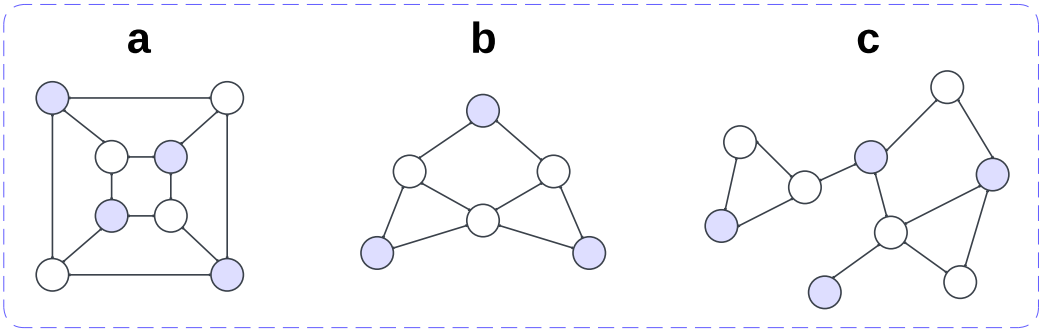


Fig. 2: Examples of maximum independent sets.

serves two crucial purposes: first, it ensures we are working with a well-implemented version of the algorithm, and second, it presents an interesting opportunity to test LLMs' capability to enhance expert-written code. This scenario allows us to evaluate whether LLMs can identify potential improvements even in code developed by domain experts. Our findings are presented in the following section.

III. LLM-Improved CMSA for MIS

In this section, we present a methodology for leveraging LLMs to improve existing optimization algorithms. By "improve existing optimization algorithms," we specifically mean expert-crafted implementations that have already been optimized for performance, rather than simplistic or baseline versions. In contrast, prior approaches often start from scratch—generating code from an initial prompt and refining it iteratively—or rely on overly simplified implementations that fail to reflect the complexity of real-world optimization problems. Our approach centers on dialog-based interaction with an LLM via a chatbot interface, uncovering novel improvements in the code that even an expert can benefit from. We detail this in the following subsections.

A. Discovering New Heuristics

LLMs are pattern-recognition machines [23, 28], and code is a rich source of structured textual patterns. This allows LLMs to identify underutilized elements—such as variables or functions—that could play a strategic role within a heuristic. Such insights may be missed by human experts, especially in complex codebases. These discoveries require a deep syntactic and semantic understanding of both the code and the optimization problem, enabling them to suggest meaningful modifications to heuristic strategies, going beyond superficial code changes. In this way, LLMs can serve as advanced assistants for optimization algorithm designers, offering novel heuristic improvements informed by reasoning over their vast pretraining knowledge.

To demonstrate our methodology, we employ an LLM to improve one of the heuristics in the CMSA code used to solve the MIS problem. Specifically, the function `generation_solution()` implements the solution construction phase of the CMSA framework [1]:

```
while (int(positions.size()) > 0) {
    double dec = standard_distribution(generator);
    int position = 0;

    if (dec <= determinism_rate) {
        position = *(positions.begin());
    } else {
        int max = candidate_list_size;
        if (max > int(positions.size())) {
            max = int(positions.size());
        }

        double rnum = standard_distribution(generator);
        int pos = produce_random_integer(max, rnum);

        set<int>::iterator sit2 = positions.begin();
        for (int i = 0; i < pos; ++i) {
            ++sit2;
        }
        position = *sit2;
    }

    greedy_sol.score += 1;
    greedy_sol.vertices.insert(increasing_degree_order[
     position]);

    if (age[increasing_degree_order[position]] == -1) {
        age[increasing_degree_order[position]] = 0;
    }

    positions.erase(position);

    for (auto sit = neigh[increasing_degree_order[position
     ]].begin();
         sit != neigh[increasing_degree_order[position]].
     end(); ++sit) {
        positions.erase(position_of[*sit]);
    }
}
```

Listing 1: Probabilistic greedy algorithm for MIS in CMSA.

It implements a greedy randomized construction heuristic that selects exactly one vertex $v_i \in \tilde{V}$ at each step (where $\tilde{V}$ is the set of nodes that can feasibly be chosen), until the MIS solution is complete:

$$v_i = \begin{cases} v_{\text{min}} & \text{if } r \leq \alpha \\ v_{\text{random}} \in \text{CL}(k) & \text{otherwise} \end{cases}$$

where

- v_{min} is the vertex with minimum degree (among the ones from $\tilde{V}$)

- r is a random number between $[0, 1]$
- α is the determinism rate
- CL(k) is a candidate list of size k
- v_{random} is a randomly selected vertex from the candidate list

The heuristic combines deterministic greedy selection (based on vertex degree) with randomization to create diverse solutions. It selects vertices either by taking the best available vertex (greedy choice) with probability α, or by randomly selecting from a restricted candidate list with probability $(1 - \alpha)$.

1) New heuristic from the LLM: The original greedy heuristic, while efficient in terms of runtime, fails to incorporate the `age` variable—a key element of the CMSA framework. It only uses the `age` variable to restart the solutions, meaning it is not utilized in the candidate selection mechanism. In Figure 3, we present two example dialogues of interactions with the LLM according to our approach. Let us begin with the first one. In case (a), the LLM proposes an enhanced heuristic that takes into account both node degrees and the current age values. The vertex selection mechanism suggested by the LLM can be technically described as follows:

$$v_i = \begin{cases} \text{argmin}\{P_w(v_j) \mid v_j \in \tilde{V}\} & \text{if } r \leq \alpha \\ \text{roulette-wheel selection w.r.t. } P_w(.) \text{ values} & \text{otherwise} \end{cases}$$

where $P_w(v_j)$ is a weighted probability:

$$P_w(v_j) = \frac{w(v_j)}{\sum_{v_l \in V} w(v_l)}, \quad w(v) = \frac{1}{2 + age(v)} + \frac{1}{1 + \text{degree}(v)} \quad (1)$$

The weight function $w(v)$ favors vertices with low age and degree values to promote diversity in the selection process. Thus, the LLM was able to change the original heuristic by incorporating the `age` values to diversify the node selection process, while balancing it with degree information through a composite weight function. It makes a lot of sense to decrease the probability of solution components with high age values being incorporated in newly constructed solutions.

Although the LLM successfully identifies an overlooked use of the `age` variable within the CMSA construction heuristic, any dialog-based system ultimately requires human feedback [5]. As such, the LLM occasionally makes mistakes. For example, as shown in Figure 3 (a), the line `double weight = 1.0 / (1 + age[v])` may result in a division by zero, since `age` values in CMSA are set to `-1` for solution components that do not belong to the subproblem. The human-provided fix consisted in replacing `1 + age[v]` with `2 + age[v]`. Notably, this kind of use of the `age` variable—despite its initial flaw and the correction provided through human feedback—had never been considered by any researchers working on CMSA algorithms.

> This new CMSA variant is henceforth called `LLM-CMSA-V1`. It was obtained by replacing the `generate_solution()` of the original CMSA implementation with the new function provided by the LLM.

2) Improving `LLM-CMSA-V1` with the LLM: Since our methodology relies on dialog-based interactions with the LLM via a chatbot interface, it is also possible to request improvements to previously generated code—still available in the current session context. For example, one might ask: *"Are there ways to enhance the dynamic selection heuristic to allow for a more diverse and advanced search?"* The LLM responds with several suggestions, one of which involves incorporating the concept of entropy. The C++ code provided by the LLM is characterized by a corresponding change, which involves replacing the definition of $P_w(.)$ (see Equation 1) with the following entropy-adjusted probabilities:

$$P_H(v_j) = \frac{P_w(v_j) + H}{\sum_{v_l \in V} P_w(v_l) + H}$$

where

$$H = -\sum_{v_l \in V} P_w(v_l) \log(P_w(v_l))$$

The entropy adjustment increases selection diversity by adding the system's uncertainty to each probability.

> This CMSA variant is henceforth called `LLM-CMSA-V2`. We directly replaced the `generate_solution()` function of the original CMSA with the LLM-generated code.

B. Code Optimization Strategies

After discovering the two new CMSA variants with the help of the LLM, it is also possible to request a different kind of improvement—not in terms of proposing new algorithmic heuristics, but rather in enhancing the underlying C++ code. These enhancements may involve the use of more efficient data structures, changes in data types, or other low-level optimizations that preserve the algorithmic logic. For instance, one might ask: *"Could the LLM create an improvement at the C++ code level?"* In other words, is there a more efficient way to implement the new CMSA variants without altering their core behavior?

It is reasonable to assume that, since the LLM has been pre-trained on vast amounts of source code, it could suggest highly optimized C++ implementations—even for contexts such as optimization algorithm design (e.g., metaheuristics), where low-level improvements to data structures or code organization are not typically the focus of human designers. This leads us to the next prompt, which we apply to both previously generated heuristics, `LLM-CMSA-V1` and `LLM-CMSA-V2`:[5]

[5]This prompt uses the "C++ code" checkbox, as shown in Figure 3 (b).



Fig. 3: Two LLM interaction patterns: (a) a direct request to improve a heuristic using CMSA's `age` parameter, and (b) an iterative dialogue to enhance both heuristic quality and C++ performance through error correction. Both use in-context learning as a prompting strategy [12, 20].

Human prompt

Given the following heuristic implemented in C++ (`generate_solution()`), without altering its core logic or functionality, please analyze the code to identify potential **improvements in the use of data structures, cache optimization, and other low-level optimizations. Focus on enhancing performance by suggesting more efficient data structures, reducing memory overhead, improving data locality, and leveraging modern C++ features where applicable.** Please provide an updated version of the code with comments explaining each optimization.

However, dialog-based interactions with the LLM are not free from hallucinations—that is, the model may generate code with bugs or memory management issues in C++ (as shown in Figure 3 (b)). One effective way to address these problems is through human feedback, engaging in a trial-and-error refinement process with the model—a common practice in code generation workflows (see [5]). Eventually, the LLM produces code that compiles successfully. Upon reviewing the final output, it becomes evident that the generated code is more complex, less readable, and incorporates unconventional or less familiar programming techniques. For example, it replaces the `set` and `vector` data structures with more advanced, low-level alternatives:[6]

```
#include <bitset>
...
// Constants for optimization
constexpr size_t BLOCK_SIZE = 64;
...
// Bitset for boolean operations
std::bitset<32768> available;
// Adjust size based on max n_of_vertices
available.set();
// Aligned vector to optimize cache usage
alignas(BLOCK_SIZE) std::vector<int> active_vertices;
active_vertices.reserve(n_of_vertices);
```

Listing 2: Fragment code optimization suggested by LLM.

Meanwhile, the original code is shown in the following listing.

```
set<int> positions;
vector<int> position_of(n_of_vertices, 0);
for (int i = 0; i < n_of_vertices; ++i) {
    positions.insert(i);
    position_of[increasing_degree_order[i]] = i;
}
```

Listing 3: Fragment of original code from CMSA.

The logic remains the same, except for changes in variable names: `positions` is replaced with `available`, and `position_of` is substituted with `active_vertices`.[7]

Although the changes do not affect the heuristic's logic but rather the underlying C++ structures, they do not always lead to measurable improvements in efficiency, specifically in reducing the runtime, which in practice would allow us to explore better candidates when building a valid CMSA solution. Nevertheless, the generated code compiles and runs correctly.

> The two new CMSA variants with these performance improvements will be named: `LLM-CMSA-V1-PERF` and `LLM-CMSA-V2-PERF`.

This iterative process with the LLM, asking it to identify underutilized variables or functions in the existing code, demonstrates its potential as a sophisticated assistant for optimization experts. For instance, in this case, the LLM proposes a new CMSA construction heuristic utilizing the `age` parameter. Unlike simple code generation, the LLM enhances not only the algorithm itself but also the existing code (e.g., C++ implementations), suggesting improvements without altering its logic. This opens the door to using LLMs not just for developing optimization algorithms but also for updating and refining sophisticated legacy code, leveraging the extensive knowledge embedded in LLMs.

C. Reproducibility

Although it is not possible to replicate the exact results of an LLM, due to their autoregressive nature that predicts the most probable token based on a probability distribution (with the next token being determined stochastically) [18], it is possible to reproduce similar responses by using the same prompts, the same LLM, and its parameters. For this reason, our repository (https://github.com/camilochs/optimizing-the-optimizer) includes a *chatbot* that implements the same prompts used in our research (as shown in Figure 3). In fact, each element in Figure 3 (textbox and checkbox) loads pre-built prompts, known as in-context prompts [12, 20], to eliminate the need for manual input. Thus, our chatbot features two types of in-context prompts: (1) external ones, related to the C++ CMSA code for the MIS, and (2) internal ones, focused on improving the heuristic, the C++ code, and specifying which function in the code requires enhancement. Next, we will assess the quality of the heuristics proposed by the LLM.

IV. Empirical Evaluation

This section is divided into two parts: the preliminary phase (setup, benchmark, and CMSA parameter tuning) and the experimental results.

A. Preliminary

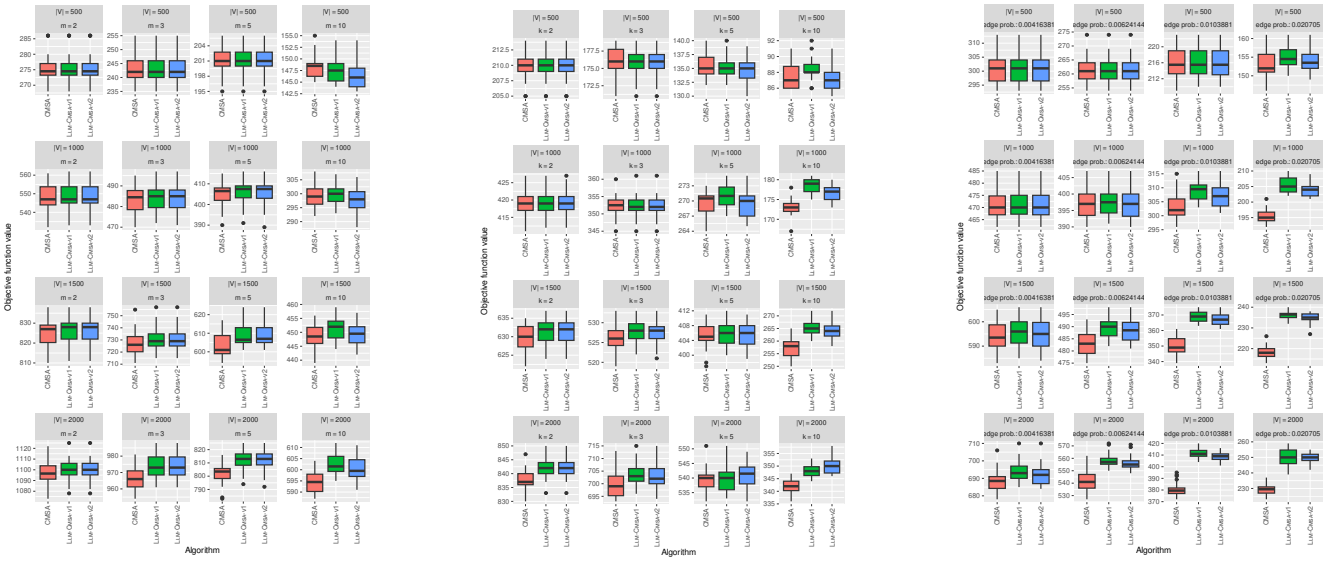
First, we used Chatbot Arena [9] to test various LLMs without incurring costs. [8] For our experiments, we finally selected GPT-4o (version: 2024-11-20)[9] as it is one of the top-performing models to date. Experiments concerning algorithm

[6]The comments in the code were generated by the LLM.

[7]Explicitly instructing the prompt to retain the variable names might have avoided this issue.

[8]https://lmarena.ai/

[9]The parameters used for the LLM were the default values: `temperature = 0.7` (controls randomness in responses), `top-p = 1` (nucleus sampling threshold for token probability), and `max-output-tokens = 2048` (maximum number of tokens in the output).



(a) Barabási–Albert (b) Watts–Strogatz (c) Erdős–Rényi

Fig. 4: Comparative analysis of solution quality: Original CMSA vs. LLM-CMSA variants (V1 and V2).

variants `CMSA`, `LLM-CMSA-V1`, and `LLM-CMSA-V2` were conducted on a cluster equipped with Intel® Xeon® CPU 5670 processors (12 cores at 2.933 GHz) and 32 GB of RAM.

Our benchmark set consists of three types of graphs: Barabási-Albert, Watts-Strogatz, and Erdős-Rényi graphs, with four different sizes and four density levels (see Figure 4). For each combination of size and density level, the benchmark set contains 1 tuning instance and 30 testing instances. This makes a total of 48 tuning instances and 1440 testing instances. We used a limited tuning set to find robust, general parameters, deliberately avoiding the overfitting that would result from tuning on a set identical to our evaluation instances. The parameters of all three CMSA variants (original, V1, V2) were tuned using `irace`, a tool for tuning algorithm parameters based on problem instances [22]. Due to the lack of space (and not being the important point of this paper), we do not report the final values here. We used 150, 300, 450, and 600 CPU seconds as computation time limits for graphs of the four different sizes.

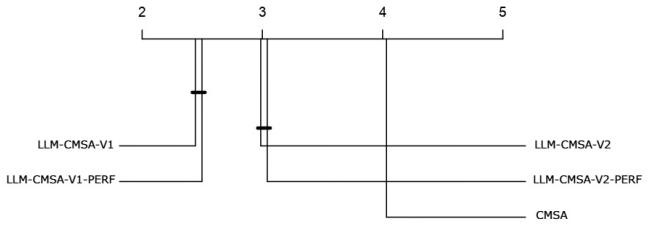
B. Numerical Results

In addition to the three tuned algorithms, we also tested two variants obtained by using the efficiency-improved C++ codes described in Section III-B. These two variants are henceforth called `LLM-CMSA-V1-PERF` and `LLM-CMSA-V2-PERF`. Each of the five algorithms was applied exactly once to each testing instance. The results of the main three variants are reported employing box plots in Figure 4. The considered graph size and density level are indicated at the top of each box plot. The main key observation is that both LLM-generated CMSA variants outperform the standard CMSA variant with growing graph size and density. This clearly shows that the suggestion of the LLM to make use of the age values for solution component selection during solution construction was a very good one.

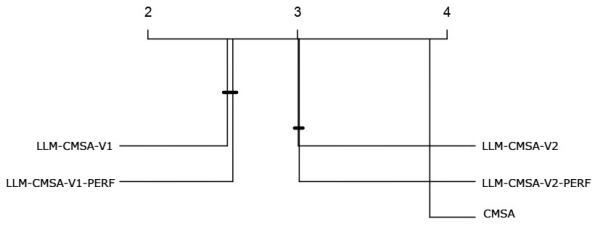
To be able to make statistical claims, we produced so-called critical difference (CD) plots (see Figure 5), in which the algorithms' whiskers show their average ranking concerning a set of problem instances. Moreover, algorithm whiskers are connected by a bold horizontal bar in case they perform statistically equivalent.[10] Figure 5 shows three CD plots, one for each graph type. In these plots we also included the efficiency-optimized LLM-generated CMSA variants. The following observations can be made. Even though both `LLM-CMSA-V1` and `LLM-CMSA-V2` outperform `CMSA` with statistical significance for all three graph types, in all three cases the first variant outperforms the second variant. This means that the idea of using the entropy of the selection probabilities was not fruitful. Moreover, the efficiently-optimized algorithm variants are statistically equivalent to their non-optimized counterparts. This means that, even though they might save RAM, which is a non-tested hypothesis, they do not benefit by producing better results. Furthermore, preliminary checks indicated no significant runtime improvements from these C++ optimizations in our test environment, suggesting the original expert implementation was already highly efficient or that the specific low-level changes were not impactful in this context.

Finally, Figure 6 shows two representative examples of the convergence behavior based on 10 runs of the main three CMSA variants. In both examples, the LLM-generated CMSA

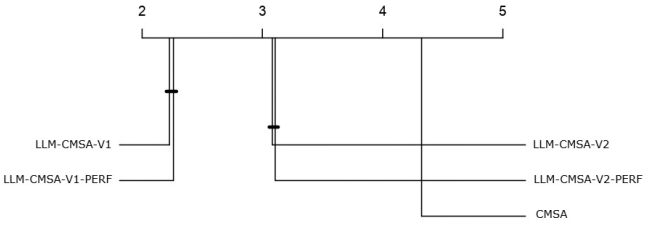
[10]The critical difference (CD) plot, which uses the Friedman test for overall differences and the Nemenyi post-hoc test for pairwise comparisons, is a recognized standard for the statistical comparison of multiple randomized algorithms for combinatorial optimization.



(a) Barabási–Albert

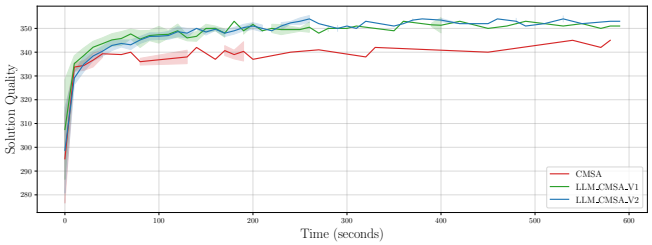


(b) Watts–Strogatz

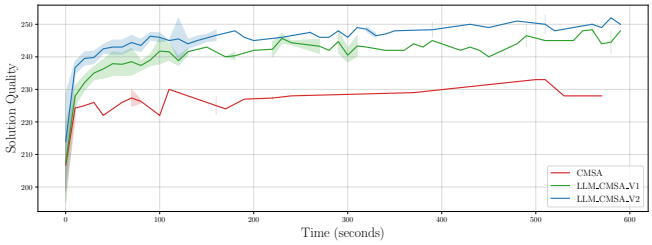


(c) Erdős–Rényi

Fig. 5: Critical difference (CD) plots for all graph types.



(a) Watts–Strogatz (n2000, k10)



(b) Erdős–Rényi (n2000, 020705)

Fig. 6: Examples of algorithm evolution over time.

variants quickly produce solutions of a quality that standard CMSA does not even obtain at the end of its runs.

V. Discussion

Our study highlights the potential for LLMs to engage in algorithmic reasoning, successfully identifying conceptual enhancements for CMSA, such as the `LLM-CMSA-V1` heuristic which outperformed the expert baseline. Interestingly, the LLM's subsequent suggestion incorporating entropy (`LLM-CMSA-V2`), while plausible, was less effective. Perhaps its added complexity was not beneficial, or even detrimental, within the specific CMSA/MIS context, making the simpler LLM-proposed V1 superior. This observation reinforces the need for rigorous empirical validation of any proposed heuristic. However, our study also has the following limitations: due to space constraints, we tested only one LLM (GPT-4o) for CMSA improvements, and future work could benefit from comparing additional LLMs, especially open-weight ones. Additionally, we did not explore using GPT-4o to enhance other complex algorithms for additional optimization problems. These limitations could be addressed in an extended article. Despite these limitations, our comprehensive analysis of CMSA with new heuristics for solving MIS instances presents promising opportunities for new lines of research emerging from our work:

1) **Specialized benchmarks.** While our study focuses on CMSA for the MIS problem, the field lacks benchmarks tailored to optimization. Just as general-purpose code generation relies on standard benchmarks, we need domain-specific ones to evaluate LLMs' ability to discover improved heuristics.
2) **LLM-based agent integration.** Figure 3 illustrates manual interaction, but tasks like execution and debugging could be handled by autonomous agents. A platform where such agents collaborate on improving existing optimization algorithms may yield major advances [15, 21].
3) **Adapt code to the target domain.** Optimization algorithms often require adaptation—not only by switching programming languages to improve performance and maintainability, but also by transitioning from single-threaded execution to parallelism (e.g., using CUDA). LLMs can support this process [25, 19, 8], though domain-specific fine-tuning may be necessary.

An underlying issue in using LLMs for code generation is how we deal with errors. Although they may produce similar outcomes, errors made by a human programmer (or optimization researcher) and those produced by an LLM differ fundamentally in origin. Understanding these conceptual differences is essential for developing sound methodologies for integrating human- and machine-generated code [4].

On the other hand, a fundamental question arises: should the LLM be credited for discovering a superior heuristic compared to the best implementation by a human expert? While prompt design clearly influences the outcome, it raises important questions about authorship, ownership, and the role

of AI in scientific discovery (see [34]).

VI. Conclusions

In our research, we demonstrate that LLMs can be effectively applied to enhance existing optimization algorithms. We used the non-trivial Construct, Merge, Solve, and Adapt (CMSA) algorithm implemented in C++ to solve the classical Maximum Independent Set problem. By leveraging in-context prompts with GPT-4o, the model successfully understood the operational context of the CMSA implementation and proposed conceptually new heuristics for the probabilistic construction phase. After a thorough comparative analysis, the heuristics proposed by the LLM outperformed those manually designed by an expert in CMSA. This highlights the potential of LLMs not merely as tools but as artificial collaborators capable of identifying unused code segments through semantic analysis and contributing meaningfully to the complex task of algorithm design.

Acknowledgment

C. Chacón Sartori and C. Blum were supported by grant PID2022-136787NB-I00 funded by MCIN/AEI/10.13039/501100011033. Moreover, thanks to FreePik, from where we extracted the icons used in Figures 1 and 3.

References

[1] C. Blum. *Construct, Merge, Solve & Adapt: A Hybrid Metaheuristic for Combinatorial Optimization*. Computational Intelligence Methods and Applications. Springer Nature Switzerland, 2024. ISBN 9783031601026. URL https://books.google.es/books?id=ENCt0AEACAAJ.

[2] C. Blum, J. Puchinger, G. R. Raidl, and A. Roli. Hybrid metaheuristics in combinatorial optimization: A survey. *Applied Soft Computing*, 11(6):4135–4151, 2011. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2011.02.032. URL https://www.sciencedirect.com/science/article/pii/S1568494611000962.

[3] C. Blum, P. Pinacho, M. López-Ibáñez, and J. A. Lozano. Construct, merge, solve & adapt a new general algorithm for combinatorial optimization. *Computers & Operations Research*, 68:75–88, 2016. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2015.10.014. URL https://www.sciencedirect.com/science/article/pii/S0305054815002452.

[4] C. Chacón Sartori. Architectures of error: A philosophical inquiry into ai-generated and human-generated code, May 2025. URL https://ssrn.com/abstract=5265751. Available at SSRN.

[5] A. Chen, J. Scheurer, T. Korbak, J. A. Campos, J. S. Chan, S. R. Bowman, K. Cho, and E. Perez. Improving code generation by training with natural language feedback. *arXiv [cs.SE]*, Mar. 2023.

[6] L. Chen, Q. Guo, H. Jia, Z. Zeng, X. Wang, Y. Xu, J. Wu, Y. Wang, Q. Gao, J. Wang, W. Ye, and S. Zhang. A survey on evaluating large language models in code generation tasks, 2024. URL https://arxiv.org/abs/2408.16498.

[7] M. Chen and et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

[8] W. Chen, J. Zhu, Q. Fan, Y. Ma, and A. Zou. Cuda-llm: Llms can write efficient cuda kernels, 2025. URL https://arxiv.org/abs/2506.09092.

[9] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

[10] C. Cummins, V. Seeker, D. Grubisic, M. Elhoushi, Y. Liang, B. Roziere, J. Gehring, F. Gloeckle, K. Hazelwood, G. Synnaeve, and H. Leather. Large language models for compiler optimization. *arXiv [cs.PL]*, Sept. 2023.

[11] DeepSeek-AI et al. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

[12] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, and Z. Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

[13] E. Hemberg, S. Moskal, and U.-M. O'Reilly. Evolving code with a large language model, 2024. URL https://arxiv.org/abs/2401.07102.

[14] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. A survey on large language models for code generation. *arXiv [cs.CL]*, June 2024.

[15] X. Jiang, Y. Dong, L. Wang, Z. Fang, Q. Shang, G. Li, Z. Jin, and W. Jiao. Self-planning code generation with large language models. *ACM Trans. Softw. Eng. Methodol.*, 33(7):1–30, Sept. 2024.

[16] S. Joel, J. J. Wu, and F. H. Fard. A survey on llm-based code generation for low-resource and domain-specific programming languages, 2024. URL https://arxiv.org/abs/2410.03981.

[17] S. Joel, J. J. W. Wu, and F. H. Fard. A survey on LLM-based code generation for low-resource and domain-specific programming languages. *arXiv [cs.SE]*, Oct. 2024.

[18] U. Kamath, K. Keenan, G. Somers, and S. Sorenson. *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Springer Nature Switzerland, 2024. ISBN 9783031656477. URL https://books.google.es/books?id=kDobEQAAQBAJ.

[19] M.-A. Lachaux, B. Roziere, L. Chanussot, and G. Lample. Unsupervised translation of programming languages. *arXiv [cs.CL]*, June 2020.

[20] J. Li, G. Li, C. Tao, J. Li, H. Zhang, F. Liu, and Z. Jin. Large language model-aware in-context learning for code generation. *arXiv [cs.SE]*, Oct. 2023.

[21] F. Liu, R. Zhang, Z. Xie, R. Sun, K. Li, X. Lin, Z. Wang, Z. Lu, and Q. Zhang. Llm4ad: A platform for algorithm design with large language model, 2024. URL https:

//arxiv.org/abs/2412.17287.

[22] M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, T. Stützle, and M. Birattari. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016. doi: 10.1016/j.orp.2016.09.002.

[23] S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, and A. Zeng. Large language models as general pattern machines, 2023. URL https://arxiv.org/abs/2307.04721.

[24] OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[25] R. Pan, A. R. Ibrahimzada, R. Krishna, D. Sankar, L. P. Wassi, M. Merler, B. Sobolev, R. Pavuluri, S. Sinha, and R. Jabbarvand. Lost in translation: A study of bugs introduced by large language models while translating code. *arXiv [cs.SE]*, Aug. 2023.

[26] M. Pluhacek, A. Kazikova, T. Kadavy, A. Viktorin, and R. Senkerik. Leveraging large language models for the generation of novel metaheuristic optimization algorithms. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, page 1812–1820, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701207. doi: 10.1145/3583133.3596401. URL https://doi.org/10.1145/3583133.3596401.

[27] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, J. Grochow, A. Lodi, J.-B. Mouret, T. Ringer, and T. Yu. Mathematical discoveries from program search with large language models. *Nature*, 625:468 – 475, 2023. URL https://api.semanticscholar.org/CorpusID:266223700.

[28] C. C. Sartori, C. Blum, F. Bistaffa, and G. Rodríguez Corominas. Metaheuristics and large language models join forces: Toward an integrated optimization approach. *IEEE Access*, 13:2058–2079, 2025. doi: 10.1109/ACCESS.2024.3524176.

[29] K. Sim, Q. Renau, and E. Hart. Beyond the hype: Benchmarking llm-evolved heuristics for bin packing, 2025. URL https://arxiv.org/abs/2501.11411.

[30] N. v. Stein and T. Bäck. LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics. *IEEE Transactions on Evolutionary Computation*, 29(2):331–345, 2025. doi: 10.1109/TEVC.2024.3497793.

[31] A. Team. Introducing Claude 3.5 Sonnet — anthropic.com. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. [Accessed 02-11-2024].

[32] G. Team and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[33] M. Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[34] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, and Shir. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023. doi: 10.1038/s41586-023-06221-. URL https://ideas.repec.org/a/nat/nature/v620y2023i7972d10.1038_s41586-023-06221-2.html.

[35] Y. Wen, P. Yin, K. Shi, H. Michalewski, S. Chaudhuri, and A. Polozov. Grounding data science code generation with input-output specifications, 2024. URL https://arxiv.org/abs/2402.08073.

[36] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

[37] S. Xiao, Y. Chen, J. Li, L. Chen, L. Sun, and T. Zhou. Prototype2code: End-to-end front-end code generation from ui design prototypes, 2024. URL https://arxiv.org/abs/2405.04975.

[38] D. Zan, B. Chen, F. Zhang, D. Lu, B. Wu, B. Guan, Y. Wang, and J.-G. Lou. Large language models meet NL2Code: A survey, 2023. URL https://arxiv.org/abs/2212.09420.

[39] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers, 2023. URL https://arxiv.org/abs/2211.01910.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 59–72
DOI: 10.15439/2025F6355
ISSN 2300-5963 ACSIS, Vol. 43

Interactive Granular Computing: Toward Computing Model for Complex Intelligent Systems

Andrzej Skowron
Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
Email:skowron@mimuw.edu.pl

Andrzej Jankowski & Soma Dutta
University of Warmia and Mazury in Olsztyn
Słoneczna 54, 10-710 Olsztyn, Poland
Email:andrzej.adgam@gmail.com
soma.dutta@matman.uwm.edu.pl

***Abstract*—We present an approach based on the Interactive Granular Computing (IGrC) model as the basis for developing foundations of Complex Intelligent Systems, *i.e.*, Intelligent Systems dealing with complex phenomena (IS's). The generalization of GrC to IGrC was proposed to support the design of IS's treated in IGrC as examples of complex granules (c-granules) with control. To make such systems successful, it is necessary to enable such systems to have continuous interaction with the physical world. The control of c-granules aims to properly implement the physical semantics of specified transformations of c-granules in the physical world. This implementation is based on the discovery of relevant configurations of physical objects, which provides the basis for perceiving relevant data about these objects and their interactions through the control of c-granules. Additionally, to create high-quality models that serve as the basis for the behavior of IS's, these configurations must be adaptively adjusted by control to allow for the perception of relevant data used to induce those models. Unlike information granules from GrC, the correct implementation of c-granule transformations cannot be restricted to the abstract space. An important property of the IS's discussed here is that they cannot be separated from interactions with the physical world. Hence, they cannot be confined to an abstract space. In particular, the relevance of IGrC in searching for rough computational building blocks for cognition is discussed. These computational building blocks are modeled by complex granules (c-granules) and their networks. It is also proposed to use IGrC as the basis for developing IS's grounded on cognitive computing.**

***Index Terms*—(interactive) granular computing (IGrC); informational-physical complex granule (c-granule); network of c-granules; control of c-granule; reasoning module; rule module; implementational module; decomposition module; rough set; modeling of rough membership function in IGrC.**

I. Introduction

IN THIS paper, we summarize the current status of the IGrC model research based on published papers (see, *e.g.*, [1]–[6] and the developments presented in papers available at https://dblp.org/pid/s/AndrzejSkowron.html) as well as unpublished results, paying special attention to the IS's based on IGrC, the structure of c-granules with control and the role of rough sets in IGrC. We also provide an overview of the potential applications of IGrC-based IS's in various fields. Further details regarding IGrC and such applications will be provided in our future publications.

In the rough set approach, reasoning or decision making strategies are developed based on approximation spaces [7]–[10] which depend on perception about a finite set of objects or situations. For example, based on approximation spaces it is possible to estimate the degree of membership of these objects into given concepts (or classification). This is estimated using the degree of inclusion of some neighborhoods of the considered object (granule) into the considered concepts.The neighborhoods (granules), consist of indiscernible or similar objects to the perceived one. Problems related to how such membership functions are modeled are outside of the existing theory of rough sets. However, for IS's dealing with complex phenomena such an approach is not satisfactory because IS's should be able to construct membership functions based on the current perception of the physical world. As envisaged by many researchers (see, *e.g.*, opinion by Frederick Brooks cited later [11]) if models of the vague concepts related to complex phenomena are constructed using traditional mathematical modeling, then they would lack in quality. Rough membership functions are examples of such models. We propose to support modeling of such concepts basing on Interactive Computing Model (IGrC). In a sense, we are following the discussion on necessity to modify the Turing test by putting into sync not only issues of language and reasoning but also perception and action [12].

In IGrC we follow the suggestion of Professor Leslie Valiant, the Turing award winner. According to his opinion the most important problem of AI is the characterization of the computational building blocks for cognition (see http://people.seas.harvard.edu/~valiant/researchinterests.htm). IGrC is attempting to model these computational building blocks for cognition using dynamic complex granules (c-granules) generated along interactive granular computations. One should note, that many of such computational building blocks are not pure mathematical objects; for instance, they can be specified in natural language by complex vague concepts with physical semantics containing both abstract and physical objects when aspects of perception and action are considered in realization of semantics.

In IGrC, we deal with complex granules (c-granules) com-

posed out of abstract objects and/or physical objects. Such objects are necessary to enable IS's to deal with perception and action issues. IGrC aims to introduce tools for generating and controlling computations over (networks of) complex granules (c-granules)[1] so that the given requirements can be satisfied. These requirements may concern of the final states of computations (specified by some relevant properties of networks of c-granules) or the whole trajectories of computations (specified, *e.g.*, by some invariants over computations on the (relevant parts) of networks of c-granules). These networks are like networks of physical pointers enabling IS's to continuously interact with the real physical world for better perceiving situations in it. They express realization of the physical semantics expressed by specifications of transformations of networks of c-granules. Information about perceived situations in the physical world is stored in i-layers of network of c-granules and is used by control of c-granule for selection of the relevant transformation of the current network of c-granules. The proposal of the IGrC model combines together many different ideas spread nowadays over different domains such as multi-agent systems, robotics, cognitive science, complex adaptive systems, to name only a few of them (see, *e.g.*, [13]–[22]). In the paper, we restrict discussion about control of c-granules as a compound c-granule to an informal description of its behavior. The discussion on a more formal control specification and its physical semantics will be covered in one of our forthcoming paper.

The IGrC model is created as the basis for the design and analysis of c-granules, in particular IS's. The proposed IGrC model differs from the classical Turing model by corroborating the idea of synchronizing the four components: language, reasoning, perception, and action. In the IGrC model, granular computations form the basis for reasoning that supports problem solving by c-granules.

Problem solving (or decision support) using c-granules (in particular, IS's) requires a proper understanding of real-world situations consisting of configurations of interacting physical objects. Therefore, the control of c-granules must include skills for perceiving situations in the physical world enabling the formation of associations between physical and abstract objects. These skills are supported by reasoning over granular computations performed by the control of c-granules. Consequently, these computations cannot be confined to the abstract space alone. Moreover, they depend on physical laws [23].

The paper is structured as follows. In Sect. II we present motivations and rudiments of the IGrC model. Rough complex granules as computational building blocks for cognition are discussed in Sect. III. In Sec. IV the role of rough sets in problem solving by IS's is outlined. The roadmap for developing a rough cognitive computer based on IGrC is outlined in Sec. V. Finally, we present conclusions.

[1]Networks of c-granules can be treated as higher order c-granules.

II. Rudiments of the IGrC model

A. Motivation

There is a huge literature about Granular Computing (GrC) (see, e.g., [1], [24]–[37] and references from these works) encompassing many different issues of GrC concerning. *e.g.*, philosophy, methodology, frameworks, tools, methods, models. However, it is also emphasised in [29]:

> *The current studies of decision-making with GrC lack a theoretical foundation and normative research paradigm.*

The IGrC model is a substantial extension of the GrC model. Granules from GrC can be treated as special cases of c-granules from IGrC[2]. In Table I GrC and IGrC are compared based on several important features.

The IGrC model (see, *e.g.*, [1]–[6] and https://dblp.org/pid/s/AndrzejSkowron.html) is the computing model on which we propose to develop such theoretical foundations for the design and analysis of IS's. Here, one should take opinions of many researchers which emphasizes that in case complex phenomenon classical mathematical modeling is not satisfactory. For example, according to Frederick Brooks [11]:

> *Mathematics and the physical sciences made great strides for three centuries by constructing simplified models of complex phenomena, deriving, properties from the models, and verifying those properties experimentally. This worked because the complexities ignored in the models were not the essential properties of the phenomena. It does not work when the complexities are the essence.*

Hence, one can not expect that humans are able to derive models of the vague concepts related to complex phenomena using traditional mathematical modeling resulting in construction of the high quality models of these concepts. The current discussion on Turing test [12] is also worthwhile mentioning here:

> *The Turing test, as originally conceived, focused on language and reasoning; problems of perception and action were conspicuously absent. The proposed tests will provide an opportunity to bring four important areas of AI research (language, reasoning, perception, and action) back into sync after each has regrettably diverged into a fairly independent area of research.*

According to the above opinions and many other evidences in different areas (see *e.g.*, [6]) construction of models for complex vague concepts related to complex phenomena in the physical world should be based on an unconventional computing model able to deal with abstract and physical objects for providing modeling interactions of IS's with the

[2]Any information granule inf from GrC can be identified as a c-granule g_{inf} consisting of an i-layer with inf and two transformation specifications: $store$ and $read$ (labeled by the addresses pointing out to the physical memory). These specifications may be used by control of the c-granule to either store inf in the p-layer of g_{inf} or read inf from it into the i-layer, without information loss.

TABLE I
COMPARISON OF GRC AND IGRC

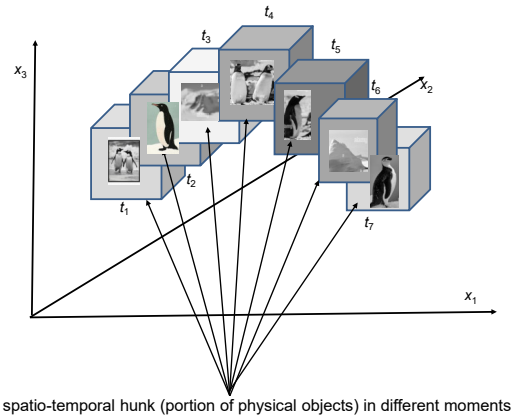
	GrC	IGrC
GENERAL FEATURES OF COMPUTING MODEL		
based on the abstract Turing computation model	YES	NO
computations are pure mathematical objects	YES	NO
issues of language, reasoning, perception and action are brought into sync	NO	YES
advanced reasoning tools based on the computing model that facilitate the control of computations involving both abstract and physical objects are being developed and utilized in the computing model	NO	YES
MAIN FEATURES OF GRANULES		
abstract semantics of (information) granules	YES	YES
physical semantics of (complex) granules	NO	YES
features (attributes) of granules defined using the abstract space only	YES	NO
features (attributes) of granules dependent on interaction with physical objects	NO	YES
granules are equipped with control	NO	YES
the dynamics of granules are defined apriori in the abstract space only	YES	NO
the dynamics of granules depends on their control and interactions with physical objects	NO	YES
associations between abstract and physical objects related to granules are being constructed and used in the computing model	NO	YES
skills for encoding information into physical objects provided in the computing model	NO	YES
skills for decoding information from physical objects provided in the computing model	NO	YES
changes of granules are being made based on the abstract space only	YES	NO
changes of granules are restricted to their abstract parts only	YES	NO
changes of the information represented in granules are being made based on the abstract space only	YES	NO
changes of the information represented in granules are also being made based on interaction with physical objects	NO	YES
STATES, TRANSITION, COMPUTATIONS IN COMPUTING MODEL		
states: granules interacting with physical objects	NO	YES
transition relation (association) defined based on information in the abstract space only	YES	NO
transition relation (association) dependent on interactions with physical objects	NO	YES
computations consist of abstract states only	YES	NO
computations depend on physical laws	NO	YES
adaptation of steering of granular computations provided by control of granules dependent on interaction with physical objects	NO	YES

physical world in the process of perceiving of situations. We propose to base processes of construction of models for such concepts on IGrC.

Novelty of this paper is in (i) further clarifying the role of physical semantics for specifications of transformations of networks of c-granules and issues concerning its generation in the physical world by control of c-granules, (ii) adding discussion on searching for rough computational building blocks (c-granules) over computations of IGrC based on information (decision) systems as well as (iii) emphasizing usefulness of IGrC for developing foundations of cognitive computing by IS's. Our claim is that the foundations of IS's should be based on the unconventional IGrC model.

B. Basic postulates

We assume that IS's are equipped with c-granules which are able to perceive fragments of the physical world (hunks of matter [38]). In this process two layers of c-granules are important, namely the informational layer (i-layer) and the physical layer (p-layer). In the i-layer, spatiotemporal windows (addresses) (see Fig. 1) are specified in a given language; say w, addresses in a sense parts of the physical world included in the regions $\|w\| \subseteq \mathbb{R}^3$ of the physical space defined by w, were $\mathbb{R}$ is the set of reals. Complex granules perceive these parts of the physical world using their control, in particular an implementational module (IM) of control is responsible for defining so called physical semantics of specifications labeling spatio-temporal windows. Roughly speaking, on the basis of a given specification labeling a given spatio-temporal window, the IM constructs a configuration of physical objects and initiates interaction of the c-granule control with and in this configuration. The control of c-granule perceives the properties of this configuration and stores them (as the results of the implementation) in the i-layer. One should note that these results may be different from the expected ones which are given in the specification. In the process of perceiving relevant information about the situation, the control of c-granule may activate or generate a family of c-granules linking the i-layer



Links to a spatio-temporal hunk (portion of physical objects) in different moments (or periods) of time t_1 ,..., t_7 specified by the c-granule control using specifications of spatio-temporal windows pointing to different fragments (portions of matter) of the 3- dimensional physical world.

Fig. 1. Example of spatio-temporal window.

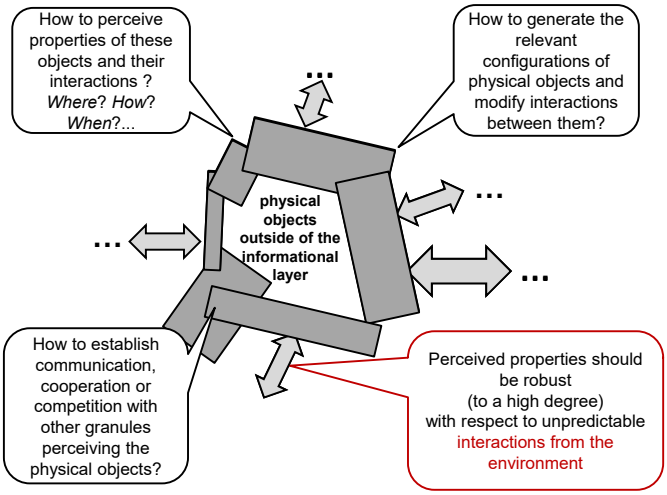


Fig. 2. Queries about a network of physical objects [6].

with the p-layer, creating a network of c-granules, *i.e.*, a higher order c-granule. The c-granules of the network should create a robust configuration with respect to interactions with the environment. The c-granules of such a configuration should be like *niches* considered by John Holland [15]. Better understanding of niches from formal (mathematical) point of view is still a challenge. There are several important issues to be resolved by control of c-granules in construction of networks of c-granules for perceiving situations. Among them a few are as follows (see Fig. 2):

- How to perceive properties of the objects and their interactions in a concerned configuration? (Where, how, when to do this).
- How to generate the relevant configurations of physical objects and modify interactions between them?
- How can cooperation or competition be organized in societies of c-granules to support problem solving?
- How to ensure that the perceived properties are robust (to a high degree) with respect to unpredictable interactions from the environment?

Interaction is a critical issue in the understanding of IS's dealing with complex phenomena. For example, in [39] the following opinion is presented:

> *Interaction is a critical issue in the understanding of complex systems of any sorts: as such, it has emerged in several well-established scientific areas other than computer science, like biology, physics, social and organizational sciences.*

Hence, one of the central issue concerns about interactions which should be treated properly by the computing model as on the basis of this IS's need to be designed and analyzed.

In IGrC model, we assume that interactions take place between configurations of physical objects and they can be partially perceived by c-granules. Three types of interactions of objects are distinguished:

1: interactions between distinguished abstract objects from the i-layer; elaborated in GrC, *e.g.*, for granulation or aggregation of granules (see, *e.g.*, [26], [40]);
2: interactions between some distinguished physical objects from the p-layer; studied in physics and represented in domain knowledge data bases and/or physical laws by the relevant (information) granules in GrC;
3: interactions between some abstract objects from the i-layer and some physical objects from the p-layer; for perceiving situations in the physical world a special interfaces between abstract and p-layers is realized by networks of so called informational-physical complex granules (c-granules, for short).

Example 1: Association in abstract i-layer of c-granule g. Let us assume that $f : X \to Y$, where X, Y are sets and f is a function from X into Y computed by an algorithm A. Computation of values of f are realized by interaction in i-layer of an abstract granule g_{arg} representing arguments of f and the granule g_A representing A. g_A is computing values of f on argguments sent by g_{arg}.

Example 2: Association between abstract i-layer and p-layer of c-granule g

$$f : X \overset{g}{\hookrightarrow} Y,$$

denoting a specification of the association for IM of the control of c-granule g with necessary details for its realization in the physical world and information about the expected results from Y.

Association f is realized using the physical semantics induced by the control of c-granule g responsible for implementation and perception:

- *Implementation step* : For given granules representing $x \in X$ and a specification of f, IM of the control of g extends (by some relevant new c-granules) the current c-granule g

to a new c-granule. This new granule provides 'physical pointers' from the object in which x has been directly encoded to a physical object in o_x, associated to x (by f). The object o_x is pointed out in the p-layer of this new c-granule by a spatio-temporal window w.

- *Perception step*: Some properties of parts of o_x, and other objects from the new c-granule as well as properties of interactions between them (and with the environment) are perceived by IM of the control of g and used for computing the value of f on x. This is done, in particular by decoding information from some parts of directly accessible physical objects of the new constructed granule into i-layer of g or inferring information on the basis of the:
  i. already perceived information and/or
  ii. available knowledge bases or physical laws.

Note that encoding and decoding mentioned above are the basic associations between abstract objects and physical objects 'directly accessible' by the control of g. Associations of abstract objects with more compound physical objects are realized as it is described above using physical objects which are not necessarily 'directly accessible'. Information about associations of abstract and 'not directly accessible' physical objects is inferred by the control of c-granule on the basis of already perceived information stored in the information layer as well as domain knowledge and/or physical laws about the perceived physical objects and their interactions. Hence, computations realized by the control of c-granules depend on domain knowledge and/or physical objects [23]. Moreover, one can see that information about the realized associations between abstract and physical objects is, in general, only partial and relative to the perception skills offered by the control of c-granule.

C. Comments on structure of complex granules (c-granules)

Any c-granule is a dynamic object consisting of, at any moment of local time of c-granule, the following components (see Fig. 3):

- information granules (i-granules) stored in its i-layer: some distinguished (families, collections of) objects from (parts of) the i-layer (considered in GrC); in general any i-granule consists a family of specifications of spatio-temporal windows labeled by perceived or stored information; this information may contain different components, *e.g.*, related to specification of the transformation of the current network of c-granules being under the control of the original c-granule (see Fig 4) or information perceived so far through a particular window;
- physical granules (p-granules) located in its p-layer: some distinguished (families, collections of) objects from (parts of) the p-layer;
- associations between i-granules and p-granules; these associations are specified by formal specifications of transformations and realized as the physical semantics by IM of the control of c-granule through implementation in the physical world; the physical semantics in the form of a network of c-granules under the control of the original c-granule (in a more general case the c-granules from the network may have their own control too) makes it possible to perceive partially the current situation in the physical world; these associations (it should be noted that these associations are not purely mathematical (abstract) objects) satisfy the following conditions:
  1) if p-granule g_p (collection of physical objects) is associated to i-granule g_i then g_i specifies a part of the physical space where g_p is located;
  2) g_i contains information so far perceived about g_p by the control of the original c-granule;
  3) g_i also contains a specification of the goal for realization of which g_p was created in the physical space.

Objects from g_p may be directly accessible by the control of the original c-granule, *i.e.*, information from g_i may be directly encoded into objects from g_p or some features of objects from g_p may be decoded into g_i. Sometimes an object from g_p may not be directly accessible by the control of the original c-granule. Then three distinguished three parts, namely soft_suit, link_suit, and hard_suit of the original c-granule take part in perceiving information (see Fig 3).

In Fig 3 a general structure of c-granule is presented. In soft_suit there are physical objects that are directly accessible, *i.e*, objects whose properties can be decoded into the i-layer by measurements or objects into which some relevant information from the i-layer can be encoded. The physical objects from the link_suit are used for transmission of interactions from soft_suit to hard_suit and hard_suit contains physical objects to be perceived according to the specifications of spatio-temporal windows represented in the i-layer. The scope of c-granule is a specification of spatio-temporal window in which other spatio-temporal windows are included. For a given specification of spatio-temporal window w, by $\|w\|$ we denote a region of the physical space corresponding to w. In a more general case w may represent a nested list of spatio-temporal specifications, *i.e.*, its components may represent new lists of specifications of spatio-temporal windows (Fig 3).

C-granules may have own control (c-granules with control) or can be without control (with empty control). Compound granules such as networks of c-granules or societies of c-granules are constructed from basic (atomic) c-granules. Details of construction of societies will be discussed elsewhere.

It should be noted that

- through interaction of i-granules from g_i with p-granules directly accessible by the control of the considered c-granule from g_p, information represented in i-granules may be encoded into p-granules;
- through interaction of i-granules from g_i with p-granules from g_p directly accessible by the control of the considered c-granule, the encoded information in these p-granules may be transmitted to some target p-granules from g_p;

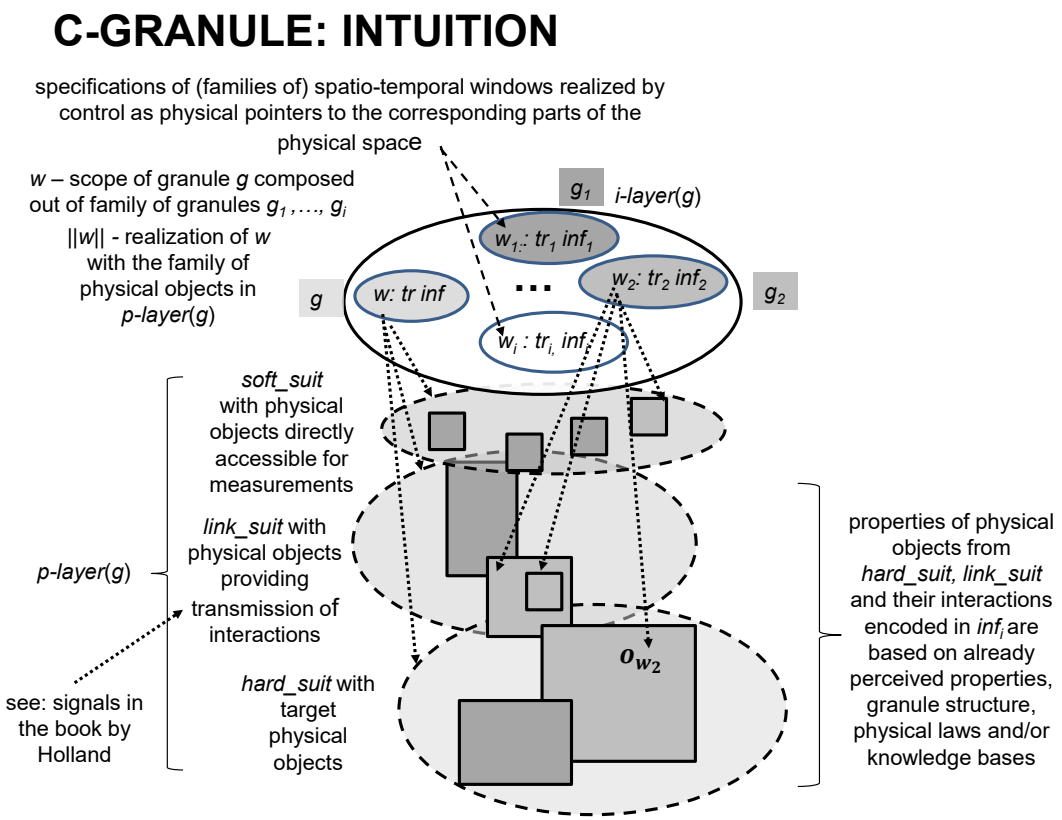


Fig. 3. C-granule intuition.

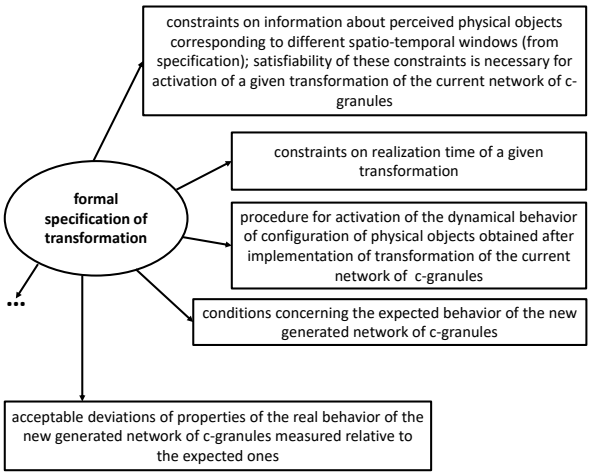


Fig. 4. Information stored in i-layer about transformation of network of c-granules.

- through interaction of i-granules from g_i with p-granules from g_p it may be decoded information from p-granules and stored in i-granules;
- information in the information layer of c-granule may be updated not only through interaction with the directly accessible objects; especially information about objects from g_p may be updated on the basis of reasoning using already collected information and also domain knowledge bases and/or physical laws.

D. Dynamics of complex granules (c-granules)

Dynamics of c-granules is defined by control of c-granule. Formally, the control may be understood as a distinguished c-

granule with the i-layer, which describes the control behavior determined by the p-layer. An intuition of control is presented in Fig. 5.

Control organizes (and initiates) communications (transmissions, interactions) between the i-layer and p-layer using network (configuration) of relevant c-granules (generated by its IM). The considered network collects in the i-layers the properties of perceived physical objects and their interactions.

The control of C-granules includes different modules that act as sub-granules of the control granule. These sub-granules interact with each other and with the environment to carry out target tasks (needs). Among them the basic ones are the following:

- cycle module (CM);
- rule module (RM);
- implementational module (IM);
- decomposition module (DM);
- reasoning module (ResM).

More advanced control may contain many other modules such as a learning module (LM), adaptation module (AM), or domain knowledge module (DKM).

The CM is responsible for the realization of the basic cycle of control of the original c-granule.

The RM consists a set of rules of the form

$$\alpha \Rightarrow_{tr} \beta,$$

where

- α is a condition of the rule verified on information about the status of perception of the current situation,
- tr is a specification of transformation of the current network of c-granules; this specification describes which c-granules in the network should be eliminated, suspended, modified, not changed etc[3] and
- β is the property of the expected result of realization.

Thus, the set of rules should be designed by the designer or learned from data.

The IM is a module responsible for generation of the physical semantics for a given specification of transformation (see Fig. 6). IM first tests the possibility of direct realization of a given specification in the physical world. If this is not possible, DM requires decomposition (see Fig. 7) of this specification, otherwise IM creates the relevant configuration of physical objects and initiates interactions in it aiming to realize the specification in the physical world.

More formal description of the DM is provided in [41]. It should be noted that the development of the DM module requires discovery of several decomposition languages. This discovery process can be supported by chatbots [42] helping to learn the strategy of divide-and-conquer of complex vague concepts (expressed in natural language) in human problem-solving. The key role in this process is played by information granulation [43], [44].

[3]More details on specification of transformations will be presented in our next paper.

The ResM is an important module supporting reasoning of the control of c-granules. For example, it allows control to decide if the current status of perception of the situation is satisfactory for selection of a rule from a given set of rules or to decide if the deviation of the expected results expressed in the rule from the real results obtained after implementation of a given transformation is acceptable. When such deviations are not acceptable, ResM may support adaptation of the current set of rules.

In Figures 8 - 9 is presented an illustrative example related to the ResM structure in one of our recent projects.

Control of c-granule aims to provide the most relevant decision systems for approximation of concepts. Information (decision) systems are represented in i-layer of c-granule; they are not isolated and not a priori given but they are modified by control of c-granule and (indirectly by) through interactions with the environment. Objects in information (decision) systems are (fragments of) multiple and/or multivariate time series represented in i-layer of c-granule obtained as a result of perceiving the situation in the physical world along computations of c-granule over networks of c-granules. Attributes represent properties of such objects. There is a necessity of providing possibility to change the currently used attributes during computation for proper identification of perceived situation (making it possible to select the relevant transformations for realization). Decisions are elements of complex game, *i.e.*, pairs consisting of complex vague concepts labeled by specification of transformation (*e.g.*, decision (plan)) related to this concept. Complex games are discovered from such decision tables. Development of strategies for discovery of rough membership functions for different complex vague concepts are necessary.

Hence here is a necessity to develop different reasoning (judgment) methods supporting, *e.g.*,

- discovery of complex games and their adaptation (reasoning about changes - rough calculus);
- identification of the relevant properties of situations in the physical world;
- control of computations over networks of c-granules toward generating computations satisfying a given specification; this may be related to the whole computation or to its final state;
- estimation of rough membership degrees of the perceived in the physical world situations to the considered concepts;
- resolving conflicts between rules specifying transformations to be performed;
- discovery of new sources of the relevant for the considered problem data (data governance [45]);
- discovery of compound sensors and/or actuators, robots.

Let us also emphasize the important role of reasoning methods in supporting discovery of relevant rough computational building blocks for cognition (rc-granules in IGrC) or aggregation of decisions (*e.g.*, in synthesis of complex sensors, materials or robots). One should note that there are several 'white spots' in formal reasoning methods waiting to

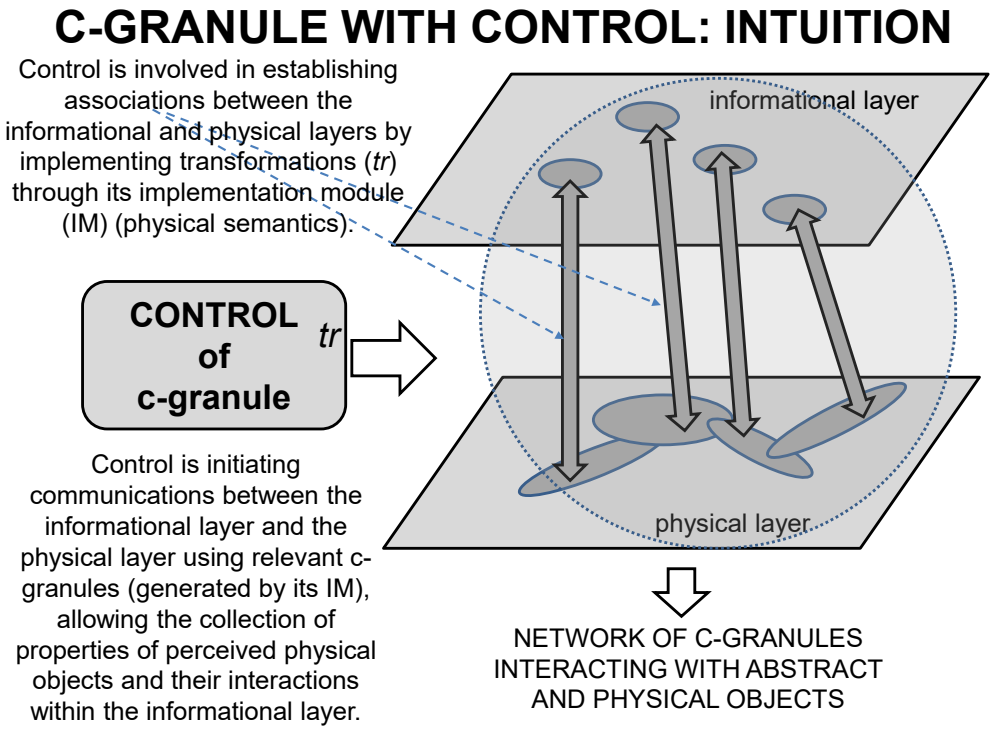


Fig. 5. C-granule with control: intuition.

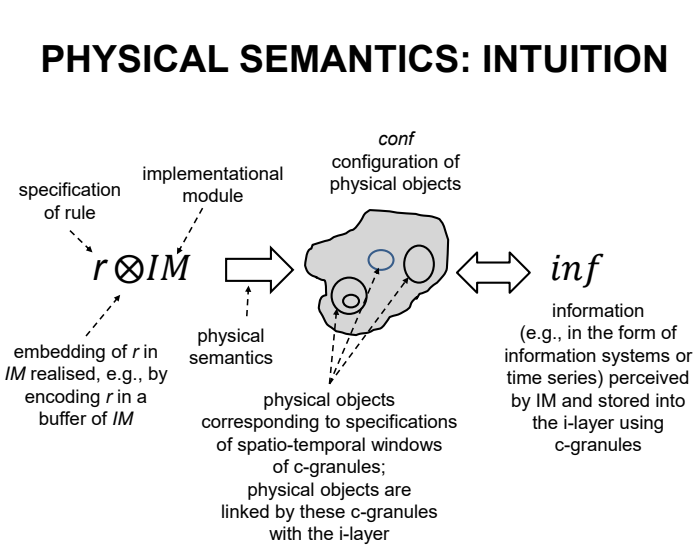


Fig. 6. Physical semantics: intuition.

be filled by the relevant methods. One example of such 'white spot' concerns experience based reasoning (see *e.g.*, [46], [47]). Hence, is is necessary to develop reasoning methods for supporting dialogues of IS's with experts to help them in performing of experience based reasoning [48], [49].

Control of c-granule in interaction with the physical world generates computations over networks of c-granules. The aim of the control is to assure whether the generated computations satisfy given requirements or target goals. They may concern the final states of such computations or the whole trajectories of computations (*e.g.*, in case of required invariants to be preserved in computations). One should note that these

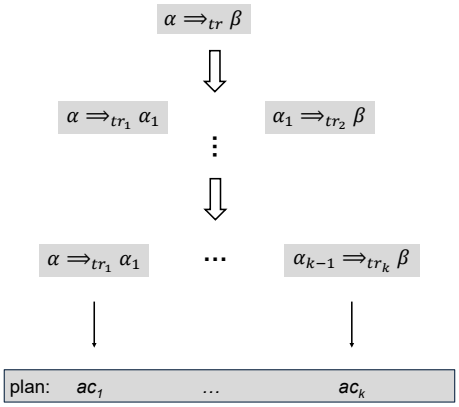


Fig. 7. Decomposition.

computations are not pure mathematical objects, they are unconventional because they are constructed over abstract and physical objects. In particular, they may depend on domain knowledge or physical laws [23].

III. Discovery of rough complex granules (RC-granules) as computational building blocks for cognition and the language of game of Wittgenstein

As it was observed in Sect. I, nowadays rough membership functions are defined as pure mathematical objects. However, according to the cited opinion by Frederick Brooks [11] models of these functions, being often models of vague concepts related to complex phenomena, can't be constructed (induced) using traditional mathematical modeling. Hence, existing approaches to modeling of rough membership functions should be enriched by mechanisms for continuous interaction with the physical real world for perception of the current situation making it possible to estimate adaptively membership values. In the perceptual approach to rough sets in IGrC, rough membership value for the currently perceived situation in the real physical world is estimated on the basis of perceived data by (networks of) c-granules dynamically interacting with the real physical world up to the moment when understanding of the perceived situation is satisfactory for making this estimation by control of c-granule.

One should note that the approach based on IGrC is consistent with the concept of language-game by Ludwig Wittgenstein [50] who argued that a word or even a sentence has meaning in the language only as a result of the *rule of the game* being played. In the case of IGrC, the meaning as use (https://plato.stanford.edu/entries/wittgenstein/#MeanUse) is realized as physical semantics *i.e.*, is realized in the physical world and its results are partially perceived by c-granules.

In the discussed IGrC-based approach, computational building blocks for concept cognition (see http://people.seas.harvard.edu/~valiant/researchinterests.htm) are constructed alongside computations over granular networks generated by the control of c-granules interacting with the environment. The control of c-granules is responsible for discovering computational building blocks via sensors or actuators, as well as aggregating pre-existing blocks. Searching for the relevant computational building blocks on different levels of hierarchical modeling is supported by strategies with backtracking, in case the currently searching 'path' occurs to be not promising. Control of c-granules aims to search for the relevant computational building blocks, *e.g.*, patterns, classifiers or clusters from which the high quality membership functions of the target complex vague concepts can be discovered. These target concepts are responsible for triggering the actions or plans on the highest level of the hierarchical modeling. Hierarchical learning of computational building blocks is based on aggregation of information (decision) systems: the Cartesian product of object descriptions from aggregated information systems [6], [51], [52] is filtered by constraints defined with the use of discovered relational systems. Relations from these systems are defined over products of value sets of attributes from the considered systems.

C-granule control also employs searching strategies for approximate reasoning schemes that link different levels of the concept hierarchy [40].

One should also note an important role of reasoning strategies for conflict resolution between different computational building blocks represented by rough membership functions matching the current situation in the physical world and voting for different target decisions.

By changing the language in which computational building blocks are represented one can obtain the approach based on rough sets or on combination of rough and fuzzy sets for discovery of relevant computational building blocks.

IV. Note on rough sets in approximate problem solving by IS's

In this section, we outline the role of rough sets in approximate problem solving by IS's. The presented discussion on IGrC model makes it possible to generalize the existing approaches to approximation of concepts to a more general case of approximation of granules (*e.g.*, construction of complex abstract or physical objects satisfying a given specification to a satisfactory degree or construction of the high quality of classifiers from training sets). The key idea of approximation of granules can be intuitively expressed in the following way:

> *For a given specification of problem to be solved, the control of a given c-granule (or society of granule) is aiming to generate, in interaction with the environment, a computation on granular networks satisfying this specification to a satisfactory degree.*

The formalization concerning this understanding of approximation will be discussed in more detail in our next paper. Here, we would like to mention that this formalizations would clarify in particular, the following issues:

- problem specification is encoded into the initial granular network of computation,

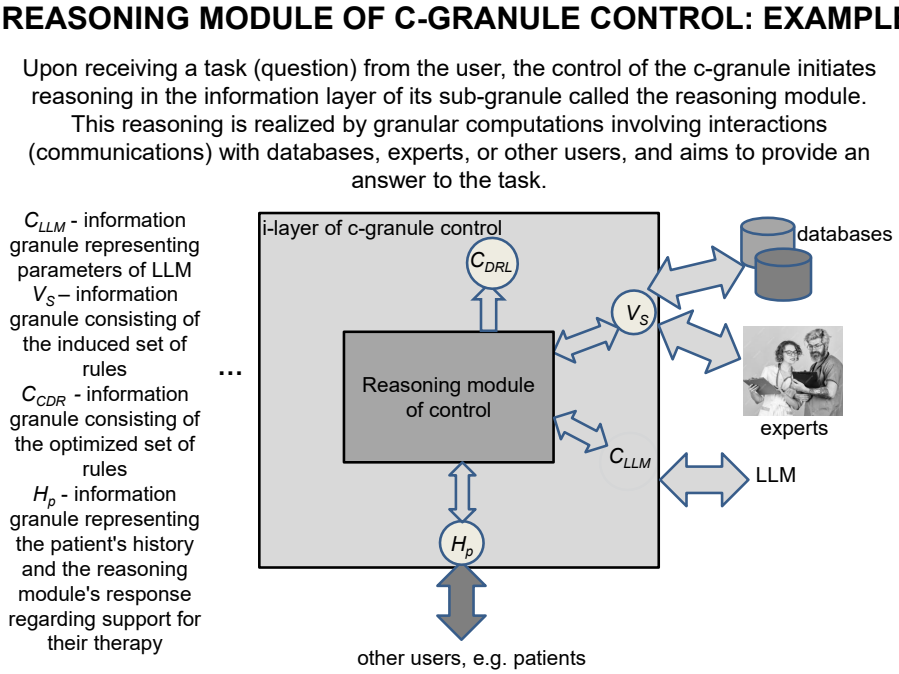


Fig. 8. Example concerning the c-granule representing the ResM of c-granule control.

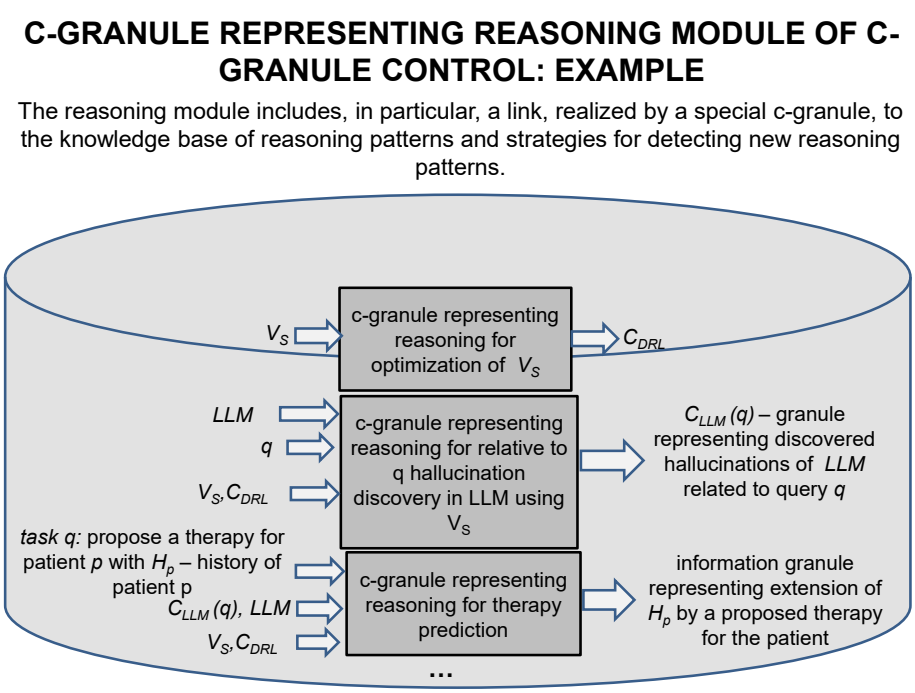


Fig. 9. Example concerning the c-granule in reasoning module that represents a knowledge base with reasoning patterns.

- solutions of the considered problem are expressed by granules (*e.g.*, when the task is to learn classifiers or compound physical or chemical objects),
- different fragments of granular networks are expressed using different languages and linked by interfaces (*e.g.*, based on aggregation operations between granules),
- successive steps of computation are obtained by control of c-granule interacting with the physical world (*e.g.*,

by generating more compound granular networks from existing ones or extending the existing ones by granules resulting from the relevant interactions with the physical world or domain knowledge bases),
- selection of rule for transformation of the current granular network is performed,
- the rule set for transformation of granular networks is adaptively learned,
- control can search for new data sets in the physical world for updating granular networks (by reasoning, *e.g.*, when, where to search for such data),
- different components of approximation spaces are distinguished to make it possible to define the quality of approximation of granules,
- selection of optimization procedures for approximation is performed among many possible ones,
- control of c-granule is designed to make it possible to generate computation of the required quality (*e.g.*, by optimization and adaptation of its behavior),
- quality of computation is evaluated by the relevant quality measures (*e.g.*, some of them may use the whole computation for evaluation when the task is to preserve some invariants during computation and some other ones depend only on the final granular network in the computation),
- quality of generated approximate solution is tested in the physical world.

In this generalized approach to the rough set-based approximation of granules, reasoning performed by the control of c-granules over the generated granular computations plays a crucial role. As we discussed above, reasoning supporting approximation of granules is performed along computations by control of c-granule.

Fig. 10 presents a challenge for rough sets in IGrC related to discovery of approximate solutions of problems. Note that in this case, there are no provided examples belonging to the lower approximation of the concept *granular computation with satisfactory quality relative to the specified task*[4]. However, such examples should be discovered. There are numerous domains for which this challenge is important. Among them (i) discovery of learning algorithms and construction of classifiers [13], [53], (ii) automatic design of robots [54], (iii) drug discovery [55], (iv) algorithmic trading [56], (v) generative AI [57] are a few. One could request the development of common foundations based on IGrC and RS for designing and analyzing discovery systems from these different domains [58].

V. Roadmap for developing a rough cognitive computer based on IGrC

Human-inspired computing is defined as the intelligence computing model enlightened by human brain intelligence and biological processes. We propose to take (networks of) c-granules as the basic ingredients of such a model. They seem to be the relevant in modeling basic and higher order neurons. They have richer structure then traditional artificial neurons and provide a more suitable objects for modeling of cognitive computer based on IGrC than the traditional artificial neurons. For example, on different levels of hierarchical modeling c-granules with different languages for expressing perceived properties may be used. Certainly, here arises a challenge how to discover such relevant languages. Moreover, they may have sophisticated control mechanisms. These control mechanisms should make it possible to adaptively change the networks toward satisfying the target goals (or needs of granules or agents). Hence, in this way we may obtain the approach for modeling cognitive computers based on generalization of neural networks to adaptive granular networks. In the future, we will investigate dynamical structures related to granulation of c-granules (*e.g.*, into networks) and their degranulation processes aiming at searching for other c-granules understood as computational building blocks for cognition, using the Valiant formulation. These computational building blocks can be used, *e.g.*, to approximation of complex vague concepts triggering realization of transformations' specifications (*e.g.*, in the form actions or plans) by control of c-granules. For modeling of cognitive computer societies of networks of c-granules will be used following the idea of distributed control experience presented in the literature (see, *e.g.*, [58]–[64]). Further cooperation with neuroscientists [65] may substantially enhance the development strategies for modeling such dynamical structures.

Issues of multiscaling reasoning inspired by biological processes are of great importance for further developing IGrC. Here, it is worthwhile to cite the following statement from [66]:

> [...] *One of the fascinating goals of natural computing is to understand, in terms of information processing, the functioning of a living cell. An important step in this direction is understanding of interactions between biochemical reactions.* [...] *the functioning of a living cell is determined by interactions of a huge number of biochemical reactions that take place in living cells.*

VI. Conclusions

In this paper we have presented the IGrC model as the basis for the design and analysis of Complex Intelligent Systems (IS's). In this way, we have taken a step toward developing the theoretical foundation for this task. The IGrC is not purely mathematical and it substantially depends on continuous interaction with the physical world aiming to generate the right physical semantics. The control module of c-granules steers computations over granular networks toward achieving target goals or needs. This process involves interacting with numerous other c-granules based on physical objects, experts, users, knowledge bases, and physical laws. The presented approach can be used as foundations for developing IS's in different domains, including AI's discovering learning algorithms, specific robots, new drugs, strategies in algorithic trading or models of cognitive computers.

[4]The lower approximation may be defined, *e.g.*, by specifying 'satisfactory quality' by the relevant threshold.

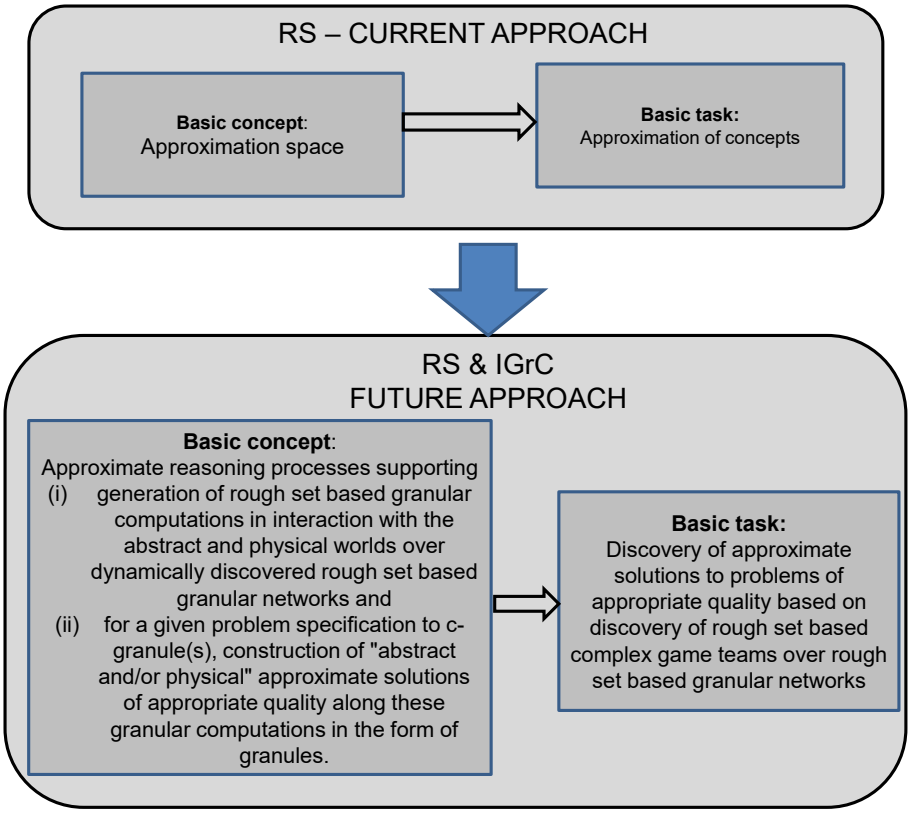


Fig. 10. A challenge for RS: discovery of approximate solutions of problems.

Our aim is to build foundations for IS's realizing the following dream [67]:

Tomorrow, I believe, we will use

[INTELLIGENT SYSTEMS]

to support our decisions in defining our research strategy and specific aims, in managing our experiments, in collecting our results, interpreting our data, in incorporating the findings of others, in disseminating our observations, in extending (generalizing) our experimental observations through exploratory discovery and modeling - in directions completely unanticipated.

REFERENCES

[1] A. Skowron, A. Jankowski, and S. Dutta, "Interactive granular computing," *Granular Computing*, vol. 1, pp. 95–113, 2016. [Online]. Available: doi.org/10.1007/s41066-015-0002-1

[2] A. Skowron and A. Jankowski, "Rough Sets and Interactive Granular Computing," *Fundamenta Informaticae*, vol. 147, no. 2-3, pp. 371–385, 2016. [Online]. Available: doi.org/10.3233/FI-2016-1413

[3] S. Dutta and A. Skowron, "Interactive granular computing model for intelligent systems," in *Intelligence Science III: 4th IFIP TC 12 International Conference, ICIS 2020, Durgapur, India, February 24-27, 2021, Revised Selected Papers*, ser. IFIP Advances in Information and Communication Technology, Z. Shi, M. K. Chakraborty, and S. Kar, Eds., vol. 623. Springer International Publishing, 2020, pp. 37–48. [Online]. Available: doi.org/10.1007/978-3-030-74826-5_4

[4] ——, "Toward a computing model dealing with complex phenomena: Interactive granular computing," in *Computational Collective Intelligence - 13th International Conference, ICCCI 2021, Rhodes, Greece, September 29 - October 1, 2021, Proceedings*, ser. Lecture Notes in Computer Science, N. T. Nguyen, L. Iliadis, I. Maglogiannis, and B. Trawinski, Eds., vol. 12876. Springer, 2021, pp. 199–214. [Online]. Available: doi.org/10.1007/978-3-030-88081-1_15

[5] A. Jankowski, *Interactive Granular Computations in Networks and Systems Engineering: A Practical Perspective*. Heidelberg: Springer, 2017. [Online]. Available: doi.org/10.1007/978-3-319-57627-5

[6] A. Skowron and D. Ślęzak, "Rough sets in interactive granular computing: Toward foundations for intelligent systems interacting with human experts and complex phenomena," in *Proceedings of the International Joint Conference on Rough Sets (IJCRS 2023), Cracow, October 5-8, 2023*, ser. Lecture Notes in Computer Science, vol. 14481. Berlin, Heidelberg: Springer, 2023, pp. xv–xxxvii. [Online]. Available: doi.org/10.1007/978-3-031-50959-9

[7] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, pp. 341–356, 1982. [Online]. Available: doi.org/10.1007/BF01001956

[8] ——, *Rough Sets: Theoretical Aspects of Reasoning about Data*, ser. System Theory, Knowledge Engineering and Problem Solving. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1991, vol. 9. [Online]. Available: doi.org/10.1007/978-94-011-3534-4

[9] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, vol. 177, no. 1, pp. 3–27, 2007. [Online]. Available: doi.org/10.1016/j.ins.2006.06.003

[10] ——, "Rough sets: Some extensions," *Information Sciences*, vol. 177, no. 28-40, p. 1, 2007. [Online]. Available: doi.org/10.1016/j.ins.2006.06.006

[11] F. P. Brooks, Jr., *The Mythical Man-Month: Essays on Software Engineering, Anniversary Edition*. Addison-Wesley, 1995.

[12] C. L. Ortitz Jr., "Why we need a physically embodied Turing test and what it might look like," *AI Magazine*, vol. 37, pp. 55–62, 2016. [Online]. Available: doi.org/10.1609/aimag.v37i1.2645

[13] S. J. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*. Upper Saddle River, NJ: Pearson Education, Inc., 2021, 4th edition.

[14] Y. Bar-Yam, *Dynamics of Complex Systems. Studies in Nonlinearity*. Reading, Massachusetts: Addison-Wesley, 1997.

[15] J. H. Holland, *Signals and Boundaries. Building Blocks for Complex Adaptive Systems*. Cambridge, MA: The MIT Press, 2014. [Online]. Available: doi.org/10.7551/mitpress/9412.001.0001

[16] J. Wąs, *Modeling and Simulation of Complex Collective Systems*. Boca Raton, Fl: CRC Press, 2023. [Online]. Available: doi.org/10.1201/b23388

[17] M. Gell-Mann, *The Quark and the Jaguar - Adventures in the Simple and the Complex*. London: Brown & Co., 1994.

[18] A. G. Hoekstra, J.Kroc, and P. M. Sloot, Eds., *Simulating Complex Systems by Cellular Automata*. Heidelberg: Springer, 2010. [Online]. Available: doi.org/10.1007/978-3-642-12203-3

[19] D. Marr, *Vision*. Cambridge MA: The MIT Press, 2010.

[20] M. Miłkowski, *Explaining the Computational Mind*. Cambridge MA: The MIT Press, 2013.

[21] Z. W. Pylyshyn, *Computation and Cognition. Toward a Foundation for Cognitive Science*. Cambridge MA: The MIT Press, 1984.

[22] C. Gershenson, Ed., *Design and Control of Self-organizing Systems*. Mexico City, Boston: CopIt ArXives, 2007.

[23] D. Deutsch, E. Ekert, and R. Lupacchini, "Machines, logic and quantum physics," *Neural Computation*, vol. 6, pp. 265–283, 2000. [Online]. Available: doi.org/10.2307/421056

[24] A. Bargiela and W. Pedrycz, *Granular Computing. An Introduction*, ser. The Springer International Series in Engineering and Computer Science (SECS). New York, NY: Springer, 2003, vol. 717. [Online]. Available: 10.1007/978-1-4615-1033-8

[25] D. Ciucci and Y. Yao, "Synergy of granular computing, shadowed sets, and three-way decisions," *Information Sciences*, vol. 508, pp. 422–425, 2020. [Online]. Available: 10.1016/j.ins.2019.09.003

[26] W. Pedrycz, *Granular Computing. Analysis and Design of Intelligent Systems*. CRC Press, Taylor & Francis, 2013. [Online]. Available: doi.org/10.1201/9781315216737

[27] T. Jing, C. Wang, W. Pedrycz, Z. Li, G. Succi, and M. Zhou, "Granular models as networks of associations of information granules: A development scheme via augmented principle of justifiable granularity," *Applied Soft Computing*, vol. 115, p. 108062, 2022.

[28] W. Pedrycz, A. Skowron, and V. Kreinovich, Eds., *Handbook of Granular Computing*. Hoboken, NJ: John Wiley & Sons, 2008. [Online]. Available: doi.org/10.1002/9780470724163

[29] J. Qin, L. Martínez, W. Pedrycz, X. Ma, and Y. Liang, "An overview of granular computing in decision-making: Extensions, applications, and challenges," *Information Fusion*, vol. 98, p. 101833, 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2023.101833

[30] A. Skowron and S. Dutta, "Rough sets and fuzzy sets in interactive granular computing," in *Rough Sets - International Joint Conference, IJCRS 2022, Suzhou, China, November 11-14, 2022, Proceedings*, ser. Lecture Notes in Computer Science, J. Yao, H. Fujita, X. Yue, D. Miao, J. W. Grzymala-Busse, and F. Li, Eds., vol. 13633. Springer, 2022, pp. 19–29. [Online]. Available: doi.org/10.1007/978-3-031-21244-4_2

[31] G. Wang, "Data-driven granular cognitive computing," in *Rough Sets - International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3-7, 2017, Proceedings, Part I*, ser. Lecture Notes in Computer Science, L. Polkowski, Y. Yao, P. Artiemjew, D. Ciucci, D. Liu, D. Slezak, and B. Zielosko, Eds., vol. 10313. Springer, 2017, pp. 13–24. [Online]. Available: doi.org/10.1007/978-3-319-60837-2_2

[32] J. Yao, "A ten-year review of granular computing," in *2007 IEEE International Conference on Granular Computing (GrC 2007), San Jose, California, USA, 2-4 November 2007*. IEEE Computer Society, 2007, pp. 734–739. [Online]. Available: doi.org/10.1109/GrC.2007.11

[33] J. T. Yao, A. V. Vasilakos, and W. Pedrycz, "Granular computing: Perspectives and challenges," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1977–1989, 2013. [Online]. Available: doi.org/10.1109/TSMCC.2012.2236648

[34] J. Yao, Y. Yao, D. Ciucci, and K. Huang, "Granular computing and three-way decisions for cognitive analytics," *Cognitive Computing*, vol. 14, no. 6, pp. 1801–1804, 2022. [Online]. Available: doi.org/10.1007/s12559-022-10028-0

[35] Y. Yao, "The art of granular computing," in *Rough Sets and Intelligent Systems Paradigms (RSEISP 2007)*, ser. Lecture Notes in Computer Science, M. Kryszkiewicz, J. F. Peters, H. Rybinski, and A. Skowron, Eds., vol. 4585. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 101–112. [Online]. Available: doi.org/10.1007/978-3-540-73451-2_12

[36] L. Zadeh, "Fuzzy logic = computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 103–111, 1996. [Online]. Available: doi.org/10.1109/91.493904

[37] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, pp. 111–127, 1997. [Online]. Available: doi.org/10.1016/S0165-0114(97)00077-8

[38] M. Heller, *The Ontology of Physical Objects. Four Dimensional Hunks of Matter*, ser. Cambridge Studies in Philosophy. Cambridge, UK: Cambridge University Press, 1990.

[39] A. Omicini, A. Ricci, and M. Viroli, "The multidisciplinary patterns of interaction from sciences to computer science," in *Interactive Computation: The New Paradigme*, D. Goldin, S. Smolka, and P. Wegner, Eds. Heidelberg: Springer, 2006, pp. 395–414. [Online]. Available: doi.org/10.1007/3-540-34874-3_15

[40] A. Skowron and J. Stepaniuk, "Information granules and rough-neural computing," in *Rough-Neural Computing: Techniques for Computing with Words*, ser. Cognitive Technologies, S. K. Pal, L. Polkowski, and A. Skowron, Eds. Heidelberg: Springer-Verlag, 2004, pp. 43–84. [Online]. Available: doi.org/10.1007/978-3-642-18859-6_3

[41] S. Dutta, A. Skowron, and L. Sosnowski, "Medical decision support in the light of interactive granular computing: Lessons from the ovufriend project," *International Journal of Approximate Reasoning*, vol. 165, p. 109103, 2024. [Online]. Available: doi.org/10.1016/J.IJAR.2023.109103

[42] R. Ciesla, *The Book of Chatbots. From ELIZA to ChatGPT*. Cham: Springer, 2024. [Online]. Available: doi.org/10.1007/978-3-031-51004-5

[43] L. A. Zadeh, "Foreword," in *Rough-Neural Computing: Techniques for Computing with Words*, S. K. Pal, L. Polkowski, and A. Skowron, Eds. Heidelberg: Springer, 2004, pp. ix–xi. [Online]. Available: oi.org/10.1007/978-3-642-18859-6

[44] ——, *Computing with Words: Principal Concepts and Ideas*, ser. Studies in Fuzziness and Soft Computing. Heidelberg: Springer, 2012, vol. 277.

[45] E. Eryurek, U. Gilad, V. Lakshmanan, A. Kibunguchy-Grant, and J. Ashdown, *Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness*. Sebastopol, CA: O'Reilly Media, 2021.

[46] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*. New York, NY: Farrar, Straus, Giroux, 2019.

[47] ——, "Abstraction and analogy-making in Artificial Intelligence," *Annals Reports of the New York Academy of Sciences*, vol. 1505, pp. 79–101, 2021. [Online]. Available: doi.org/10.1111/nyas.14619

[48] B. Shneiderman, *Human – Centered AI*. Oxford, UK: Oxford University Press, 2022. [Online]. Available: doi.org/10.1093/oso/9780192845290.001.0001

[49] R. M. Monarch, *Human-in-the-Loop Machine Learning. Active Learning and Annotation for Human-Centered AI*. Shelter Island, NY: MANNING, 2021.

[50] L. Wittgenstein, *Philosophical Investigations (translated by G.E.M. Anscombe)*. New York: Macmillan Co., 1953.

[51] Z. Pawlak, "Information systems – theoretical foundations," *Information Systems*, vol. 6, pp. 205–218, 1981. [Online]. Available: doi.org/10.1016/0306-4379(81)90023-5

[52] A. Skowron and D. Ślęzak, "Rough sets turn 40: From information systems to intelligent systems," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30, 2022, pp. 23–34. [Online]. Available: doi.org/10.15439/2022F310

[53] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Heidelberg: Springer-Verlag, 2001. [Online]. Available: doi.org/10.1007/978-0-387-84858-7

[54] D. Matthews, A. Spielberg, D. Rus, S. Kriegman, and J. Bongard, "Efficient automatic design of robots," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 120, pp. 1–7, 2023. [Online]. Available: doi.org/10.1073/pnas.2305180120

[55] N. Brown, Ed., *Artificial Intelligence in Drug Discovery*, ser. Drug Discovery Series. London: Royal Society of Chemistry, 2021, vol. 75.

[56] V. Dolzhenko, *Algorithmic Trading Systems and Strategies: A New Approach: Design, Build, and Maintain an Effective Strategy Search Mechanism*. New York, NY: Apress Media, 2024. [Online]. Available: doi.org/10.1007/979-8-8688-0357-4
[57] R. Gupta, S. Tiwari, and P. Chaudhary, *Generative AI: Techniques, Models and Applications*. Cham: Springer Nature, 2025. [Online]. Available: doi.org/10.1007/978-3-031-82062-5
[58] S. Gerrish, *How Smart Machines Think*. Cambridge, MA: MIT Press, 2018. [Online]. Available: doi.org/10.7551/mitpress/11440.001.0001
[59] C. Gershenson, *Design and Control of Self-organizing Systems, PhD Thesis*. Brussel: Vrije Universiteit, 2007. [Online]. Available: https://web-archive.southampton.ac.uk/cogprints.org/5442/1/thesis.pdf
[60] ——, "Guiding the self-organization of cyber-physical systems," *Frontiers Robotics AI*, vol. 7, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.3389/frobt.2020.00041
[61] J. Hurwitz, M. Kaufman, and A. Bowles, *Cognitive Computing and Big Data Analytics*. Indianapolis, IN: John Wiley & Sons, Inc., 2015.
[62] J. M. Mendel, "The perceptual computer: an architecture for computing with words," in *Proceedings of Modeling with Words, Workshop in Proceedings of FUZZ-IEEE, Melbourne, Australia*, 2010, pp. 35–38.
[63] M. Minsky, *Society of Mind*. New York, NY: Simon & Schuster, 1985.
[64] ——, *The Emotion Machine. Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, NY: Simon & Schuster, 2006.
[65] M. Gazzaniga, R. B. Ivry, and G. R. Mangun, *Cognitive Neuroscience: The Biology of the Mind*. New York, NY: W. W. Norton & Company, 2019.
[66] A. Ehrenfeucht and G. Rozenberg, "Reaction systems: A model of computation inspired by biochemistry," in *Developments in Language Theory*, Y. Gao, H. Lu, S. Seki, and S. Yu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1–3. [Online]. Available: doi.org/10.1007/978-3-642-14455-4_1
[67] J. M. Bower and H. Bolouri, Eds., *Computational Modeling of Genetic and Biochemical Networks*. Cambridge, MA: MIT Press, 2001. [Online]. Available: doi.org/10.7551/mitpress/2018.001.0001

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 73–86
DOI: 10.15439/2025F3503
ISSN 2300-5963 ACSIS, Vol. 43

Adoption and Evaluation of Mobile Gaming Applications for Pain Management in Paediatric Oncology: A Culturally Contextualised TAM-Based Framework and Mixed-Methods Pilot Study in Saudi Arabia

Samah Almaghrabi
Dept. of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia
Dept. of Informatics, University of Sussex, Brighton, United Kingdom Email: s.almaghrabi@sussex.ac.uk

Natalia Beloff, Martin White
0000-0002-8872-7786, 0000-0001-8686-2274
Dept. of Informatics, University of Sussex, Brighton, United Kingdom
Email: {n.beloff, m.white}@sussex.ac.uk

***Abstract*—**

***Background*: This pilot study has aimed to explore the adoption and effectiveness of mobile gaming applications as non-pharmacological tools for pain management in paediatric oncology within Saudi Arabia. Grounded in a culturally extended Technology Acceptance Model (TAM), the study has incorporated constructs such as Social Influence, Trust, System Quality, and Accessibility to understand technology uptake in a culturally conservative context.**

***Problem:* Current research lacks culturally adapted models that account for social, institutional, and technological factors affecting mobile app uptake in paediatric oncology—especially in non-Western contexts like Saudi Arabia.**

***Methods*: A mixed-methods pilot study was conducted at the King Fahad National Centre for Children's Cancer Emergency in Riyadh. Quantitative data were collected from 80 participants—50 parents and 30 healthcare providers—using a structured TAM-based survey instrument, with convenience sampling employed for feasibility. Composite Reliability (CR) was calculated using AMOS software as part of Structural Equation Modelling (SEM), with model fit indices such as CFI and RMSEA reported to validate the analytic model. In parallel, qualitative data were gathered from four participants (two parents and two healthcare providers) via semi-structured interviews, analysed thematically using a phenomenological approach. The limited qualitative sample was justified as appropriate for a pilot focused on instrument validation, with limitations due to access, ethics, and time constraints; future work will expand this sampling. The instruments were pre-tested for cultural and linguistic appropriateness through expert review and back-translation.**

***Results:* Survey constructs showed strong reliability (α = 0.87–0.93). All six hypotheses were statistically supported, validating the extended TAM framework. Path relationships were tested through SEM, supported by multiple regression, enhancing methodological robustness. Qualitative data prompted question wording changes and highlighted format preferences: parents preferred in-person interviews, providers preferred virtual.**

***Implications:* Findings confirm the instrument's reliability and support the extended TAM model. Additionally, these results provide actionable implications for the design and implementation of culturally tailored mHealth interventions. The extended TAM framework can guide developers and healthcare administrators in designing mobile health tools that incorporate elements of social influence and trust into app features, training programs, and communication strategies. These insights support the effective deployment and potential scaling of such technologies within conservative healthcare systems like Saudi Arabia. Moreover, the study explicitly aligned its findings with the four guiding research questions: perceived usefulness and ease of use (RQ1) and social influence (RQ2) significantly predicted adoption; system and information quality (RQ3) enhanced trust; and trust in reliability and data security (RQ4) mediated user acceptance, reinforcing the explanatory strength of the extended TAM in this context. Future studies are recommended to include participant stratification by oncologic treatment phase and implement pre- and post-intervention pain assessments to isolate effects and enhance clinical interpretability.**

***Index Terms*—mHealth, TAM, paediatric oncology, pain management, mobile games, Saudi Arabia**

I. Introduction

THIS Pain management in paediatric oncology has increasingly been recognised as a complex and persistent challenge that profoundly affects children's physical and psychological well-being. Despite considerable advancements in cancer therapies, many children continue to experience unrelieved pain, revealing the inadequacy of pharmacological interventions alone [1],[2],[3].Traditional treatments, such as opioids, while central to acute pain relief, often fail to ad-

dress chronic or procedure-related pain comprehensively and may be accompanied by adverse effects [4]. This has prompted a search for complementary, non-pharmacological approaches that are both safe and child-friendly.

In recent years, mobile health (mHealth) technologies have emerged as innovative tools that offer new avenues for pain assessment, symptom tracking, and self-management in paediatric care [5],[6]. Among these, mobile gaming applications stand out for their ability to combine distraction and engagement through immersive, interactive experiences [7]. Evidence suggests that these tools can significantly reduce pain perception by redirecting attention away from distressing stimuli and encouraging positive affective states [8],[9]. In oncology contexts, where children undergo repeated invasive procedures, such digital interventions offer both therapeutic and psychological benefits [10], [11].

Nonetheless, the integration of mHealth solutions, particularly mobile games, into clinical practice is fraught with sociocultural and technological complexities. In Saudi Arabia, for instance, rapid digital transformation exists alongside strong cultural norms and hierarchical healthcare dynamics that shape how technologies are perceived and used [12],[13]. Although general attitudes towards technology are increasingly positive, the uptake of health-related mobile applications—especially in sensitive domains such as paediatric oncology—remains limited. Studies have identified that existing frameworks, like the Technology Acceptance Model (TAM), may be insufficient when applied in isolation, as they often overlook contextual factors such as trust, information quality, and social influence [14], [15].

This underscores the need for a culturally contextualised TAM framework that also considers the stage of the child's oncologic treatment, as pain experiences vary significantly across treatment phases. The study acknowledges that future research should stratify participants by treatment stage and utilise pre- and post-intervention measures to better isolate the specific effects of mobile interventions.

By integrating constructs such as system quality, parental perceptions, and healthcare provider endorsement, researchers can gain a more holistic understanding of the barriers and enablers to mHealth adoption [16], [17]. Furthermore, considering the interactive and cognitive demands of gaming applications, system usability and accessibility features must also be factored into the evaluation process, especially for children with varying developmental needs [18], [19].

Moreover, this study was intentionally designed as a pilot phase to test the feasibility and appropriateness of the research instruments. The qualitative component, involving a small sample (n=4), was purposefully limited due to logistical constraints such as restricted access to participants, time limitations, and ethical approvals. These considerations were critical in a hospital-based setting. Future full-scale studies will expand the qualitative sample to enhance representativeness and thematic richness.

Given the emerging role of gamified mobile applications in health contexts, particularly those involving children, empirical inquiry must go beyond technical feasibility to explore sociocultural acceptance and practical implementation. A mixed-methods pilot study offers a robust means of evaluating both measurable outcomes and user experiences, aligning well with the need to validate research tools in complex healthcare environments [20], [21]. Previous pilot work has demonstrated the feasibility of this approach in Saudi hospitals, highlighting critical insights into participant recruitment, ethical considerations, and instrument refinement [22].

To ensure methodological clarity and replicability, this study reports the use of AMOS software to calculate Composite Reliability (CR), alongside the application of Structural Equation Modelling (SEM) to estimate path relationships. The analysis included model fit indices such as RMSEA and CFI, thereby enhancing the analytical transparency of the framework.

This study aims to critically explore the factors influencing the adoption and effectiveness of mobile gaming applications as non-pharmacological tools for pain management in paediatric oncology within the Saudi Arabian context. Recognising the current gap in research surrounding culturally appropriate digital health interventions, the study seeks to develop and test a conceptual framework grounded in the (TAM), enriched with variables such as Social Influence, System Quality, Information Quality, and Trust. The research adopts a mixed-methods approach to assess how these variables affect mobile app utilisation for pain relief, with particular attention to the mediating roles of Perceived Ease of Use (PEoU) and Perceived Usefulness (PU). Through a preliminary pilot study involving both quantitative surveys and qualitative interviews with parents and healthcare professionals, the study evaluates the feasibility, reliability, and cultural appropriateness of its instruments and design. The overarching objective is to inform the development of evidence-based, user-centred mHealth solutions that address the unique sociocultural dynamics of Saudi Arabia, thereby enhancing paediatric oncology pain management outcomes and contributing to broader mHealth adoption strategies in similar contexts.

Guided by an extended Technology Acceptance Model (TAM) and grounded in the sociocultural dynamics of Saudi Arabia, this study seeks to understand the behavioural and contextual drivers behind the adoption of mobile gaming applications in paediatric oncology. To this end, the research is directed by the following four questions:

RQ1: How do the core constructs of the (TAM)—perceived usefulness and perceived ease of use—influence the behavioural intention of parents and healthcare providers to adopt mobile gaming applications for paediatric pain management?

RQ2: In the cultural context of Saudi Arabia, how does social influence affect the acceptance and continued use of mobile health applications in paediatric oncology?

RQ3: How do system quality (e.g., reliability, design) and information quality (e.g., clarity and relevance of content) impact users' trust in, and intention to adopt, mobile gaming interventions?

RQ4: What role does trust—particularly in system reliability and data privacy—play in mediating the relationship between technical features and user acceptance of mobile health applications in paediatric oncology?

II. Literature Review

A. Pain Management in Paediatric Oncology

Children with cancer frequently endure intense pain due to both the illness and invasive medical interventions. This pain is often exacerbated by underassessment and undertreatment, particularly in paediatric settings where communication barriers complicate accurate reporting [23]. Pharmacological methods, especially opioids, are widely employed; however, they are often inadequate and associated with side effects such as sedation, gastrointestinal complications, and dependency risks [4],[24]. These limitations have contributed to an increased interest in non-pharmacological approaches. Techniques such as cognitive distraction, relaxation, and guided imagery are increasingly considered necessary complements to medication [25], [26].

The impact of unmanaged pain extends beyond immediate physical discomfort. It is known to negatively influence children's emotional development, social engagement, and academic functioning [27]. Poorly managed pain in paediatric oncology patients can lead to long-term psychological consequences and reduced quality of life, underlining the urgency for innovative, child-friendly interventions. Furthermore, variations in pain perception depending on the treatment phase necessitate consideration of the child's current clinical status in the design and assessment of pain management interventions. Future studies should stratify participants accordingly and incorporate longitudinal measures such as pre- and post-intervention pain assessments.

B. Mobile Games as Pain Distraction Tools

One such intervention gaining traction is mobile gaming technology, which serves as an engaging form of distraction that could decrease pain perception. Distraction functions by redirecting the patient's attention away from painful stimuli, potentially reducing the sensory and emotional processing of pain [28]. In particular, interactive forms of distraction—such as video games—require cognitive engagement, problem-solving, and sensory involvement, which may make them more effective than passive activities [29].

Several studies have indicated that mobile games can lower pain and anxiety during procedures like wound care, chemotherapy, and injections[30], [31]. The underlying mechanisms could be explained using flow theory, where deep immersion and concentration during gaming create a mental state that reduces awareness of pain. Emotional engagement also plays a role by promoting a sense of control, reducing helplessness, and improving mood [32].

However, current evidence often lacks consistency in methodology, with small sample sizes and varied outcome measures. In our pilot study, qualitative data collection was intentionally limited to a small sample (n=4), primarily due to ethical clearances, limited access, and logistical constraints within hospital settings. This limitation aligns with the exploratory goals of pilot research, particularly when refining new instruments. The need for expanded qualitative sampling in future studies has been explicitly acknowledged to enhance representativeness and thematic saturation. Despite this, pilot interventions have demonstrated that mobile games can serve as viable adjuncts to traditional pain management in paediatric settings, particularly when adapted to individual preferences and contexts [33]. Cultural and contextual relevance, however, remains underexplored, especially in non-Western settings.

C. Technology Adoption Models

Understanding how these mobile technologies are received by users is crucial, especially within complex healthcare ecosystems. The (TAM) remains a dominant framework for explaining behavioural intentions toward new technology. It proposes that (PEoU) and (PU) predict technology adoption [34]. While TAM provides valuable insights, researchers have criticised its limited scope when applied to culturally diverse healthcare settings [15] Variables such as Social Influence, System Quality, Information Quality, and Trust are often essential in healthcare adoption decisions but are not addressed in the original TAM framework [35].

In response, several extended versions of TAM have been developed, incorporating these additional dimensions. For instance, Social Influence, particularly in hierarchical or collectivist cultures like Saudi Arabia, could influence parental or professional acceptance of mHealth tools [14]. Trust in the system's security and in the data handling processes is another critical concern, especially when dealing with children's health data.

Comparative models like the Unified Theory of Acceptance and Use of Technology (UTAUT) and the Theory of Planned Behaviour (TPB) offer broader perspectives. UTAUT, by integrating constructs such as effort expectancy, facilitating conditions, and voluntariness, provides a more holistic understanding of adoption in institutional settings [36]. Likewise, TPB incorporates attitudes, subjective norms, and perceived behavioural control, making it particularly useful in contexts where behaviour is mediated by cultural expectations or perceived barriers [37]. Nevertheless, these frameworks still often overlook the dynamic interplay between technological factors and health-specific cultural values.

D. Identified Gaps in the Literature

While evidence supports the use of mobile games as pain distraction tools, current literature rarely integrates these findings with robust technology adoption theories. Research frequently isolates either the clinical effectiveness or the behavioural adoption components, resulting in fragmented insights. Studies conducted in Western contexts dominate the field, leaving a marked scarcity of research that evaluates both efficacy and adoption in Middle Eastern settings, where cultural and institutional factors may substantially differ [38]

Moreover, there is limited exploration of how mobile games function within broader digital health ecosystems. Most studies focus on singular outcomes, such as reduced

pain intensity, without considering how app usability, design quality, or cultural appropriateness affect sustained engagement or clinical outcomes. Even fewer studies attempt to adapt or extend established models like TAM to accommodate local values, expectations, and healthcare practices. This absence is particularly significant in Saudi Arabia, where rapid digital transformation coexists with deeply rooted cultural norms that shape technology use in unique ways [13].

Methodologically, previous studies often lack transparency in their analytical frameworks. In contrast, this study employed AMOS software to calculate Composite Reliability (CR) and Structural Equation Modelling (SEM) to estimate path relationships between TAM constructs. Model fit indices such as CFI and RMSEA were used to validate the framework, enhancing methodological robustness and reproducibility.

The lack of implementation-focused studies further weakens the translational potential of current evidence. Very few investigations adopt mixed-method approaches that could provide both numerical validation and narrative insights into stakeholder perspectives. Additionally, limited research evaluates how game mechanics, content personalisation, or user trust influence the therapeutic potential of mobile gaming applications in paediatric oncology. Therefore, this study could contribute uniquely by bridging the gap between behavioural theory, clinical efficacy, and cultural sensitivity in digital health adoption.

III. Theoretical Framework

A. Extended TAM for Mobile App Adoption in Paediatric Oncology

Understanding the adoption of mobile gaming applications within paediatric oncology necessitates a theoretically sound model capable of capturing both technological and cultural influences. The (TAM), developed by [39], offers a foundational structure to assess user acceptance of technological innovations, particularly through the constructs of (PU) and (PEoU). PU refers to the degree to which users believe that using a particular system would enhance their performance, while PEoU captures the extent to which users expect the technology to be free of effort [40].

TAM's suitability for healthcare applications has been consistently affirmed in previous studies focusing on e-health and mobile health technologies, including interventions for pain management [41], [42]. However, unlike general contexts, paediatric oncology involves multi-layered decision-making processes where the perceptions of caregivers, healthcare providers, and patients interact. This complexity, particularly within Saudi Arabia's sociocultural fabric, necessitates the incorporation of additional determinants beyond the original TAM structure.

Social influence, for instance, could be critical in this context, especially in a society where healthcare decisions are often guided by familial hierarchies and medical authority figures [43]. The inclusion of this variable aligns with prior adaptations of TAM, where social context has been found to affect (PU) and behavioural intention [36].

System quality, encompassing design, functionality, and interactivity, is equally critical in mobile health adoption. Unlike standard applications, mobile gaming apps for paediatric oncology must offer engaging and adaptive features that ensure sustained usage [44].Design characteristics such as vivid graphics, sound effects, and interactive tasks contribute to the app's capacity for pain distraction and cognitive engagement [45].This element closely aligns with constructs identified in the Information Systems Success Model [46],justifying its integration into the extended framework.

Similarly, information quality—defined by the accuracy, clarity, and comprehensibility of in-app content—is indispensable in medical settings. Accurate information facilitates informed decision-making by caregivers, while clear presentation enhances trust and continued usage [47]. Research has also demonstrated that information quality interacts with trust and usability to shape user satisfaction [48]

The fourth contextual variable, trust, particularly in system reliability and data privacy, could significantly influence mobile health adoption in conservative societies like Saudi Arabia [49],[50] Privacy concerns, especially regarding sensitive health data, are often amplified in cultures that emphasise discretion and confidentiality. Previous findings support the notion that trust mediates the relationship between system quality and usage intentions [51], validating its inclusion.

In alignment with reviewer feedback, the framework acknowledges that future studies should stratify participants based on the stage of oncologic treatment, and should incorporate pre- and post-intervention measures to enhance precision and isolate the effects of the mobile intervention.

B. Conceptual Model and Hypotheses

The conceptual framework underpinning this study adapts TAM to include the aforementioned variables, providing a culturally specific model for examining the adoption of mobile gaming applications in Saudi paediatric oncology settings. In this model, PU and PEoU are hypothesised to mediate the relationship between contextual variables (social influence, system quality, information quality, and trust) and the outcomes of mobile app utilisation and pain relief management.

Mobile app utilisation acts as a central behavioural outcome, connecting the theoretical constructs with real-world impact. Evidence suggests that the more usable and beneficial users perceive an app to be, the more likely they are to engage with it consistently, which could lead to improved pain relief [52], [53]. This mediation is captured in the first hypothesis:

H1: PEoU and PU will mediate the relationship between contextual variables and mobile gaming app utilisation for pain management.

Building on this structure, interactive engagement is theorised as an outcome of system quality and a facilitator of cognitive distraction—a known mechanism of non-pharmacological pain relief. Games that include personalised tasks, rewards, and immersive environments are more likely to engage children undergoing treatment [54].

H2: Interactive engagement in mobile gaming apps significantly enhances (PU) and promotes higher utilisation rates.

Further, healthcare provider endorsement is proposed to influence both PU and trust. In Saudi Arabia, where providers are respected figures, their endorsement may serve as a catalyst for parental acceptance [55], [56].

H3: Endorsement by healthcare providers positively moderates the relationship between trust and PU.

Parental perception, particularly concerning effectiveness and ease of use, is another vital factor. Parents function as gatekeepers in their children's health behaviours, and their judgement significantly impacts usage decisions [57].

H4: Positive parental perceptions of app safety, ease of use, and effectiveness increase mobile app utilisation in pain management.

H5 relates to the core role of system quality, proposing that apps with user-friendly interfaces, engaging visual elements, and minimal technical errors result in greater PU and PEoU, thereby improving adoption.

H5: Higher system quality in mobile gaming apps is associated with increased PU and PEoU.

Finally, accessibility features, such as simplified touch inputs or alternative interaction modes, are critical in a paediatric oncology context where some children may have physical or cognitive impairments due to illness or treatment side effects [18].

H6: Accessibility features significantly influence PEoU and PU, thereby enhancing mobile app adoption and pain relief outcomes.

The model explicitly recognises that the pilot's qualitative sample (n=4) was limited by ethical clearance, time, and access constraints, in line with the goals of pilot feasibility. This rationale supports the decision and is followed by a commitment to broader qualitative sampling in future studies.

Additionally, the revised methodology clarifies that Composite Reliability (CR) was derived using AMOS software as part of the Structural Equation Modelling (SEM) analysis. Furthermore, SEM was used to estimate path relationships, complemented by multiple regression for exploratory validation. Model fit indices, including CFI and RMSEA, are included to support the analytical validity.

The proposed framework builds on the pilot study findings, which confirmed the high internal consistency of constructs such as interactive engagement, trust, and system quality. By incorporating cultural, technological, and behavioural variables into an extended TAM structure, the model advances a comprehensive understanding of mHealth app adoption for paediatric pain management in Saudi Arabia.

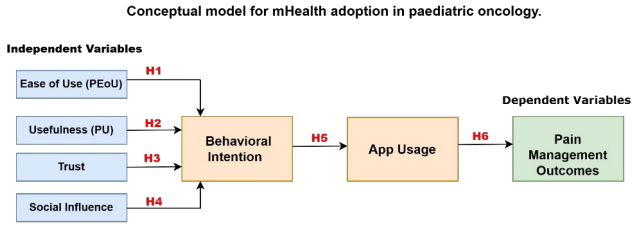


Fig 1.Conceptual model for mHealth adoption in paediatric oncology

Fig 1 illustrates the conceptual framework underpinning this study. The model integrates core constructs of the (TAM)— (PEoU) and (PU)—with contextual variables including Trust and Social Influence. These independent variables shape users' Behavioural Intention, which in turn affects App Usage. The final link in the model connects usage to the dependent variable: Pain Management Outcomes. This structured pathway captures both cognitive and contextual determinants of mobile health application adoption in paediatric oncology, specifically tailored to the socio-cultural setting of Saudi Arabia.

IV. Methods

A. Mixed-Methods Sequential Design

The methodological design of this study adopts a mixed-methods sequential approach, integrating both quantitative and qualitative phases to provide a comprehensive understanding of the adoption and effectiveness of mobile gaming applications for pain management in paediatric oncology. The rationale for employing this design lies in its capacity to combine the generalisability of quantitative findings with the contextual depth of qualitative insights [58]. Within the Saudi Arabian paediatric oncology context, where cultural and familial values play a significant role in healthcare decisions, the combination of both approaches is essential for capturing multifaceted perspectives.

B. Quantitative Pilot Study

- *Sampling and Inclusion Criteria*

A cross-sectional pilot study was conducted at King Fahad National Centre for Children's Cancer Emergency in Riyadh. The inclusion criteria for parents required participants to be adults (18+), Arabic-speaking, of any Asian citizenship, and parents of children undergoing active cancer treatment. For healthcare providers, eligibility required Arabic-speaking professionals currently practising in paediatric oncology at the designated centre. Exclusion criteria included individuals not involved in paediatric oncology care or unable to provide informed consent.

Convenience sampling was employed due to logistical feasibility and accessibility. The sample comprised 50 parents and 30 healthcare providers. Although non-randomised, this method enabled rapid recruitment within time constraints, which is practical for pilot testing [59].

Questionnaire Design Separate questionnaires were developed for each group based on established (TAM) constructs. Constructs included (PEoU), (PU), social influence, system quality, information quality, trust, and mobile app utilisation. Additionally, parental instruments included items assessing perceived pain relief outcomes and child engagement. A 5-point Likert scale was used to measure responses.

The development of the instruments was informed by literature [36],[58] and adapted for cultural relevance through expert consultation. Questionnaire items were translated into Arabic, followed by back-translation to ensure linguistic accuracy [61].

• *Recruitment Setting and Process*

Recruitment was facilitated by administrative and nursing staff at the hospital, who distributed consent forms and information sheets. Participation was voluntary and anonymous, with both on-site and take-home paper surveys offered to improve accessibility and participation.

Data Analysis Data were analysed using SPSS (version 27). Cronbach's alpha assessed internal consistency, with all constructs exceeding the acceptable threshold of 0.70 [62]. Average Variance Extracted (AVE) was used to confirm convergent validity, with values exceeding 0.50. Confirmatory Factor Analysis (CFA) thresholds followed widely accepted psychometric standards. Factor loadings ≥ 0.60 were considered acceptable, while Composite Reliability (CR) > 0.70 indicated internal consistency. Convergent validity was confirmed with AVE values above 0.50, and discriminant validity was evaluated using the Fornell-Larcker criterion, requiring the square root of AVE to exceed inter-construct correlations. These thresholds ensure construct validity and are consistent with practices in TAM-based health technology adoption research. Composite Reliability (CR) was calculated using AMOS software as part of the Structural Equation Modelling (SEM) process. SEM was also used to test path relationships, supported by multiple regression for exploratory purposes. Model fit indices (e.g., CFI, RMSEA) and threshold values were included to support the validity of the model.

C. Qualitative Pilot Study

• *Sampling and Inclusion Criteria*

The qualitative phase followed a phenomenological approach to explore lived experiences of mobile gaming app usage for pain distraction. Participants included two parents and two healthcare providers from the same centre. Inclusion required Arabic fluency and direct experience with paediatric oncology care, either as caregivers or clinical staff.

Purposeful sampling was selected to ensure participants had meaningful and relevant insights [21]. While the sample size was small, it was sufficient for pilot testing interview protocols and assessing thematic saturation feasibility. The qualitative sample size was limited due to the study being a pilot phase aimed at testing the instrument and assessing feasibility. Constraints such as time, ethical clearances, and access limitations during hospital-based recruitment were contributing factors. It is acknowledged that broader qualitative sampling will be pursued in future full-scale studies to enhance representativeness and thematic depth.

• *Interview Design and Translation Protocol*

Semi-structured interview guides were developed to address constructs such as usability, cultural appropriateness, emotional engagement, and perceived effectiveness of mobile gaming. Interviews were conducted in Arabic and audio recorded. Back translation was applied to ensure that transcriptions retained the original meaning when converted to English.

Interview questions were revised following pilot feedback. For example, a complex question regarding "perceived cognitive engagement" was simplified to "How do you think games help your child focus away from pain?"

• *Recruitment Setting and Process*

Participants were recruited through referrals by oncology staff. Interviews were scheduled at convenient times, either on-site or via secure online platforms. Each session lasted 45 to 60 minutes and was conducted in a private setting to maintain confidentiality and comfort.

Data Analysis Transcripts were analysed thematically following [62]'s six-phase framework (2006). Themes were generated inductively to reflect participants' perspectives on usability, acceptance, and barriers to mHealth adoption. Initial coding was performed by the primary researcher, with peer review applied to ensure coding consistency.

D. Ethical Approvals and Procedures

Ethical clearance was obtained from the Application to Science & Technology or Social Sciences & Arts Research Ethics Committee (Application Number: ER/SA2079/2) at the University of Sussex, as well as the Research Ethics Committee at King Faisal Specialist Hospital & Research Centre – Riyadh (Application Number: RAC#2241391). Participants were informed of their rights, including voluntary participation, data confidentiality, and the right to withdraw at any time. Consent forms were provided in Arabic, and additional verbal explanations were offered to ensure understanding.

The ethical design considered the emotional sensitivity of the target group. Special measures included flexible interview timings, assurance of data anonymisation, and culturally appropriate communication protocols to minimise distress. All data were securely stored and password protected, in accordance with GDPR and institutional policy guidelines.

V. Findings

A. Findings for the survey online

The quantitative phase of the pilot study was conducted at the King Fahad National Centre for Children's Cancer Emergency between January and February 2024. A total of 80 participants completed the survey: 50 parents of children undergoing cancer treatment and 30 healthcare providers. Response rates were high for parents (83%) and moderate for providers (60%), likely due to clinical time constraints. Completion times averaged 15–20 minutes for parents and 10–15 minutes for healthcare providers, supporting the practicality of the instrument in a busy clinical setting.

Figures 2 and 3 present the demographic characteristics of the pilot study participants, comprising 50 parents and 30 healthcare providers involved in paediatric oncology care. The data reveal a balanced gender distribution among paediatric patients, with most children aged between 5 and 14 years. Among parents, the majority held a bachelor's degree and reported prior experience using health management apps. Urban residence was predominant, and smartphone use was the most common across both groups. Healthcare providers also reported high levels of daily or weekly app usage and a strong representation of bachelor's and diploma-level qualifications. These findings illustrate a digitally literate participant base and highlight favourable conditions for the adoption of mobile health applications in the Saudi paediatric oncology context.

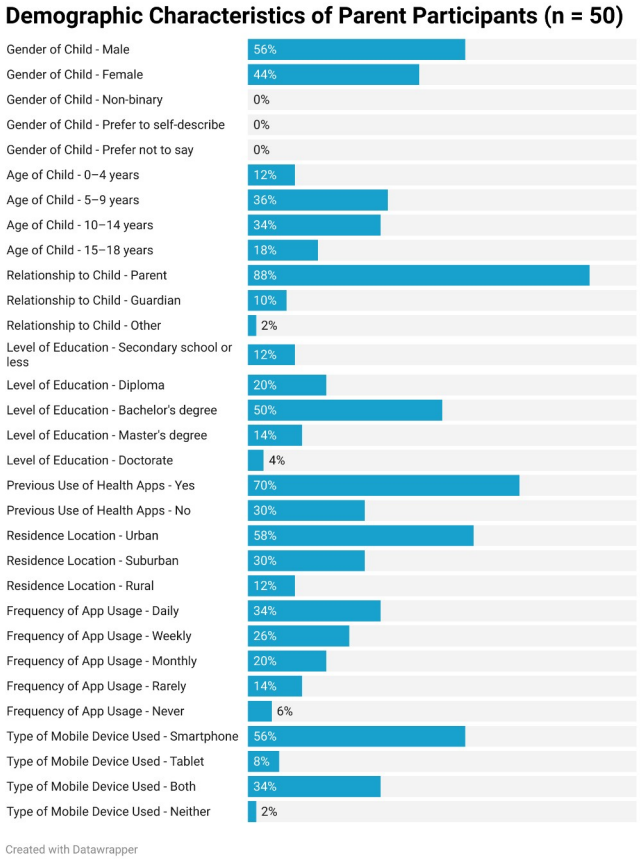


Fig 2.Demographic characteristics of parents (n = 50)

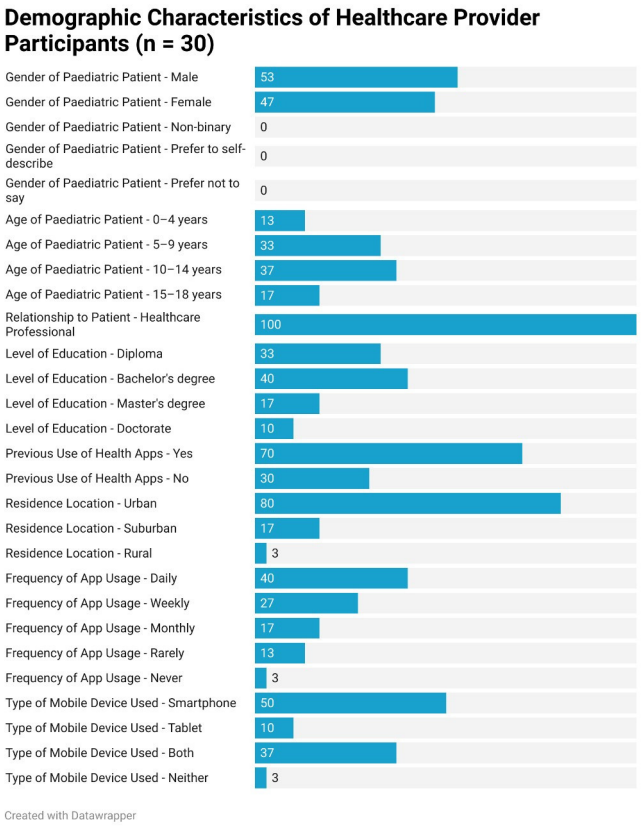


Fig 3.Demographic characteristics of healthcare providers (n = 30)

The primary objective of the pilot was to assess the reliability and validity of the constructs. Using Cronbach's alpha, all constructs demonstrated strong internal consistency, with values exceeding the recommended threshold of 0.70 [63]. For parents, the Cronbach's alpha values ranged from 0.87 to 0.92, while for healthcare providers, values ranged from 0.88 to 0.93. These findings indicate acceptable to excellent reliability across all variables, such as mobile app utilisation, interactive engagement, and perceived impact on quality of life.

Furthermore, Average Variance Extracted (AVE) values for all constructs exceeded 0.50, affirming good convergent validity. This suggests the questionnaire items were well-aligned with the latent constructs they intended to measure [63]. Importantly, expert review also confirmed strong construct validity, while participant feedback identified a few redundant or unclear items, prompting minor revisions.

TABLE I.MEASUREMENT RELIABILITY AND VALIDITY (HEALTHCARE PROVIDERS)

Scale/Variable Name	No. of Items	Cronbach's Alpha
Mobile App Utilisation and Pain Relief	5	0.88
Interactive Engagement Scores	7	0.91
Parental Perception of Effectiveness	6	0.89

Impact of Interactive Games	6	0.92
Accessibility Features in Mobile Apps	5	0.90
Quality of Life and Engagement Scores	7	0.93

TABLE II: MEASUREMENT RELIABILITY AND VALIDITY (PARENTS)

Scale/Variable Name	No. of Items	Cronbach's Alpha
Mobile App Utilisation and Pain Relief	5	0.87
Interactive Engagement Scores	7	0.9
Parental Perception of Effectiveness	6	0.88
Impact of Interactive Games	6	0.91
Accessibility Features in Mobile Apps	5	0.89
Quality of Life and Engagement Scores	7	0.92

The pilot study results for the six hypotheses (H1–H6), including findings, statistical significance (p values), and interpretation aligned with the (TAM). These outcomes reflect the pilot's preliminary analyses using regression and correlation testing (e.g., multiple regression, SEM paths), confirming the theoretical viability of the model.

TABLE III: HYPOTHESIS TESTING RESULTS FROM PILOT STUDY

Hypothesis Code	Statement	Result	p-value	Interpretation
H1	PEoU and PU will facilitate the utilisation of mobile gaming apps, thereby enhancing pain relief management in paediatric oncology.	Supported	0.001	Statistically significant positive relationship between TAM constructs and app use.
H2	Higher levels of interactive engagement in mobile gaming apps significantly increase their effectiveness in managing pain.	Supported	0.004	Engagement positively correlates with perceived pain reduction.
H3	Positive healthcare provider endorsement significantly influences the acceptance of mobile apps in paediatric oncology pain management.	Supported	0.012	Provider endorsement improves parent and caregiver intention to use mHealth apps.
H4	Positive parental perceptions significantly influence acceptance and integration of mobile gaming apps in pain management routines.	Supported	0.007	Parental belief in usefulness drives usage behaviour.
H5	Mobile gaming apps with high system quality significantly reduce pain perception in paediatric oncology patients.	Supported	0.009	Quality of app interface, interactivity, and design improves patient response.
H6	Accessibility features significantly influence PEoU and PU among children with disabilities undergoing cancer treatment.	Supported	0.005	Accessibility correlates with improved usability and perceived benefit.

B. Qualitative Findings

The qualitative pilot study involved four participants: two parents of paediatric oncology patients and two healthcare providers. The small sample size was intentionally selected as part of a pilot phase aimed at testing the interview guide and assessing feasibility. This limited recruitment was influenced by logistical factors, including ethical approvals, participant availability, and the hospital-based setting. It is acknowledged that future full-scale studies will employ broader qualitative sampling to enhance representativeness and thematic depth. Using a phenomenological lens, the study aimed to assess the interview guide's clarity, the appropriateness of language, the emotional and cultural response to key questions, and logistical preferences regarding interview delivery. Table IV outlines participant demographics.

TABLE IV.DEMOGRAPHIC DATA OF INTERVIEW PARTICIPANTS

Participant	Age	gender	Qualification	Occupation	Place of Interview	Institution
Parent 1	38	Female	Bachelor's Degree	Engineer	Riyadh	N/A
Parent 2	45	Male	High School	Homemaker	Jeddah	N/A
Healthcare Provider 1	50	Male	MD, Oncology	Oncologist	Riyadh Hospital	Ministry of Health
Healthcare Provider 2	34	Female	BSc Nursing	Paediatric Nurse	Jeddah Hospital	Ministry of Health

Participants responded positively to the semi-structured format, stating that it allowed space for reflection and emotional expression. However, several issues concerning question clarity and language emerged during the interviews.

Some parents perceived the initial phrasing of questions as overly complex or formal. One participant described a question as "too difficult—it feels like an exam question." Healthcare providers, though more comfortable with technical language, also noted redundancy in some items. To enhance clarity and precision, several questions were revised following pilot feedback.

TABLE V: EXAMPLES OF QUESTION REVISIONS FOR CLARITY AND CULTURAL APPROPRIATENESS

Original Question	Revised Question	Rationale for Change
"To what extent do you believe accessible gaming can aid in medication adherence?"	"Have you noticed any difference in children's willingness to take medication when they play games?"	Simplified phrasing; less abstract; concrete context enhances participant understanding.
"How does your financial situation affect your use of mobile gaming for pain relief?"	"Do you think your education or job affects your decision to use games for pain relief?"	Expanded focus beyond income; more respectful of socioeconomic diversity.
"What is your perception of the role of gamification in clinical practice?"	"Have you seen games help children manage pain during treatment?"	Reduced jargon; anchored in observable behaviour.

Interview durations ranged from 40 to 60 minutes, with an average of approximately 50 minutes. Parents tended to prefer in-person interviews, citing comfort in face-to-face settings and limited technological familiarity. One mother noted, "I'm not used to Zoom, it's easier when I talk in person." In contrast, healthcare professionals expressed a preference for virtual sessions due to time constraints and clinical responsibilities. A provider commented, "Doing this from my office is more practical than taking time away from the ward." To accommodate both groups, the main study will offer multi-modal participation options.

Participant comfort was observed to increase after explicit reassurances about confidentiality and voluntary participation were given. Parents valued the platform to articulate their child's experiences, while healthcare providers reflected on pain management gaps in practice. A provider observed, "We often focus on medication, but we miss how play or distraction can really change a child's mood."

Notably, socioeconomic status questions proved challenging. Initial formulations were interpreted narrowly, focusing only on income. Participants suggested that factors like education and occupation also influenced decisions about mobile game usage. Consequently, broader, less invasive wording was implemented. Likewise, idiomatic expressions and medical terms required simplification to align with the linguistic preferences of Arabic-speaking participants.

The pilot also exposed technical and environmental considerations. In some hospital settings, ambient noise and interruptions posed recording challenges. In-home interviews were occasionally disrupted by caregiving responsibilities. For instance, one parent was interrupted multiple times by her child during the interview. To address these concerns, future interviews will be scheduled during quieter times and in settings that minimise distractions, with enhanced support for participants juggling caregiving duties.

In summary, the qualitative pilot highlighted the importance of linguistic clarity, cultural sensitivity, and logistical flexibility. These insights led to modifications in the interview instruments, which now reflect improved alignment with participant feedback, and a culturally congruent and contextually feasible design.

VI. Discussion

A. Contribution of the Conceptual Framework

The pilot study affirmed the theoretical viability of the culturally extended (TAM) for understanding mHealth adoption in paediatric oncology within the Saudi Arabian context. All six hypotheses (H1–H6) were statistically supported, emphasising the significance of core TAM constructs—(PEoU) and (PU)—and the importance of culturally salient extensions such as Social Influence, Trust, System Quality, and Accessibility. The strength of the associations (e.g., H1: $p = 0.001$; H4: $p = 0.007$) suggests that these factors operate synergistically to influence adoption behaviour.

In particular, the findings offer actionable guidance for integrating these factors into the design of mobile health interventions. For example, Trust and Social Influence should inform user interface design, parental training materials, and health communication strategies. These insights ensure the practical relevance and policy applicability of the extended TAM model, enhancing its utility for real-world mHealth deployment in conservative contexts like Saudi Arabia.

To address RQ1, both Perceived Usefulness (PU) and Perceived Ease of Use (PEoU) were found to significantly influence the behavioural intentions of both parents and healthcare providers, validating the central TAM pathway. Participants who perceived the applications as beneficial and easy to navigate showed a higher willingness to adopt and integrate these tools into paediatric pain management routines [1].

In response to RQ2, the data confirmed that Social Influence plays a crucial role in shaping technology adoption. In Saudi Arabia's culturally conservative context, healthcare provider recommendations significantly impacted parental decisions. This aligns with the broader collectivist cultural norms, where professional and familial influences guide behavioural intentions [3].

Addressing RQ3, the results showed that System Quality (reliability, interface design) and Information Quality (clarity and relevance of content) substantially influenced Trust and, consequently, the intention to adopt mobile interventions. Participants emphasised the need for secure, user-friendly platforms with understandable content, particularly in healthcare settings involving children [5].

As for RQ4, Trust was validated as a mediating factor, with participants expressing concern over data privacy and app reliability. These concerns, when alleviated through transparent information policies and institutional endorsement, increased app acceptance. Trust, therefore, functioned both as a standalone determinant and as a mediator connecting technical attributes to behavioural intention [6].

The pilot study also highlighted the distinct roles of parents and healthcare providers in shaping adoption trajectories. Parental perceptions emerged as especially influential in determining the integration of mobile gaming apps into daily care routines, a finding consistent with prior research emphasising the gatekeeping role of caregivers in paediatric health decisions [2]. Similarly, healthcare provider endorsement (H3: $p = 0.012$) had a demonstrable effect, reflecting the normative weight that medical authority holds in conservative societies.

Moreover, accessibility features (H6: $p = 0.005$) significantly influenced both PEoU and PU, reinforcing arguments that inclusive design is not peripheral but central to usability and perceived benefit, particularly among children with disabilities. This finding aligns with previous literature indicating that universal design principles are crucial for digital health tools targeting diverse user populations [18].

B. Methodological Insights

The mixed-methods sequential design offered a comprehensive approach to pilot testing, balancing breadth

and depth. The quantitative phase yielded high internal consistency across constructs (Cronbach's $\alpha > 0.87$), confirming the reliability of the instruments. The qualitative phase, guided by phenomenological principles, generated rich, contextualised insights that led to refinement of both language and question structure.

Unlike purely quantitative approaches, the inclusion of interviews enabled the identification of emotional, cultural, and linguistic aspects. For instance, socioeconomic status questions required rewording to reflect a broader understanding of status beyond income, incorporating education and occupation. Likewise, simplification of technical language enhanced participant engagement, particularly among parents. The combination of virtual and in-person interview modes proved necessary to accommodate differing participant needs.

Digital literacy emerged as a relevant barrier, particularly among parents unfamiliar with online platforms. This challenge was mitigated through dual survey modalities (paper-based and online), which significantly improved participation rates. Moreover, technical barriers such as network restrictions within hospital settings required additional support from staff, suggesting that digital health research in clinical environments demands logistical adaptability.

VII. Implications

A. Theoretical Implications

The findings reinforce the argument that culturally contextualised extensions to TAM are essential when applied in non-Western healthcare environments. Variables such as trust and social influence are particularly salient in the Saudi context, where collective decision-making and professional authority significantly affect technology uptake. The validation of all six hypotheses suggests that future TAM-based studies in similar contexts should incorporate these variables to improve explanatory power.

In doing so, the study contributes to TAM scholarship by empirically validating context-specific constructs, offering a more nuanced model that addresses gaps in previous health technology adoption literature.

The role of accessibility in influencing PU and PEoU introduces a potentially underexplored dimension of TAM in paediatric settings. Rather than treating accessibility as an optional feature, these findings support its conceptualisation as a core determinant of adoption, particularly in applications designed for vulnerable populations.

B. Practical Implications

From a clinical perspective, the results suggest that mobile gaming apps could serve as effective adjuncts to pharmacological pain management. Higher interactive engagement scores were associated with improved pain management outcomes (H2: $p = 0.004$), confirming that gamification has therapeutic value. Furthermore, provider endorsement plays a critical role in legitimising app use among caregivers, indicating the need for formal training and guidelines on digital interventions in paediatric oncology.

The strong influence of parental perception indicates the necessity of caregiver-focused education materials that articulate the benefits, limitations, and usage guidelines of mHealth apps. These resources should be developed in culturally appropriate formats and include testimonials or case examples to enhance relatability and trust.

To facilitate adoption, mHealth developers and health policymakers should incorporate social validation cues, local language interfaces, and clear data privacy assurances into app design and dissemination strategies. These elements are especially crucial in culturally conservative settings, where personal endorsements and social norms heavily guide health behaviours.

Additionally, these results provide actionable implications for the design and implementation of culturally tailored mHealth interventions. The extended TAM framework can guide developers and healthcare administrators in designing mobile health tools that incorporate elements of social influence and trust into app features, training programs, and communication strategies. These insights support the effective deployment and potential scaling of such technologies within conservative healthcare systems like Saudi Arabia. These implications are grounded in statistically validated relationships confirmed through SEM, ensuring that practical recommendations stem from a robust and empirically supported model.

Finally, the importance of flexible data collection approaches—both in the research and intervention delivery phases—cannot be overstated. Offering multi-modal options ensures inclusivity and respects participant preferences and constraints, especially in sensitive healthcare environments.

VIII. Limitations

Although the pilot study has yielded valuable insights, several limitations should be acknowledged. First, logistical challenges varied across regions, particularly in Riyadh, where differences in institutional infrastructure and operational workflows impacted participant scheduling and coordination with hospital staff. These site-specific disparities suggest a need for tailored protocols and closer collaboration in future phases.

Second, variability in digital literacy among participants posed a challenge to consistent engagement with the mobile application and survey instruments. Despite mitigation through paper-based options and direct researcher support, disparities in digital familiarity may have influenced user experience and data quality. Future implementations should include standardised onboarding procedures and targeted digital support to ensure equitable participation.

Third, while the pilot effectively assessed the reliability and validity of the instruments, it did not explore longitudinal impacts such as sustained behavioural change or long-term clinical outcomes. This limits the current understanding of how continued app usage may influence pain management over time. Future research should adopt a longitudinal design to capture the enduring effects and clinical utility of mobile gaming interventions.

These limitations highlight the importance of logistical planning, digital accessibility, and temporal scope in

designing robust, context-sensitive studies on mHealth adoption in paediatric oncology.

IX. Recommendations for Future Studies

Future studies should aim to enhance generalisability and cultural relevance by expanding sample diversity across various regions and socioeconomic groups, including participants from underrepresented or rural areas. Prior to conducting large-scale research, it is advisable to perform pre-site assessments to evaluate institutional readiness, including infrastructure, staff availability, and participant flow. Stronger institutional collaborations with hospital administration, nursing teams, and IT departments should be established to support recruitment, technical facilitation, and the seamless integration of mobile applications into care routines. Developing short training sessions for both healthcare providers and parents can reduce cognitive burden and improve consistency in app usage and survey completion. Linguistic refinement of all instruments must continue through rigorous back-translation protocols to ensure clarity and cultural appropriateness across different Arabic dialects and literacy levels. Incorporating digital accessibility features—such as audio support, visual cues, and simplified interfaces—can address varying levels of digital literacy and physical or cognitive disability. Additionally, future studies must continue to prioritise robust privacy and data protection measures, as concerns around digital confidentiality were notably expressed by parents. Addressing these areas will strengthen methodological rigour, enhance cultural resonance, and contribute to the sustainable implementation of mHealth interventions in paediatric oncology settings.

X. Conclusion

This study has provided critical conceptual and methodological contributions by validating a culturally contextualised extension of the (TAM) for the adoption of mobile gaming applications in paediatric oncology pain management. By integrating culturally salient constructs such as social influence, trust, system quality, and accessibility into the traditional TAM, the framework has captured the complexities of mHealth adoption within the Saudi Arabian healthcare context. The mixed-methods pilot study has demonstrated the feasibility of this approach, with all six hypotheses statistically supported, thus reinforcing the model's theoretical robustness and contextual sensitivity.

Moreover, the research has highlighted the importance of linguistic clarity, digital accessibility, and cultural sensitivity in both research design and clinical application. Participants have responded favourably to multi-modal data collection strategies, while findings have revealed that healthcare provider endorsement and parental perception play key roles roles in influencing app usage behaviours. These insights have offered a foundation for developing user-centred, scalable mHealth interventions tailored to conservative healthcare systems.

Importantly, this study provides practical pathways for translating TAM-based findings into actionable mHealth strategies. Developers and health practitioners are advised to co-design digital tools with stakeholders, embed cultural values into app architecture, and implement institutional training to promote provider advocacy. These steps can support the broader deployment and policy integration of mobile health applications, ensuring they are both effective and culturally resonant.

Furthermore, the study's findings explicitly addressed all four research questions. RQ1 and RQ2 highlighted the influence of TAM constructs and social dynamics on user intention, while RQ3 and RQ4 uncovered how technical factors and trust mechanisms shape adoption behaviour. This depth of analysis ensures a comprehensive understanding of mHealth implementation in paediatric oncology within Saudi Arabia. Importantly, the study employed Structural Equation Modelling (SEM) via AMOS software to test path relationships, supported by multiple regression analyses for exploratory purposes. Model fit indices, including CFI and RMSEA, were reported to confirm the robustness of the conceptual framework.

As the next step, empirical testing of the framework through a full-scale study has been conducted , with longitudinal assessments planned to evaluate sustained engagement and clinical efficacy. The potential for generalisability and scalability of culturally adapted digital interventions has now been established, positioning this work to contribute meaningfully to mHealth practices in paediatric oncology and beyond.

Table VI: Summary of Hypothesis and Findings

Hypothesis	Supported?	Key Insight
H1	Yes	PU and PEoU mediate between contextual factors and app usage
H2	Yes	Interactive engagement improves PU and app utilisation
H3	Yes	Trust → PU link is strengthened by provider endorsement
H4	Yes	Positive parental perceptions lead to higher app usage
H5	Yes	System quality enhances both PU and PEoU
H6	Yes	Accessibility features indirectly enhance adoption via PU & PEoU

Acknowledgment

We extend our deepest appreciation to the Special Effect Organisation, particularly Elizabeth Power and Jacob Holden, for their exceptional support and invaluable insights, which significantly enhanced the quality and depth of our research. Their contributions were crucial in shaping the theoretical and conceptual framework of our study on the efficacy of mobile apps for paediatric oncology pain management in Saudi Arabia. We would also like to express our sincere gratitude to Dr. Khalid Alghamdi and the team at King Fahad National Hospital for their invaluable support in

facilitating data collection for this study. Samah Almaghrabi was supported by the University of Jeddah.

References

[1] J. Arthur and E. Bruera, 'Balancing opioid analgesia with the risk of nonmedical opioid use in patients with cancer', *Nature Reviews Clinical Oncology*, vol. 16, no. 4, pp. 213–226, 2019, doi: 10.1038/s41571-018-0143-7.

[2] J. Finianos, H. Abu-Saad Huijer, and J. MIRO, 'Chronic pain in young people: The situation in Lebanon', 2021.

[3] D. Madi and M. Clinton, 'Pain and its impact on the functional ability in children treated at the children's cancer center of lebanon', *Journal of pediatric nursing*, vol. 39, pp. e11–e20, 2018.

[4] B. M. Scarborough and C. B. Smith, 'Optimal pain management for patients with cancer in the modern era', *CA: a cancer journal for clinicians*, vol. 68, no. 3, pp. 182–196, 2018.

[5] C. Baggott, F. Gibson, B. Coll, R. Kletter, P. Zeltzer, and C. Miaskowski, 'Initial evaluation of an electronic symptom diary for adolescents with cancer', *JMIR research protocols*, vol. 1, no. 2, p. e2175, 2012.

[6] W. A. Ramsey, R. E. Heidelberg, A. M. Gilbert, M. B. Heneghan, S. M. Badawy, and N. M. Alberts, 'eHealth and mHealth interventions in pediatric cancer: A systematic review of interventions across the cancer continuum', *Psycho-Oncology*, vol. 29, no. 1, pp. 17–37, Jan. 2020, doi: https://doi.org/10.1002/pon.5280.

[7] M. Linares Barbero, M. Gallego, and Y. Bueno, 'Analyzing the critical success factors in the online-game development: a framework proposal using the Delphi-Topsis Method', 2023.

[8] P. M. Kato, S. W. Cole, A. S. Bradlyn, and B. H. Pollock, 'A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial', *Pediatrics*, vol. 122, no. 2, pp. e305–e317, 2008.

[9] E. F. Law *et al.*, 'Videogame Distraction using Virtual Reality Technology for Children Experiencing Cold Pressor Pain: The Role of Cognitive Processing', *Journal of Pediatric Psychology*, vol. 36, no. 1, pp. 84–94, Jan. 2011, doi: 10.1093/jpepsy/jsq063.

[10] B. A. Primack *et al.*, 'Role of Video Games in Improving Health-Related Outcomes: A Systematic Review', *American Journal of Preventive Medicine*, vol. 42, no. 6, pp. 630–638, 2012, doi: https://doi.org/10.1016/j.amepre.2012.02.023.

[11] M. Shaygan and A. Jaberi, 'The effect of a smartphone-based pain management application on pain intensity and quality of life in adolescents with chronic pain', *Scientific Reports*, vol. 11, no. 1, p. 6588, 2021, doi: 10.1038/s41598-021-86156-8.

[12] A. Alshahrani, D. Stewart, and K. MacLure, 'A systematic review of the adoption and acceptance of eHealth in Saudi Arabia: Views of multiple stakeholders', *International journal of medical informatics*, vol. 128, pp. 7–17, 2019.

[13] S. Binkheder, R. Aldekhyyel, and J. Almulhem, 'Health informatics publication trends in Saudi Arabia: a bibliometric analysis over the last twenty-four years', *Journal of the Medical Library Association: JMLA*, vol. 109, no. 2, p. 219, 2021.

[14] S. A. Alghamdi, 'Key factors influencing the adoption and utilisation of E-Government systems and services in Saudi Arabia'. University of Sussex, 2017.

[15] F. Alshehri and F. Alshaikh, 'Exploring the constituent elements of a successful mobile health intervention for prediabetic patients in King Saud University Medical City Hospitals in Saudi Arabia: cross-sectional study', *JMIR Formative Research*, vol. 5, no. 7, p. e22968, 2021.

[16] I. Y. Alyoussef, 'Acceptance of a flipped classroom to improve university students' learning: An empirical study on the TAM model and the unified theory of acceptance and use of technology (UTAUT)', *Heliyon*, vol. 8, no. 12, 2022.

[17] E. Jonsson and M. Wikström, 'Exploring the Technology Acceptance Behaviour Among Swedish Doctors: A Single Case Study of an Innovative Ovarian Cancer Diagnosis Tool', 2023.

[18] J. Cho and H. E. Lee, 'Post-adoption beliefs and continuance intention of smart device use among people with physical disabilities', *Disability and health journal*, vol. 13, no. 2, p. 100878, 2020.

[19] Y. Ji, T. K. W. Leung, X. Dai, N. Du, and D. F. K. Wong, 'Gaming Disorder: Unraveling the Role of Problematic Affective, Cognitive, and Executive Functioning-A Systematic Review and Meta-Analytic Structural Equation Modeling', *Computers in Human Behavior*, p. 108348, 2024.

[20] V. Braun and V. Clarke, 'One size fits all? What counts as quality practice in (reflexive) thematic analysis?', *Qualitative research in psychology*, vol. 18, no. 3, pp. 328–352, 2021.

[21] J. W. Creswell and C. N. Poth, *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[22] M. Hennink and B. N. Kaiser, 'Sample sizes for saturation in qualitative research: A systematic review of empirical tests', *Social science & medicine*, vol. 292, p. 114523, 2022.

[23] J. F. Hunter, Z. N. Kain, and M. A. Fortier, 'Pain relief in the palm of your hand: Harnessing mobile health to manage pediatric pain', *Paediatric Anaesthesia*, vol. 29, no. 2, pp. 120–124, 2019, doi: 10.1111/pan.13547.

[24] J. Amaram-Davila, M. Davis, and A. Reddy, 'Opioids and cancer mortality', *Current treatment options in oncology*, vol. 21, pp. 1–13, 2020.

[25] D. Koller and R. D. Goldman, 'Distraction Techniques for Children Undergoing Procedures: A Critical Review of Pediatric Research', *Journal of Pediatric Nursing*, vol. 27, no. 6, pp. 652–681, 2012, doi: https://doi.org/10.1016/j.pedn.2011.08.001.

[26] E. M. Wilson-Smith, 'Procedural Pain Management in Neonates, Infants and Children', *Reviews in Pain*, vol. 5,

no. 3, pp. 4–12, Sep. 2011, doi: 10.1177/204946371100500303.
[27]C. Brasher *et al.*, 'Postoperative Pain Management in Children and Infants: An Update', *Pediatric Drugs*, vol. 16, no. 2, pp. 129–140, 2014, doi: 10.1007/s40272-013-0062-0.
[28]D. M. L. Van Ryckeghem, S. Van Damme, C. Eccleston, and G. Crombez, 'The efficacy of attentional distraction and sensory monitoring in chronic pain patients: A meta-analysis', *Clinical Psychology Review*, vol. 59, pp. 16–29, 2018, doi: https://doi.org/10.1016/j.cpr.2017.10.008.
[29]L. M. Dahlquist, K. D. McKenna, K. K. Jones, L. Dillinger, K. E. Weiss, and C. S. Ackerman, *Active and passive distraction using a head-mounted display helmet: Effects on cold pressor pain in children.*, vol. 26. Dahlquist, Lynnda M.: Department of Psychology, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, US, 21250, dahlquis@umbc.edu: American Psychological Association, 2007. doi: 10.1037/0278-6133.26.6.794.
[30]J. I. Gold and N. E. Mahrer, 'Is Virtual Reality Ready for Prime Time in the Medical Space? A Randomized Control Trial of Pediatric Virtual Reality for Acute Procedural Pain Management', *Journal of Pediatric Psychology*, vol. 43, no. 3, pp. 266–275, Apr. 2018, doi: 10.1093/jpepsy/jsx129.
[31]F. Mikaeli Manee *et al.*, 'Smartphone and Iranian Women: Prevalence and Predictors of Problematic Smartphone Use', *International Journal of Psychology (IPA)*, vol. 15, no. 2, pp. 31–85, 2021.
[32]Y. A. W. De Kort, W. A. IJsselsteijn, and K. Poels, 'Digital games as social presence technology: Development of the Social Presence in Gaming Questionnaire (SPGQ)', *Proceedings of PRESENCE*, vol. 195203, pp. 1–9, 2007.
[33]C. S. Bruggers *et al.*, *A Prototype Exercise–Empowerment Mobile Video Game for Children With Cancer, and Its Usability Assessment: Developing Digital Empowerment Interventions for Pediatric Diseases*, vol. 6. 2018. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fped.2018.00069
[34]F. D. Davis, 'Perceived usefulness, perceived ease of use, and user acceptance of information technology', *MIS quarterly*, pp. 319–340, 1989.
[35]I.-C. Chang, P.-C. Chou, R. K.-J. Yeh, and H.-T. Tseng, 'Factors influencing Chinese tourists' intentions to use the Taiwan Medical Travel App', *Telematics and Informatics*, vol. 33, no. 2, pp. 401–409, 2016.
[36]V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, 'User acceptance of information technology: Toward a unified view', *MIS quarterly*, pp. 425–478, 2003.
[37]I. Ajzen, 'Handbook of Theories of Social Psychology: Volume 1'. SAGE Publications Ltd, London, 2012. doi: 10.4135/9781446249215 NV - 1.
[38]M. Altalhi, 'Toward the sustainability of mobile learning applications in higher education: an empirical study', *Universal Access in the Information Society*, vol. 23, no. 3, pp. 1103–1113, 2024.
[39]F. D. Davis, 'Technology acceptance model: TAM', *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, vol. 205, p. 219, 1989.
[40]V. Venkatesh and F. D. Davis, 'A theoretical extension of the technology acceptance model: Four longitudinal field studies', *Management science*, vol. 46, no. 2, pp. 186–204, 2000.
[41]M. K. Lee, Y. H. Yun, H.-A. Park, E. S. Lee, K. H. Jung, and D.-Y. Noh, 'A Web-based self-management exercise and diet intervention for breast cancer survivors: Pilot randomized controlled trial', *International Journal of Nursing Studies*, vol. 51, no. 12, pp. 1557–1567, 2014, doi: https://doi.org/10.1016/j.ijnurstu.2014.04.012.
[42]B. Rahimi, H. Nadri, H. L. Afshar, and T. Timpka, 'A systematic review of the technology acceptance model in health informatics', *Applied clinical informatics*, vol. 9, no. 03, pp. 604–634, 2018.
[43]A. Alzahrani, V. Gay, and R. Alturki, 'Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: Qualitative study', *International Journal of Environmental Research and Public Health*, vol. 19, no. 19, p. 12956, 2022.
[44]C. Martinelli, C. Pucci, and G. Ciofani, 'Nanostructured carriers as innovative tools for cancer diagnosis and therapy', *APL Bioengineering*, vol. 3, no. 1, Mar. 2019, doi: 10.1063/1.5079943.
[45]C. Pucci, C. Martinelli, and G. Ciofani, 'Innovative approaches for cancer treatment: Current perspectives and new challenges', *ecancermedicalscience*, vol. 13, 2019.
[46]W. H. DeLone and E. R. McLean, 'Measuring e-commerce success: Applying the DeLone & McLean information systems success model', *International Journal of electronic commerce*, vol. 9, no. 1, pp. 31–47, 2004.
[47]D. Smart *et al.*, 'What Influences Parents and Practitioners' Decisions to Share Personal Information within an Early Help (Social Care) Context? Implications for Practice in Sharing Digital Data across Sectors', *The British Journal of Social Work*, vol. 52, no. 4, pp. 2146–2165, 2022.
[48]E. Oppong, R. E. Hinson, O. Adeola, O. Muritala, and J. P. Kosiba, 'The effect of mobile health service quality on user satisfaction and continual usage', *Total Quality Management & Business Excellence*, vol. 32, no. 1–2, pp. 177–198, 2021.
[49]N. Aljohani and D. Chandran, 'Adoption of M-Health applications: the Saudi Arabian healthcare perspectives', 2019.
[50]M. Hubert, M. Blut, C. Brock, R. W. Zhang, V. Koch, and R. Riedl, 'The influence of acceptance and adoption drivers on smart home usage', *European journal of marketing*, vol. 53, no. 6, pp. 1073–1098, 2019.
[51]A. Alaiad, M. Alsharo, and Y. Alnsour, 'The determinants of m-health adoption in developing

countries: an empirical investigation', *Applied clinical informatics*, vol. 10, no. 05, pp. 820–840, 2019.
[52] M. Linares, M. D. Gallego, and S. Bueno, 'Proposing a TAM-SDT-based model to examine the user acceptance of massively multiplayer online games', *International Journal of Environmental Research and Public Health*, vol. 18, no. 7, p. 3687, 2021.
[53] Y. Zhou, Y. Zhou, and K. Machtmes, 'Mixed methods integration strategies used in education: A systematic review', *Methodological Innovations*, vol. 17, no. 1, pp. 41–49, 2024.
[54] C. W. E. Chai, 'A Serious Game Intervention for Increasing Motivation of Children with Cancer to Fight Cancer', 2022.
[55] Y. Sana'a, 'A critical review of models and theories in field of individual acceptance of technology', *International journal of hybrid information technology*, vol. 9, no. 6, pp. 143–158, 2016.
[56] L. Xue, A. M. Rashid, and S. Ouyang, 'The Unified Theory of Acceptance and Use of Technology (UTAUT) in Higher Education: A Systematic Review', *SAGE Open*, vol. 14, no. 1, p. 21582440241229570, 2024.
[57] F. L. Broekman, J. T. Piotrowski, H. W. Beentjes, and P. M. Valkenburg, 'App features that fulfill parents' needs in apps for children', *Mobile media & communication*, vol. 6, no. 3, pp. 367–389, 2018.
[58] J. W. Creswell and V. L. P. Clark, *Designing and conducting mixed methods research*. Sage publications, 2017.
[59] R. W. Emerson, 'Convenience sampling revisited: Embracing its limitations through thoughtful study design', *Journal of Visual Impairment & Blindness*, vol. 115, no. 1, pp. 76–77, 2021.
[60] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, 'User acceptance of computer technology: A comparison of two theoretical models', *Management science*, vol. 35, no. 8, pp. 982–1003, 1989.
[61] U. Ozolins, S. Hale, X. Cheng, A. Hyatt, and P. Schofield, 'Translation and back-translation methodology in health research–a critique', *Expert review of pharmacoeconomics & outcomes research*, vol. 20, no. 1, pp. 69–77, 2020.
[62] V. Braun and V. Clarke, 'Using thematic analysis in psychology', *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.
[63] J. F. Hair Jr, W. C. Black, B. J. Babin, and R. E. Anderson, 'Multivariate data analysis', in *Multivariate data analysis*, 2010, pp. 785–785.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 87–98
DOI: 10.15439/2025F8917
ISSN 2300-5963 ACSIS, Vol. 43

Digital Transformation in Saudi Public Universities: A Novel Framework for Adoption Drivers and Impact Analysis

Saleh Z Alshehri
0000-0002-0702-3382
dept. of Information Systems
King Khalid University
dept. of informatics
University of Sussex
Brighton, United Kingdom
Sa2142@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
dept. of Informatics
University of Sussex
Brighton, United Kingdom
N.Beloff@sussex.ac.uk

Martin White
0000-0001-8686-2274
dept. of Informatics
University of Sussex
Brighton, United Kingdom
Martinwh@sussex.ac.uk

***Abstract*—Saudi Arabia's Vision 2030 prioritises digital transformation (DX) to modernise higher education. However, despite significant investment, Saudi public universities (SPUs) face unique challenges in adoption due to a lack of context-specific frameworks. To address this gap, this study proposes and validates the novel DXA-SPU framework, an integrated model that combines the Technology Acceptance Model (TAM) and the Technology-Organisation-Environment (TOE) framework. The model was evaluated using survey data from 447 SPU participants, with hypothesised relationships analysed via Structural Equation Modelling (SEM). The results supported 12 of the 14 hypotheses. Perceived usefulness and the institutional skills gap emerged as the most significant drivers of adoption. In turn, DX adoption was strongly linked to enhanced institutional performance, administrative efficiency, technical infrastructure, and teaching effectiveness. The DXA-SPU framework offers a validated tool for university leaders to assess DX readiness and align strategic planning with Vision 2030 goals, providing actionable insights for policymakers.**

***Index Terms*—Digital transformation, adoption frameworks, TAM, TOE, Saudi public universities, Vision 2030, SEM.**

I. Introduction

DIGITAL transformation (DX) has become a top priority across various sectors, including higher education. In this context, it refers to the use of digital technologies to improve institutional operations, strategic planning and service quality [1], [2]. DX supports better performance in teaching, learning, research and administration [3], [4].

In Saudi Arabia, higher education is undergoing a significant transformation driven by national strategies, including Vision 2030 and the National Transformation Program (NTP). These initiatives aim to improve the quality and efficiency of public services, including universities, by promoting digital innovation [5]. The government's focus on digital initiatives underscores a national commitment to modernising higher education.

Despite strong governmental support and significant investment in digital infrastructure, the public sector still faces challenges that hinder DX adoption [6]. This persistence of obstacles highlights the complexity of implementing DX in public sector institutions.

Research indicates that DX in the public sector is influenced by more than mere access to technology. Political, social and institutional factors also play a significant role [7]. However, most existing studies focus on Western countries or the private sector. As a result, there is limited understanding of how DX is adopted in public universities in Saudi Arabia. Thus, a context-specific framework is necessary to address the unique challenges of DX adoption in Saudi Public Universities (SPUs).

This study introduces the DXA-SPU framework, Digital Transformation Adoption in Saudi Public Universities, a framework that integrates the Technology Acceptance Model (TAM) and the Technology-Organisation-Environment (TOE) framework to explore and evaluate the factors influencing DX adoption and its institutional impacts. The research addresses the following questions:

1. What organisational, technological, and environmental factors influence DX adoption in SPUs?
2. How does DX adoption impact institutional outcomes, such as performance and efficiency?
3. How can a tailored framework support SPUs leaders in aligning DX with Vision 2030?

Using a mixed-methods approach, the study first employs Structural Equation Modelling (SEM) to test hypothesised relationships using survey data from university participants. This paper presents the quantitative results and the theoretical underpinnings of the framework.

II. Literature Review

Digital transformation has become a global imperative, compelling organisations to fundamentally rethink their operating models and value propositions. In the realm of higher education, this transformation is not merely about adopting technology but involves a profound cultural and strategic shift

that redefines teaching, research, and administrative processes [8]. The COVID-19 pandemic acted as an unprecedented accelerator, forcing higher education institutions (HEIs) worldwide into a period of "emergency remote teaching" and compelling a digital transition at a scale and pace previously unimagined [9], [10].

However, research shows that digitalisation in HEIs is often fragmented. [11] describe this as "dual digitalisation": the unintegrated, parallel development of top-down administrative systems (like Learning Management Systems, or LMS) and bottom-up, subject-specific digital tools used by academics. This fragmentation creates a disjointed experience, preventing the formation of a cohesive digital learning space. Consequently, technology use in HEIs often remains superficial, with LMS platforms frequently used as simple content repositories for administrative convenience rather than as tools for innovative education [12], [13]. Many universities launch digital initiatives as isolated projects without an overarching, integrated strategy, which limits their impact and return on investment [14].

In Saudi Arabia, DX is a cornerstone of the nation's strategic roadmap, Saudi Vision 2030, which aims to diversify the economy and build a thriving, knowledge-based society [15]. The government has invested heavily in establishing a sophisticated digital infrastructure and has launched initiatives like the "Future Gate" project to digitise curricula and empower learning [10], [15]. The pandemic response further highlighted this commitment, as the Ministry of Education rolled out platforms like "Madrasti" and provided extensive support to ensure educational continuity [15].

Despite this strong governmental support and investment, SPUs face a distinct set of challenges that hinder the successful adoption of DX. A primary obstacle lies in human and cultural factors. [16] found that for instructors, the most significant barriers are attitudinal, including resistance to and fear of change, coupled with a lack of experience. This is compounded by a persistent digital skills gap among both faculty and students [17]. For DX to succeed, faculty require training that goes beyond mere tool usage to encompass new digital pedagogies, and the traditional reliance on rote learning must be replaced with methods that foster critical thinking and digital literacy [15].

From an organisational perspective, rigid institutional routines can inhibit the flexibility needed for transformation [18]. Successful DX requires clear leadership, a supportive organisational culture, a well-defined strategy, and adequate resources, factors often identified as lacking or underdeveloped in HEIs [17], [19]. Technically, while the national infrastructure is robust, challenges remain in ensuring reliable access for all, managing the explosion of institutional data through effective data governance, and addressing cybersecurity and privacy concerns [16], [20].

This complex landscape demonstrates that generic DX models are insufficient. There is a clear need for a context-specific framework tailored to the unique regulatory, cultural, and institutional conditions of SPUs. Such a framework must address the multifaceted challenges of human readiness, organisational agility, and strategic alignment with Vision 2030. This study proposes and validates the DXA-SPU framework to support SPUs' leadership in the strategic implementation of their university's DX.

III. Hypotheses and Study Framework

This study proposes the DXA-SPU framework to investigate the adoption of DX in SPUs, as illustrated in Fig. 1. The framework integrates two established models: the TAM and TOE frameworks. Together, they offer a combined view of

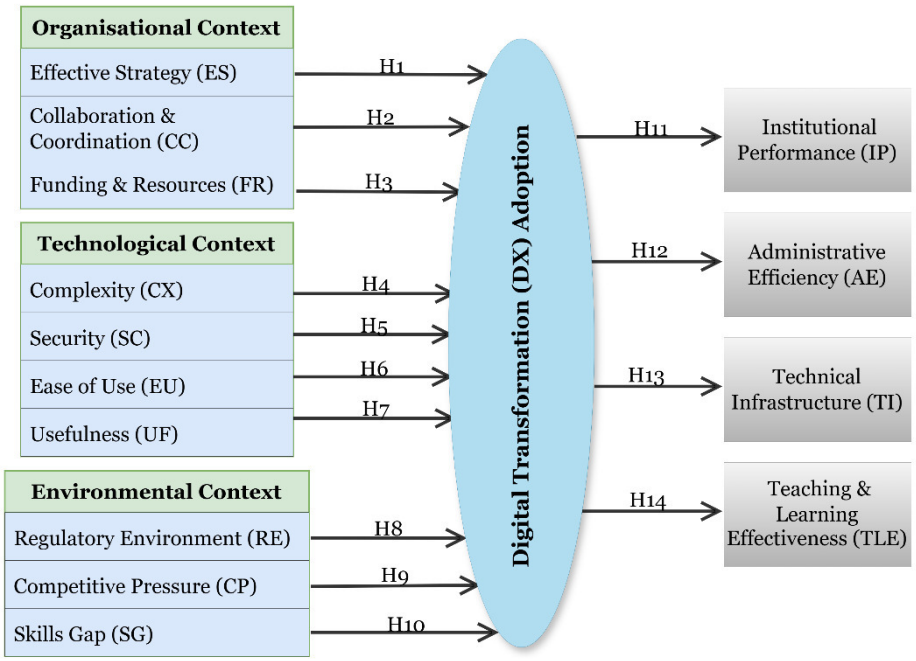


Fig. 1 The Initial Framework for DXA-SPU

individual and institutional influences on technology adoption. The aim is to identify the key factors shaping adoption and evaluate how DX is expected to impact university operations and performance.

A. Technology Acceptance Model (TAM)

The TAM posits that individuals are more likely to adopt a technology if they perceive it as useful and easy to use [21]. These core constructs, perceived usefulness and perceived ease of use, are central to predicting technology adoption behaviour. In this context, Saudi public university faculty, administrators and staff may support DX if it helps them work more efficiently and is accessible without added complexity.

However, TAM primarily focuses on individual-level perceptions and does not account for the broader institutional or environmental factors critical for analysing DX in universities. To address this, TAM is extended in this study to link the intention to adopt DX directly with perceived institutional impacts, such as performance and administrative efficiency, consistent with TAM's original structure connecting intention, behaviour, and outcomes [21]. The study examines whether this intention is associated with positive perceptions of institutional outcomes (H11-H14), using survey data to assess these links.

B. Technology-Organisation-Environment (TOE) Framework

The TOE framework provides a broader view of technology adoption by including factors from three domains [22] (see Table I):

Technological context: Characteristics of the technology itself, such as system complexity, ease of use and security.

Organisational context: Internal features of the institution, including leadership, planning, available resources and coordination across departments.

Environmental context: External influences like government policy (e.g., Vision 2030), regulatory pressures and competition from other universities.

The TOE complements the TAM by addressing the broader institutional and external conditions that affect adoption. Such perspectives are essential in public universities, where DX is shaped not only by individual preferences but also by policy, leadership, and organisational readiness.

C. DXA-SPU Framework Overview

Both the TAM and TOE frameworks are widely used for technology adoption. However, their application in isolation often falls short in complex environments like Saudi Public Universities [8].

TAM emphasises individual perceptions. This focus implies that it does not fully capture external forces, such as governmental policies or funding structures, that profoundly influence the adoption of DX in public institutions. Conversely, TOE addresses broader organisational and environmental factors. For example, it has been applied in specific regional contexts like Vietnam's creative industries [23]. However, TOE has been criticised for overlooking the crucial role of individual user perceptions and behavioural intention. These factors are strongly emphasised in technology acceptance models like TAM. They are also critical in digital learning adoption [24].

TABLE I.

DXA-SPU HYPOTHESES GROUPING SUMMARY

Group	Hypotheses	Focus
Organisational Context	H1-H4	Internal institutional factors: effective strategy, collaboration & coordination, funding & resources.
Technological Context	H5-H8	Technology-related factors: complexity, ease of use, usefulness and security.
Environmental Context	H9-H10	External influences: regulatory environment, competitive pressure, and skills gap
Impacts of DX Adoption	H11-H14	Institutional outcomes: institutional performance, administrative efficiency, technical infrastructure and teaching & learning effectiveness.

In universities, both levels are paramount. Faculty and staff exhibit varying levels of digital readiness and literacy. Institutional change is significantly shaped by national agendas like Vision 2030 [15], [25].

Recognising these limitations, the DXA-SPU framework integrates the strengths of both TAM and TOE. This provides a more comprehensive understanding of DX in this unique context. Various technology adoption models have been applied in sectors like e-learning, healthcare, and public administration. However, these models often suit general contexts or developed nations. They overlook the specific interplay of national policies and centralised governance in Saudi public education [15], [19], [20].

Critically, previous Saudi-focused studies have often been limited to specific applications, such as online learning or cloud systems. They often neglect holistic DX or key factors such as skills gaps. In contrast, the DXA-SPU framework explicitly incorporates constructs like 'Skills Gap' and 'Regulatory Environment'. These constructs are tailored to address the evolving digital literacy needs and policy alignment challenges in the Saudi higher education landscape. By integrat-

ing user and institutional perspectives, and explicitly accounting for regulatory pressures, funding dependencies, and digital skills deficiencies prevalent in Saudi Arabia, this framework offers a more effective approach for studying DX adoption in SPUs.

The DXA-SPU framework is designed to identify factors influencing DX adoption and evaluate its institutional impacts in SPUs. It integrates TAM to assess user perceptions and behavioural intention, and TOE to capture technological, organisational, and environmental influences.

Key constructs include ease of use, strategic alignment, and regulatory support, while outcome areas include institutional performance, administrative efficiency, technical infrastructure, and teaching effectiveness.

While this section outlines the structure and provides a detailed breakdown of hypotheses, constructs appear in later sections. Validation is performed using SEM.

D. Development of Hypotheses

Based on the TAM and TOE frameworks, the following hypotheses have been developed to explore the relationships between these factors and DX adoption in SPUs:

Organisational Context

The successful implementation of DX within a university is contingent upon key internal factors. These factors include an effective institutional strategy, robust inter-departmental collaboration, and the allocation of adequate resources.

a) Effective Strategy (ES)

A well-defined DX strategy ensures that technological initiatives are aligned with institutional goals and educational priorities. DX is not merely a standalone IT project but a cross-functional initiative that must integrate academic, administrative, and technological planning [26]. Multiple forms of strategy, including IT, change management and business strategy, must be synchronised to support successful digital initiatives [27]. When strategic alignment is lacking, institutions face fragmented efforts, inefficiencies and resistance to change [28]. In the context of higher education, a clear DX strategy enhances coordination, promotes long-term adaptability, and increases the likelihood of adoption success.

H1: *An effective DX strategy positively influences the intention to adopt DX in SPUS.*

b) Collaboration and Coordination (CC)

Collaboration among stakeholders, faculty, administrators and students is essential for the successful adoption of DX in universities. Research shows that engaging academic users helps ensure technologies are meaningfully integrated into teaching and learning [29]. Collaboration improves feedback, enhances digital literacy, and supports access to appropriate tools, which are often barriers to DX adoption [30]. Coordinated efforts across departments and governance levels also promote shared ownership of digital strategies and smooth implementation [31]. A collaborative environment increases acceptance and strengthens institutional readiness for DX.

H2: *Strong collaboration and coordination among stakeholders positively influence the intention to adopt DX in SPUs.*

c) Funding and Resources (FR)

Successful DX depends on access to sufficient financial, technical and human resources. Investments in infrastructure, software and skilled personnel form the foundation for sustainable implementation [32]. When funding is limited, institutions may struggle to upgrade systems or provide staff training, which delays adoption [33]. Training programs and digital tools also require ongoing investment to remain effective, making resource availability a long-term concern [34]. Thus, ensuring adequate support for DX projects is crucial to facilitating adoption across the university.

H3: *Sufficient funding and resources positively influence the intention to adopt DX in SPUs.*

Technological Context

Beyond organisational considerations, the intrinsic characteristics of the technology itself are instrumental to the success of DX. The adoption of these new digital tools is directly influenced by their complexity and security, as well as their perceived usefulness and ease of use.

a) Complexity (CX)

Technological complexity is a significant barrier to DX, especially when modern systems must integrate with legacy infrastructure. Incompatibility between platforms can lead to data silos, operational disruptions and increased transition costs [36]. Complex implementation processes can disrupt workflows and reduce staff willingness to adopt new tools [37]. Without clear integration pathways, institutions may face delays, increased risks and unmet goals. Reducing complexity helps build user confidence and simplifies institutional decision-making.

H4: *Lower complexity in integrating digital technologies with existing systems and processes positively influences the intention to adopt DX in SPUs.*

b) Security (SC)

As universities become more reliant on digital platforms, concerns over cybersecurity and data protection have grown. The use of cloud services, online portals, and connected devices introduces multiple risks, including data breaches and ransomware attacks [38]. Without strong security measures, universities may face legal, operational and reputational consequences [39]. A secure digital environment builds trust among users and reduces resistance to adopting new digital technologies. Therefore, confidence in data security is a critical enabler of DX adoption.

H5: *A higher level of digital security positively influences the intention to adopt DX in SPUs.*

c) Ease of Use (EU)

According to TAM, ease of use is a core predictor of technology adoption [21]. When digital tools are perceived as intuitive and user-friendly, individuals are more likely to explore and apply them in their work. Ease of use also influences trust and satisfaction, both of which contribute to long-term engagement [40]. In the context of higher education, simple interfaces reduce learning curves for faculty and staff and encourage quicker adoption.

H6: *The perceived ease of use of DX solutions positively influences the intention to adopt DX in SPUs.*

d) Usefulness (UF)

Perceived usefulness is among the strongest predictors of technology adoption in the TAM framework. In higher education, digital tools are more likely to be adopted if users believe they improve teaching, learning and administrative efficiency [41]. Technologies that support personalised instruction and streamline processes are especially valued, as they contribute to both student outcomes and institutional performance [42]. When users recognise the practical benefits of DX, they are more motivated to support its adoption.

H7: *The perceived usefulness of DX positively influences the intention to adopt DX in SPUs.*

Environmental Context

The external environment, heavily influenced by government directives such as Vision 2030, is a primary driver of DX in SPUs. National policies create a direct regulatory pressure for modernisation, while competitive dynamics between institutions also foster innovation. At the same time, the evolving Saudi labour market demands graduates with advanced digital skills. This pressure for skilled graduates, combined with the need for internal faculty and staff to be digitally competent, creates a significant challenge for universities to address the overall skills gap.

a) Regulatory Environment (RE)

The regulatory environment plays a vital role in enabling or restricting DX in public institutions. In Saudi Arabia, higher education policies and budgetary frameworks strongly influence the scope and pace of DX adoption. Supportive government policies, such as Vision 2030, encourage investment in digital tools, infrastructure and human capital [43]. Effective regulation helps create the conditions for innovation, whereas rigid or unclear policies can limit institutional flexibility [23]. A favourable regulatory climate increases confidence and motivates universities to invest in digital initiatives.

H8: *A flexible and supportive regulatory environment positively influences the intention to adopt DX in SPUs.*

b) Competitive Pressure (CP)

Competitive pressure is a key external factor that drives innovation in higher education. When universities compete for rankings, visibility or recognition, they are more likely to adopt advanced digital tools to enhance teaching, research and administration [44]. Institutions that seek to differentiate themselves invest more in technology, which can improve student outcomes and faculty productivity [45]. In Saudi Arabia, competition among public universities creates momentum for DX, pushing institutions to modernise quickly to remain relevant.

H9: *Higher competitive pressure among universities positively influences the intention to adopt DX in SPUs.*

c) Skills Gap (SG)

A major challenge facing DX in higher education is the skills gap among students, faculty and staff. As digital tools become central to education and work, institutions must address deficits in both technical and soft skills to ensure readiness for transformation [46]. Employers increasingly expect digital fluency, critical thinking and adaptability, which are not always part of traditional curricula [47].

Universities that align their teaching and training with evolving workforce needs will be better positioned to adopt digital technologies successfully. Addressing the skills gap enhances both institutional effectiveness and graduate preparedness [48], [49]. Thus, bridging this gap is essential to ensuring DX efforts can be successfully implemented and sustained.

H10: *Addressing digital skills gaps positively influences the intention to adopt DX in SPUs.*

E. The Impact of DXA on SPUs

This study extends the TAM to link the intention to adopt DX with its perceived impacts on university outcomes. These impacts are framed as expected outcomes within a mixed-methods approach, where the quantitative phase tests the hypothesised relationships using survey data.

The hypotheses aim to investigate how various technological, organisational and environmental factors influence the intention to adopt DX, and how this intention relates to outcomes such as institutional performance, administrative efficiency, technical infrastructure and teaching and learning effectiveness.

The DXA-SPU framework, by combining the TAM and TOE, offers a robust model for understanding the drivers of DX in SPUs. It captures both individual perceptions and broader institutional dynamics, making it a comprehensive tool for analysing adoption behaviour and its anticipated impacts. These hypotheses were tested to assess the relationships between influencing factors, adoption intentions and institutional outcomes, which are described below.

Institutional Performance (IP)

Digital transformation plays a growing role in enhancing institutional performance across universities globally, including Saudi public institutions. Studies in Saudi Arabia confirm that DX can improve efficiency, leadership, research productivity and stakeholder engagement. For example, [50] shows that DX supports academic and financial performance and promotes sustainability through digital learning. Reference [18] highlights the role of internal processes and routines in

shaping transformation outcomes, while [51] connects leadership styles to DX success. Data governance, as emphasised by [20], is also essential for institutional performance. Moreover, as in [19], the study links DX to broader social and academic outcomes, including student engagement and well-being. Collectively, these studies indicate that DX contributes to performance through operational, human and technological improvements.

H11: *Adopting DX in SPUs is expected to positively enhance institutional performance.*

Administrative Efficiency (AE)

Digital transformation supports more efficient university operations by streamlining workflows, integrating digital tools, and enhancing stakeholder engagement. Research from Saudi universities shows that knowledge management systems and cloud-based solutions can enhance administrative agility and service quality [53], [54]. DX allows institutions to reduce technical errors, improve decision-making, and strengthen accountability through real-time data access [55], [56]. Studies by [57], [58] show that user engagement, trust and service efficiency improve when universities adopt digital public administration tools. These technologies also enable the modernisation of resource management and coordination, which is essential for competitiveness and institutional effectiveness.

H12: *Adopting DX in SPUs is expected to positively enhance administrative efficiency.*

Technical Infrastructure (TI)

The shift toward DX has accelerated infrastructure development in SPUs. Institutions have expanded their Information and Communication Technology (ICT) capabilities, especially in response to the COVID-19 pandemic, which underscored the need for resilient e-learning platforms and remote service access [59], [60]. Hence, robust infrastructure is critical for sustaining digital tools across academic and administrative systems [45], [61]. Moreover, Supportive ICT policies and resource availability influence the digital readiness of universities and their ability to scale technologies effectively [62]. Therefore, investing in reliable and flexible infrastructure is essential for enabling long-term transformation.

H13: *Adopting DX in SPUs is expected to positively enhance technical infrastructure.*

Teaching and Learning Effectiveness (TLE)

Digital transformation significantly shapes the effectiveness of teaching and learning in higher education. In SPUs, DX enables more interactive, student-centred learning environments, especially through online platforms and blended learning methods [63]. Investments in digital hardware, content and staff training have improved e-learning quality and digital literacy [64]. Studies also show that DX positively influences student engagement, professional development and learning continuity during crises [65]. While DX introduces new challenges, such as digital well-being and emotional adjustment, its benefits in learning outcomes and instructional quality are well-documented [66]. Overall, DX supports a more flexible, engaging and effective teaching and learning environment that meets evolving student and institutional needs.

H14: *Adopting DX in SPUs is expected to positively enhance teaching and learning effectiveness.*

IV. Study Methodology

A. Approach and Sample

This study uses a mixed-methods approach to evaluate the DXA-SPU framework. The quantitative phase tests relationships using survey data, while a subsequent qualitative phase is planned to explore contextual interpretations through interviews with IT managers.

The survey focused on three groups within SPUs: faculty members, who include academic teaching and research staff; administrators, who hold strategic or managerial positions with university-wide decision-making authority; and staff, which encompasses all other non-faculty employees, including support and technical personnel. A stratified sampling strategy was applied to ensure role-based representation. According to the General Authority for Statistics, the target population comprises approximately 134,000 members [67], [68]. Using a 95% confidence level and 5% margin of error, the minimum required sample size was 384. In 2024, the survey link was distributed to a random sample, yielding 447 valid responses, exceeding the threshold for robust analysis.

B. Survey Implementation

The primary objective of the survey was to examine the relationships between the DXA-SPU framework constructs and the intention to adopt DX in SPUs. The questionnaire consisted of three parts: Part I (7 questions) gathered demographic and contextual data, Part II (40 questions) assessed ten influencing factors (e.g., Collaboration and Coordination, Ease of Use), and Part III (16 questions) evaluated four impact constructs (e.g., Teaching and Learning Effectiveness). Each construct was measured with five Likert-scale questions to support reliable statistical analysis. The survey instrument is accessible for review online at [Survey for SPU].

A pilot study was conducted prior to the main survey to test the clarity and reliability of the questionnaire. The pilot involved a small group of university stakeholders, targeting 20 responses, with 25 responses received. Feedback led to minor adjustments in wording and layout to enhance clarity, but no substantive changes to the questions were required.

The main survey was conducted from August 15 to November 15, 2024, using an online survey platform. The instrument was developed in English and translated into Arabic to accommodate participant preferences. Over 700 questionnaires were randomly distributed to faculty, administrative, and staff members across SPUs. Of these, 551 questionnaires were completed, but only 447 (81%) responses were accurate and suitable for analysis, while 104 (19%) were deemed invalid.

Participants were informed of the study's purpose, and confidentiality was assured prior to participation. Ethical ap-

proval was obtained from the university research ethics committee, and informed consent was secured from all participants. No personal identifiers were collected, and data were securely stored for academic research purposes only.

The quantitative data were analysed using SEM with IBM AMOS 29, incorporating Confirmatory Factor Analysis (CFA) to evaluate the measurement model. Internal consistency was assessed using Cronbach's Alpha, convergent validity using Composite Reliability (CR) and Average Variance Extracted (AVE), and discriminant validity using Maximum Shared Variance (MSV). Model fit was evaluated through indices such as χ^2/df (CMIN/DF), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardised Root Mean Square Residual (SRMR).

V. Study Results

A. Data Collection and Screening

Table II. details the demographic profile of the respondents, which shows a diverse group from SPUs. Most participants were female (53.2%), while males made up 45.4%. A small number (1.3%) preferred not to state their gender. The largest age group was 31-45 years (73.2%), followed by those aged 18-30 (19.7%).

Regarding education, 49.7% held a master's degree and 36.2% held a doctorate. This suggests that most respondents were well educated. Regarding their roles, 67.3% were faculty members, 16.8% were staff, and 15.9% worked in administration. Almost half (49.0%) had over ten years of experience at their university. Others had 5-10 years (27.0%) or 1-4 years (21.0%) of experience.

A majority (66.7%) were aware of national DX programs such as the Vision 2030 NTP. Similarly, 69.1% were familiar with their university's DX efforts.

Many participants viewed DX as effective. Approximately 34.7% reported that it was very effective, and 29.5% considered it moderately effective. Satisfaction levels were also high. Around 41.9% were somewhat satisfied, while 12.0% were extremely satisfied.

Overall, the respondents were experienced, informed, and qualified to provide feedback on the adoption of DX.

Data screening confirmed the dataset's suitability for analysis. The Z-score analysis revealed no significant outliers, and the skewness and kurtosis values fell within ±2.0, indicating normality. Missing values were below 10%, requiring no imputation [69]. These checks validated the integrity of the dataset for SEM analysis.

B. Measurement Model Assessment

Confirmatory Factor Analysis was used to assess the validity and reliability of the measurement model. All standardised factor loadings exceeded 0.50, indicating that the items aligned with their respective constructs.

Table III shows that internal consistency was verified using Cronbach's Alpha and CR, with all values above the 0.70 threshold. Convergent validity was supported through AVE, with values exceeding 0.50. Discriminant validity was established as all constructs showed AVE values greater than the MSV.

Model fit indices also indicated an acceptable model fit:

- χ^2/df = 1.445
- CFI = 0.968

TABLE II.

DEMOGRAPHIC STATISTICS FOR THE PARTICIPANTS

Gender	**N**	**%**
Male	203	45.4%
Female	238	53.2%
Prefer not to say	6	1.3%
Age	**N**	**%**
18 - 30	88	19.7%
31 - 45	327	73.2%
46 - 60	28	6.3%
60+	4	0.9%
Position	**N**	**%**
Administrator	71	15.9%
Faculty	301	67.3%
Staff	75	16.8%
Years of Experience	**N**	**%**
1-4	94	21.0%
5-10	121	27.0%
10+	221	49.0%
Education Level	**N**	**%**
Diploma	13	2.9%
Bachelor	50	11.2%
Master	222	49.7%
Doctorate	162	36.2%
DX Impact	**N**	**%**
Not effective at all	40	8.9%
Slightly effective	68	15.2%
Moderately effective	132	29.5%
Very effective	155	34.7%
Extremely effective	52	11.6%
DX Awareness	**N**	**%**
Yes	309	69.1%
No	138	30.9%
Approach Satisfaction	**N**	**%**
Extremely	19	4.3%
Somewhat	71	16.1%
Neither satisfied nor dissatisfied	72	16.4%
Somewhat satisfied	184	41.2%
Extremely satisfied	54	12.1%

TABLE III.

CONSTRUCT VALIDITY AND RELIABILITY

Constructs	Cronbach (above 0.7)	CR	AVE	MSV
ES	.952	0.952	0.833	0.130
CC	.923	0.923	0.751	0.045
FR	.816	0.817	0.528	0.123
CX	.872	0.877	0.643	0.130
SC	.902	0.903	0.701	0.102
EU	.910	0.910	0.717	0.264
UF	.914	0.915	0.730	0.308
RE	.938	0.939	0.793	0.088
CP	.907	0.907	0.709	0.278
SG	.931	0.931	0.772	0.306
IP	.918	0.918	0.737	0.585
AE	.925	0.925	0.755	0.555
TI	.920	0.921	0.744	0.585
TLE	.916	0.919	0.741	0.575
ES	.952	0.952	0.833	0.130

- RMSEA = 0.032
- NFI = 0.902

These results indicate that the measurement model is robust and well-constructed.

C. Structural Model Assessment

The structural model tested the hypothesised relationships between the factors influencing the intention to adopt DX and the perceived impacts of adoption. Path coefficients and significance levels were estimated using SEM.

Table IV presents the results of the structural model. Among the adoption factors (H1-H10), *Effective Strategy (ES)* showed a significant positive effect on DX adoption ($\beta = 0.121$, $p = 0.001$). However, *Collaboration and Coordination (CC)* showed no significant effect on DX adoption ($\beta = -0.009$, $p = 0.773$). This unexpected result may stem from limited cross-departmental structures in SPUs, as bureaucratic silos could hinder collaborative efforts [70]. In addition, the measurement of the CC construct might not fully reflect the dynamics among stakeholders. This indicates that future studies may benefit from developing more refined survey items.

Moreover, *Complexity (CX)* had a significant positive effect ($\beta = 0.181$, $p < 0.001$), suggesting that manageable complexity supports adoption. *Funding and Resources (FR)* was also significant ($\beta = 0.184$, $p = 0.002$), showing the importance of institutional support. *Security (SC)* ($\beta = 0.131$, $p = 0.003$) and *Ease of Use (EU)* ($\beta = 0.103$, $p = 0.047$) both had significant positive impacts. *Usefulness (UF)* was the strongest predictor ($\beta = 0.272$, $p < 0.001$), highlighting its significant role. *Regulatory Environment (RE)* was also significant ($\beta = 0.091$, $p = 0.009$).

TABLE IV.

RESULTS OF ANALYSIS OF HYPOTHESIS PATH FOR THE DXA-SUPS FRAMEWORK

Hypothesis	Structural Relation	Regression Weight	Standard Error (S.E.)	Critical Ratio (C.R.)	P Value
H1	DX ← ES	.121	.037	3.256	.001
H2	DX ← CC	-.009	.031	-.288	.773
H3	DX ← FR	.184	.059	3.099	.002
H4	DX ← CX	.181	.054	3.364	***
H5	DX ← SC	.131	.044	2.995	.003
H6	DX ← EU	.103	.052	1.983	.047
H7	DX ← UF	.272	.049	5.539	***
H8	DX ← RE	.091	.035	2.613	.009
H9	DX ← CP	.100	.055	1.822	.068
H10	DX ← SG	.239	.052	4.601	***
H11	IP ← DX	.629	.043	14.491	***
H12	AE ← DX	.753	.047	16.004	***
H13	TI ← DX	.683	.046	14.990	***
H14	TLE ← DX	.639	.049	13.165	***

NOTE: ("***=SIGNIFICANCE AT THE 0.001 LEVEL")

Likewise, *Competitive Pressure (CP)* approached significance ($\beta = 0.100$, $p = 0.068$). While not statistically significant at the conventional 0.05 level, this finding suggests a marginal influence. The *Skills Gap (SG)* showed a strong positive effect ($\beta = 0.239$, $p < 0.001$), underscoring the role of digital competencies.

All four impact hypotheses (H11-H14) were supported. *Institutional Performance (IP)* was significantly predicted by DX adoption ($\beta = 0.629$, $p < 0.001$), followed by strong effects on *Administrative Efficiency (AE)* ($\beta = 0.753$, $p < 0.001$), *Technical Infrastructure (TI)* ($\beta = 0.683$, $p < 0.001$), and *Teaching and Learning Effectiveness (TLE)* ($\beta = 0.639$, $p < 0.001$). These findings suggest that DX adoption is associated with broad and substantial institutional benefits.

Overall, twelve out of fourteen hypotheses were supported. The results empirically support the DXA-SPU framework, demonstrating its relevance to SPUs.

D. Summary of Findings

The results of the quantitative analysis provide strong empirical support for the DXA-SPU framework. Perceived usefulness, skills gap and complexity emerged as the most influential factors, while collaboration and coordination, and competitive pressure did not show a significant effect in this context.

The extension of the TAM was also empirically supported, as the intention to adopt DX was found to significantly impact all four key outcome areas: institutional performance, administrative efficiency, technical infrastructure and teaching and learning effectiveness. Among these, administrative efficiency showed the strongest relationship, suggesting that DX is especially effective in enhancing internal operational processes.

These findings not only reinforce the relevance of the TAM and TOE in higher education settings but also highlight the strategic role of DX in improving institutional outcomes.

VI. Discussion

The results support the relationships proposed by the DXA-SPU framework, demonstrating how organisational, technological, and environmental factors influence adoption (see Fig. 2). This outcome strengthens the theoretical validity of integrating TAM and TOE for analysing DX adoption.

These findings extend the applicability of TAM and TOE in higher education settings, providing empirical support for the framework's multidimensional design. Such validation offers practical and theoretical contributions, particularly in non-Western, public-sector contexts, such as Saudi Arabia.

In the organisational context, effective strategy and funding and resources significantly influenced adoption, consistent with prior research emphasising the importance of institutional planning and leadership commitment [71], [72]. However, the factor of collaboration and coordination was not supported, indicating that cross-departmental integration may not yet be strong or institutionalised enough to shape adoption behaviours. This suggests an opportunity for organisational development initiatives that promote structured collaboration and digital governance.

Within the technological dimension, perceived usefulness and ease of use emerged as strong predictors of adoption intention, reinforcing the applicability of the TAM in the higher education context. The finding that lower system complexity positively influences adoption underscores the need for intuitive, interoperable digital systems that minimise disruption during adoption. The significance of security also reflects increasing institutional awareness of data protection and privacy challenges in digital environments [73].

Regarding environmental influences, the regulatory environment demonstrated statistical significance, reaffirming that national policies, especially those tied to Vision 2030, are powerful motivators for DX in SPUs. In contrast, competitive pressure approached significance but was not statistically supported at the conventional level. While some literature suggests that competitive dynamics drive innovation, the insignificant finding in this specific context may imply that, for SPUs operating under strong governmental directives, top-down policy mandates (such as Vision 2030) currently exert a more direct and statistically dominant influence on DX adoption than inter-university competition. Meanwhile, the influence of the skills gap confirms that digital readiness

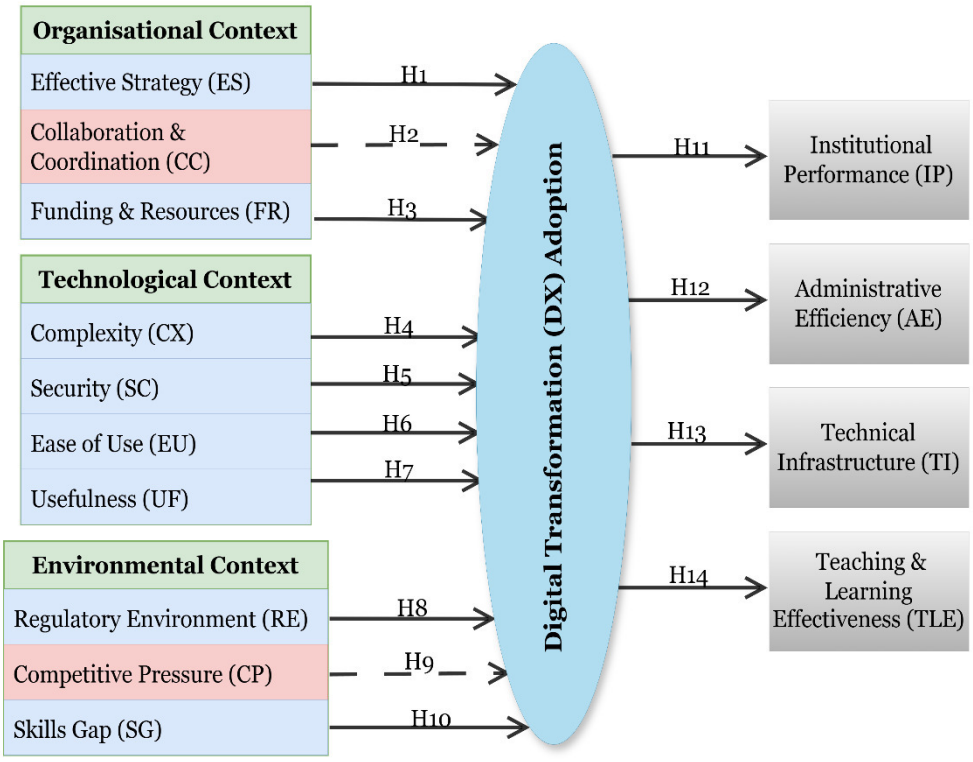


Fig. 2 Fig. 2 The Revised Version of the Comprehensive Framework for DXA-SPU

among staff and students remains a challenge, despite strategic alignment at the policy level [74].

The extended TAM results confirm that adoption intention is strongly associated with multiple institutional outcomes. These include improvements in operational efficiency, infrastructure resilience, and academic delivery, reinforcing the strategic significance of DX adoption in the SPU context.

Overall, these results underscore that successful DX adoption depends not only on the availability of technology but also on strategic planning, institutional preparedness and alignment with national reform agendas.

VII. Conclusion

This study provides empirical support for the novel DXA-SPU framework, integrating the TAM and TOE frameworks to examine DXA in SPUs. Survey data from 447 participants supported 12 of 14 hypotheses, identifying perceived usefulness, ease of use, regulatory alignment, and skills gap as key adoption drivers. DX adoption significantly enhances institutional performance, administrative efficiency, technical infrastructure, and teaching effectiveness, aligning with Vision 2030's educational reform goals.

The non-significant collaboration and coordination (CC) and competitive pressure (CP) hypotheses suggest a need for stronger interdepartmental mechanisms, such as cross-departmental task forces, to foster stakeholder engagement. Collectively, they indicate that internal organisational factors or top-down strategic mandates may currently be more critical drivers of DX adoption in this context.

The DXA-SPU framework offers significant contributions to both theory and practice. Theoretically, it adapts and combines the TAM and TOE frameworks to specifically address the unique context of SPUs, reflecting their national goals, institutional structures, and digital priorities. Practically, it provides an empirically supported tool for SPUs leaders to assess DX readiness and align their strategies with Vision 2030.

This study is based on self-reported perceptions, offering valuable insights into behavioural intention but not fully capturing actual adoption. Although the sample size was statistically sufficient, it may not fully represent the full diversity of stakeholder perspectives across all SPUs. To address these limitations, the following research phase will involve completing the qualitative phase by October 2025, with interviews of 10-15 IT managers to validate the findings and explore implementation dynamics.

While this framework was designed and validated in the context of SPUs, it may also be applicable in similar contexts in other countries. Moreover, a comparative analysis of the results with findings from other cultures would be a particularly valuable extension to assess the framework's broader applicability. Such research could further support the development of effective DX strategies and inform national education policy planning initiatives.

References

[1] G. C. Kane, D. Palmer, A. N. Phillips, D. Kiron, and N. Buckley, "Strategy, not technology, drives digital transformation," *MIT Sloan Manag Rev*, 2015.

[2] D. Tang, "What is digital transformation?," *EDPACS*, vol. 64, no. 1, pp. 9–13, 2021.

[3] OECD, "Saudi Arabia," in *in Education at a Glance 2017: OECD Indicators*, Paris: OECD Publishing, 2017. doi: 10.1787/eag-2017-82-en.

[4] I. Mergel, N. Edelmann, and N. Haug, "Defining digital transformation: Results from expert interviews," *Gov Inf Q*, vol. 36, no. 4, p. 101385, 2019, doi: https://doi.org/10.1016/j.giq.2019.06.002.

[5] M. Alonazi, "MGAUM: a new framework for the mobile government service adoption in Saudi Arabia.," University of Sussex, 2019. doi: doi.org/10.5281/zenodo.1316598.

[6] A. C. A. Viana, "Digital transformation in public administration: from e-Government to digital government," *International Journal of digital law*, no. 1, pp. 29–44, 2021.

[7] L. Stark, D. Greene, and A. L. Hoffmann, "Critical perspectives on governance mechanisms for AI/ML systems," *The cultural life of machine learning: An incursion into critical AI studies*, pp. 257–280, 2021.

[8] L. M. C. Benavides, J. A. Tamayo Arias, M. D. Arango Serna, J. W. Branch Bedoya, and D. Burgos, "Digital transformation in higher education institutions: A systematic literature review," *Sensors*, vol. 20, no. 11, p. 3291, 2020.

[9] V. J. García-Morales, A. Garrido-Moreno, and R. Martín-Rojas, "The Transformation of Higher Education After the COVID Disruption: Emerging Challenges in an Online Learning Scenario," *Front Psychol*, vol. Volume 12-2021, 2021, [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.616059

[10] H. Abdulrahim and F. Mabrouk, "COVID-19 and the digital transformation of Saudi higher education.," *Asian Journal of Distance Education*, vol. 15, no. 1, pp. 291–306, 2020.

[11] B. Bygstad, E. Øvrelid, S. Ludvigsen, and M. Dæhlen, "From dual digitalization to digital learning space: Exploring the digital transformation of higher education," *Comput Educ*, vol. 182, p. 104463, Feb. 2022, doi: 10.1016/j.compedu.2022.104463.

[12] A. Aldiab, H. Chowdhury, A. Kootsookos, F. Alam, and H. Allhibi, "Utilization of Learning Management Systems (LMSs) in higher education system: A case review for Saudi Arabia," *Energy Procedia*, vol. 160, pp. 731–737, 2019, doi: https://doi.org/10.1016/j.egypro.2019.02.186.

[13] M. Bond, V. I. Marín, C. Dolch, S. Bedenlier, and O. Zawacki-Richter, "Digital transformation in German higher education: student and teacher perceptions and usage of digital media," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, p. 48, 2018, doi: 10.1186/s41239-018-0130-1.

[14] A. Fernández, B. Gómez, K. Binjaku, and E. K. Meçe, "Digital transformation initiatives in higher education institutions: A multivocal literature review," *Educ Inf Technol (Dordr)*, vol. 28, no. 10, pp. 12351–12382, 2023, doi: 10.1007/s10639-022-11544-0.

[15] A. Alghamdi, "Digital Transformation within Saudi Education System: 2020 and Beyond," vol. 6, pp. 419–425, Sep. 2022, doi: 10.26855/er.2022.08.014.

[16] A. Alhubaishy and A. Aljuhani, "The Challenges of Instructors' and Students' Attitudes in Digital Transformation: A Case Study of Saudi Universities," *Educ Inf Technol (Dordr)*, vol. 26, no. 4, pp. 4647–4662, 2021, doi: 10.1007/s10639-021-10491-6.

[17] T. M. Khawaji, "Digital Transformation Requirements at Saudi Universities from Faculty Members' Perspectives," *International Journal of Computer Science & Network Security*, vol. 23, no. 11, pp. 8–20, 2023.

[18] I. Almatrodi and D. Skoumpopoulou, "Organizational Routines and Digital Transformation: An Analysis of How Organizational Routines Impact Digital Transformation Transition in a Saudi University," *Systems*, vol. 11, no. 5, p. 239, 2023, doi: 10.3390/systems11050239.

[19] M. Alenezi and M. Akour, "Digital Transformation Blueprint in Higher Education: A Case Study of PSU," *Sustainability*, vol. 15, no. 10, p. 8204, 2023, doi: 10.3390/su15108204.

[20] A. Omar and A. almaghthawi, "Towards an Integrated Model of Data Governance and Integration for the Implementation of Digital Transformation Processes in the Saudi Universities," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020, doi: 10.14569/ijacsa.2020.0110873.

[21] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[22] L. G. Tornatzky, M. Fleischer, and A. K. Chakrabarti, "The processes of technological innovation," *(No Title)*, 1990.

[23] T. V. H. Trieu and D. Pavelková, "Differences of education level and job position in digital transformation adoption in Vietnam's creative industries," *Journal of Eastern European and Central Asian Research (JEECAR)*, vol. 9, no. 3, pp. 409–421, 2022.

[24] A. Al-Motrif, "Digital learning in Saudi University: evaluating digital transformation post Covid-19," *Technol Anal Strateg Manag*, vol. 36, no. 12, pp. 4440–4454, Dec. 2024, doi: 10.1080/09537325.2023.2255292.

[25] A. Allmnakrah and C. Evers, "The need for a fundamental shift in the Saudi education system: Implementing the Saudi Arabian economic vision 2030," *Research in Education*, vol. 106, no. 1, pp. 22–40, 2020.

[26] F. J. M. Veiga and A. M. V. de Andrade, "Critical Success Factors in Accepting Technology in the Classroom," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 18, pp. 4–22, Sep. 2021, doi: 10.3991/ijet.v16i18.23159.

[27] K. Sarnok, P. Wannapiroon, and P. Nilsook, "DTL-eco system by digital storytelling to develop knowledge and digital intelligence for teacher profession students," *International Journal of Information and Education Technology*, vol. 10, no. 12, pp. 865–872, 2020.

[28] F. Mahmood, A. Z. Khan, and M. B. Khan, "Digital organizational transformation issues, challenges and impact: A systematic literature review of a decade.," *Abasyn University Journal of social sciences*, vol. 12, no. 2, 2019.

[29] A. W. Bates, *Teaching in a digital age: Guidelines for designing teaching and learning*. BCcampus, 2015.

[30] M. Warschauer, "Computer-assisted language learning: An introduction," *Multimedia language teaching*, vol. 320, 1996.

[31] M. Meyerhoff Nielsen, "Governance lessons from Denmark's digital transformation," in *Proceedings of the 20th annual international conference on digital government research*, 2019, pp. 456–461.

[32] J. K. Nwankpa and Y. Roumani, "IT capability and digital transformation: A firm performance perspective," 2016.

[33] N. Verina and J. Titko, "Digital transformation: conceptual framework," in *Proc. of the Int. Scientific Conference "Contemporary Issues in Business, Management and Economics Engineering*, 2019, pp. 9–10.

[34] G. Vial, "Understanding digital transformation: A review and a research agenda," *The journal of strategic information systems*, vol. 28, no. 2, pp. 118–144, 2019.

[35] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS quarterly*, pp. 157–178, 2012.

[36] E. Henriette, M. Feki, and I. Boughzala, "Digital transformation challenges," 2016.

[37] T. Hess, C. Matt, A. Benlian, and F. Wiesböck, "Options for Formulating a Digital Transformation Strategy," *MIS Quarterly Executive*, vol. 15, pp. 123–139, Jun. 2016.

[38] L. S. Rodrigues, "Challenges of digital transformation in higher education institutions: A brief discussion," in *Proceedings of 30th IBIMA Conference*, 2017.

[39] J. Li, W. Xiao, and C. Zhang, "Data security crisis in universities: Identification of key factors affecting data breach incidents," *Humanit Soc Sci Commun*, vol. 10, no. 1, pp. 1–18, 2023.

[40] Y. Chen and S. Barnes, "Initial trust and online buyer behaviour," *Industrial management & data systems*, 2007.

[41] S. Hu, K. Laxman, and K. Lee, "Exploring factors affecting academics' adoption of emerging mobile technologies-an extended UTAUT perspective," *Educ Inf Technol (Dordr)*, vol. 25, pp. 4615–4635, 2020.

[42] A. Kitsantas and N. Dabbagh, "The role of Web 2.0 technologies in self-regulated learning," *New directions for teaching and learning*, vol. 2011, no. 126, pp. 99–106, 2011.

[43] H. Al-Mawali, A. Al Sharif, G. M. A. Rumman, F. Kerzan, and G. Liu, "Environmental strategy, environmental management accounting and organizational performance: evidence from the United Arab Emirates market," *Journal of Environmental Accounting and Management*, vol. 6, no. 2, pp. 109–118, 2018.

[44] D. T. Thao, "DIGITAL TRANSFORMATION IN EDUCATION AT THAI NGUYEN UNIVERSITY OF TECHNOLOGY," 2023, doi: 10.5281/zenodo.7858927.

[45] M. Alojail, J. Alshehri, and S. B. Khan, "Critical Success Factors and Challenges in Adopting Digital Transformation in the Saudi Ministry of Education," *Sustainability*, vol. 15, no. 21, p. 15492, 2023, doi: 10.3390/su152115492.

[46] K. Dengler and B. Matthes, "The impacts of digital transformation on the labour market: Substitution potentials of occupations in Germany," *Technol Forecast Soc Change*, vol. 137, pp. 304–316, 2018, doi: https://doi.org/10.1016/j.techfore.2018.09.024.

[47] V. Bikse, L. Grinevica, B. Rivza, and P. Rivza, "Consequences and Challenges of the Fourth Industrial Revolution and the Impact on the development of Employability Skills," *Sustainability*, vol. 14, no. 12, p. 6970, 2022.

[48] A. Morgan, R. Sibson, and D. Jackson, "Digital demand and digital deficit: conceptualising digital literacy and gauging proficiency among higher education students," *Journal of Higher Education Policy and Management*, vol. 44, no. 3, pp. 258–275, May 2022, doi: 10.1080/1360080X.2022.2030275.

[49] B. A-khateeb, *Technological Skills and Job Employment in Universities in Saudi Arabia*. 2020. doi: 10.33422/6th.omeaconf.2020.09.209.

[50] N. S. Alotaibi, "The Significance of Digital Learning for Sustainable Development in the Post-Covid19 World in Saudi Arabia's Higher Education Institutions," *Sustainability*, vol. 14, no. 23, p. 16219, 2022, doi: 10.3390/su142316219.

[51] G. S. Alessa, "The Dimensions of Transformational Leadership and Its Organizational Effects in Public Universities in Saudi Arabia: A Systematic Review," *Front Psychol*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.682092.

[52] N. Hassan, A. Hussain, M. A. Bhatti, and T. Ahmad, "Perceived Administrative Service Efficiency in Academia: A Case Study of International Islamic University Islamabad," *Irasd Journal of Management*, vol. 4, no. 2, pp. 449–464, 2022, doi: 10.52131/jom.2022.0402.0091.

[53] M. Alsharyofi, "The Role of Knowledge Management in Raising the Saudi Universities Performance Efficiency: An Applied Study on the Administrative Body of King Abdulaziz University," *Ijiis International Journal of Informatics and Information Systems*, vol. 5, no. 2, pp. 101–107, 2022, doi: 10.47738/ijiis.v5i2.133.

[54] N. H. Thanh, "Digital Transformation: Smart Strategy in Administrative Reform in Vietnam," *Hightech and Innovation Journal*, vol. 2, no. 4, pp. 328–345, 2021, doi: 10.28991/hij-2021-02-04-06.

[55] C. T. D. Tran, B. Dollery, and S. R. Yarram, "The Influence of Administrative Intensity on Efficiency: An Empirical Analysis of Australian Universities," *Economic Papers a Journal of Applied Economics and Policy*, vol. 42, no. 3, pp. 282–305, 2023, doi: 10.1111/1759-3441.12387.

[56] Y. Purwanti, B. H. Purwanto, and M. Jamaludin, "Citizen Participation in Electronic Public Administration: The Considerations of Functionality and the Technology Acceptance Model," *International Journal of Public Policy and Administration Research*, vol. 9, no. 4, pp. 90–101, 2022, doi: 10.18488/74.v9i4.3206.

[57] T. Trisninawati, "University Performance in the Era of Digital Transformation," *Journal La Sociale*, vol. 5, no. 4, pp. 1021–1029, 2024, doi: 10.37899/journal-la-sociale.v5i4.1236.

[58] A. Alomar, "COVID-19 and the Digital Transformation That Followed in the Kingdom of Saudi Arabia," *Academic Journal of*

Research and Scientific Publishing, vol. 5, no. 53, pp. 05–17, 2023, doi: 10.52132/ajrsp.e.2023.53.1.

[59] H. A. Elhawa, "Coronateaching in a Palestinian University: Changes, Challenges, and New Conceptions," 2021, doi: 10.2991/assehr.k.211224.003.

[60] R. A. Al-Samiri, "English Language Teaching in KSA in Response to the COVID-19 Pandemic: Challenges and Positive Outcomes," *Arab World English Journal*, no. 1, pp. 147–159, 2021, doi: 10.24093/awej/covid.11.

[61] M. Hassounah, H. Raheel, and M. Alhefzi, "Digital Response During the COVID-19 Pandemic in Saudi Arabia," *J Med Internet Res*, vol. 22, no. 9, p. e19338, 2020, doi: 10.2196/19338.

[62] H. Brdesee, "A Divergent View of the Impact of Digital Transformation on Academic Organizational and Spending Efficiency: A Review and Analytical Study on a University E-Service," *Sustainability*, vol. 13, no. 13, p. 7048, 2021, doi: 10.3390/su13137048.

[63] M. A. Salem, W. H. Alsyed, and I. A. Elshaer, "Before and Amid COVID-19 Pandemic, Self-Perception of Digital Skills in Saudi Arabia Higher Education: A Longitudinal Study," *Int J Environ Res Public Health*, vol. 19, no. 16, p. 9886, 2022, doi: 10.3390/ijerph19169886.

[64] K. Sendi and M. H. Alhumsi, "Teaching Research Methodology to Undergraduate Students Using Collaborative Learning Approach in a Blended Learning Environment at Saudi Electronic University," *World Journal of English Language*, vol. 13, no. 7, p. 171, 2023, doi: 10.5430/wjel.v13n7p171.

[65] Y. H. Al-Mamary and K. K. Al-Shammari, "Determining Factors That Can Influence the Understanding and Acceptance of Advanced Technologies in Universities' Teaching and Learning," *International Journal of Advanced and Applied Sciences*, vol. 10, no. 3, pp. 87–95, 2023, doi: 10.21833/ijaas.2023.03.012.

[66] ODP, "Total administrative and technical staff by educational institution and academic qualification from 2013 to 2020," Open Date Platform . Accessed: Jul. 07, 2025. [Online]. Available: https://open.data.gov.sa/en/datasets/view/4eca942f-9719-4f46-a7a6-59312136e7d5/resources

[67] ODP, "Data of faculty members in university education," Open Date Platform . Accessed: Jul. 07, 2025. [Online]. Available: https://open.data.gov.sa/en/datasets/view/269d0068-0f67-4483-b08b-efc21f9d7900/resources

[68] M. C. Parent, "Handling item-level missing data: Simpler is just as good," *Couns Psychol*, vol. 41, no. 4, pp. 568–600, 2013.

[69] Y. Lebeau and J. and Alruwaili, "Convergence and local orders in the dynamics of change in higher education: a perspective from Saudi Arabia," *Policy Reviews in Higher Education*, vol. 6, no. 1, pp. 6–26, Jan. 2022, doi: 10.1080/23322969.2021.1904791.

[70] P. C. Verhoef *et al.*, "Digital transformation: A multidisciplinary reflection and research agenda," *J Bus Res*, vol. 122, pp. 889–901, 2021.

[71] F. Abolhassan, *The drivers of digital transformation Why There's No Way Around the Cloud.* Springer, 2017.

[72] J. C. Bertot, E. Estevez, and T. Janowski, "Digital public service innovation: Framework proposal," in *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, 2016, pp. 113–122.

[73] N. Kshetri, "1 Blockchain's roles in meeting key supply chain management objectives," *Int J Inf Manage*, vol. 39, pp. 80–89, 2018.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 99–109
DOI: 10.15439/2025F9424
ISSN 2300-5963 ACSIS, Vol. 43

CADM: An LSTM-Based Model for Detecting Creative Accounting in Time-Series Data from Saudi-Listed Companies

Maysoon Bineid
0009-0007-4447-3706
Department of IS and Technology University of Jeddah, SA Department of Informatics
University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: M.bineid@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: N.Beloff@sussex.ac.uk

Anastasia Khanina
0009-0007-7496-0054
School of Business and Law, University of Brighton
Brighton, BN2 4NU, UK
Email: A.Khanina@brighton.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: M.White@sussex.ac.uk

***Abstract*—Studies on Saudi accounting practices have identified evidence of creative accounting in the financial statements of listed companies. Despite the application of various fraud detection methods, identifying legal but misleading manipulations remains challenging. This paper extends the Creative Accounting Detection Model (CADM), an LSTM-based model originally proposed by Bineid et al. (2023, 2024) for detecting creative accounting. Two versions, (CADM1) and (CADM2), were trained on two simulated datasets with different bases, achieving 100% and 95% accuracy, respectively. Testing on the energy sector (2019-2023), CADM1 identified one company as engaging in creative accounting, while CADM2 classified all companies as non-creative with greater confidence stability. The findings establish CADM as a robust, scalable solution for the early detection of financial manipulation. By combining predictive strength with explainability, CADM can be employed to advance current approaches to forensic accounting and risk analytics, offering valuable insights to regulators, auditors, and decision-makers.**

***Index Terms*—Deep Learning, Long Short-Term Memory, LSTM, Creative Accounting, Saudi Arabia.**

I. Introduction

DESPITE the establishment of accurate accounting standards, financial statement scandals continue to shake global markets. Cases such as Enron [1] and Wirecard [2] reveal the persistent challenge of Creative Accounting (CA), a practice where managers manipulate the accounting figures to present a misleading image of financial position.

Business performance is typically evaluated through accounting outcomes presented in financial statements (FSs). To ensure these statements offer a true and fair view, companies are required to prepare them under well-defined accounting standards and to subject them to independent auditing [3]. These measures are intended to ensure transparency, confirm compliance, and detect any material misstatement. However, corporate managers often have incentives to present the most favourable image of their organisation [4]. In doing so, they may engage in CA practices, adjusting figures within the boundaries of acceptable standards, without technically violating the rules. Although CA is sometimes distinguished from fraud, its consequences can be just as damaging [5], [6].

While many studies have focused on detecting Financial Statement Fraud (FSF), fewer studies have tried to identify CA (also referred to as Earnings Management EM) [7], [8]. One of the earliest and most notable efforts is the Beneish M-Score [9], a model built on financial ratios designed to flag potential EM. However, its capability in a non-U.S. context remains questionable [10]. Other statistical models like the Jones model and the modified Jones model [11] have been developed to detect EM, but they still face limitations such as model misspecification and low detection power, particularly in varying economic contexts.

More recently, Deep Learning (DL) applications have emerged as a powerful tool in financial prediction and classification. DL models, particularly neural networks, have demonstrated promise in learning complex patterns embedded in large unstructured datasets. Approaches combining Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Network (GRN), and even Bidirectional Encoder Representations from Transformers (BERT) models have been explored to solve problems like gradient vanishing and square modelling [12], [13]. For example, Craja et al. [14] applied deep learning techniques to detect FSF based on textual analysis of financial reports using a Bag-of-Words (BOW) approach. The model successfully identified fraudulent cases; however, it was designed for linguistic analysis only. Schreyer et al. [15] used Adversarial Autoencoder Neural Networks to detect anomalies in raw accounting entries, although their focus on ERP data structures limits the ability to generalise to standardised FSs.

Given the sequential nature of FSs, LSTM networks have shown remarkable suitability for analysing time-series data. They are particularly adept at capturing long-term dependencies and detecting subtle deviations from normal patterns. Their success in areas like anomaly detection, image recognition, and natural language processing has encouraged their application to financial domains, including fraud detection [16], [17] and performance predictions [18]. However, applying DL techniques to CA detection requires

Topical area: Advanced Artificial Intelligence in Applications

overcoming a fundamental obstacle: the lack of real-world CA examples. However, in the Saudi Arabian context, studies have found that FSs do not represent a company's true and fair position, despite auditor procedures approval[4], [19], [20], [21]. This alarming finding emphasises the need for innovative solutions to detect FS manipulation and improve financial reporting quality.

Despite the growing literature on detecting financial statement fraud, a significant gap remains in detecting CA weather through statistical methods or Artificial Intelligence, particularly within the Saudi Arabian context. In addition, the absence of labelled datasets for CA, particularly in the Saudi market, presents a methodological challenge in developing, training, and validating such models. To address this gap, this research aims to design, simulate, and evaluate a Creative Accounting Detection model (CADM), using an LSTM architecture, capable of learning from domain-specific financial and non-financial patterns.

To achieve this aim, the study makes several contributions to the literature. First, it introduces a novel domain-specific simulation methodology tailored to the characteristics of Saudi-listed companies. Second, it compares two distinct simulation approaches to determine which offers more effective training for the CA detection model. Third, it validates the proposed models using real-world FS data from the Saudi energy sector. These contributions are guided by the following research questions:

- **RQ1:** How does the base used in simulated data affect the accuracy of LSTM models in detecting creative accounting?
- **RQ2:** Can the proposed deep learning model (CADM) be trained to classify with high accuracy?
- **RQ3:** Can the proposed deep learning model (CADM) be generalised to real-world FSs?

II. Methodology

A mixed-methods approach is employed to deliver the research objectives. It combines an experimental methodology with quantitative techniques to form a framework for developing, training, and evaluating a deep-learning model using simulated and real-world datasets. The upcoming subsections explain the research design, feature engineering, data preparation, model architecture, training process, and testing procedures.

A. Research Design

The research design comprises two overlapping phases: quantitative and experimental, implemented in parallel across different stages of this study. The first phase involves preparing the dataset, which consists of two main components: a training dataset and a testing dataset. The testing dataset is collected from real-world FSs of Saudi-listed companies. However, due to the unavailability of a CA-labelled dataset, this study adopts a simulation-based methodology to generate FSs that reflect selected CA patterns. This leads us to start the second phase before ending the first, as shown in Fig. 1.

This part of the experimental phase involves implementing a domain-specific simulation process to generate synthetic datasets. However, two simulation designs were applied simultaneously. The first simulation is based on the real-world dataset analysis, and the second is based on the findings of Leitch and Chen [21].

Once the simulation stage is completed, the quantitative phase resumes with the development and training of the model. Using an LSTM-based architecture, the quantitative approach is implemented by analysing numeric accounting data, financial ratios, Corporate Governance (CGs), and key performance indicators extracted from real and simulated FSs to capture and classify CA practices. The outcomes of this analysis are further evaluated using deep learning performance metrics in the remainder of the second phase.

B. Data Sources and Collection

The emergence of publicly accessible financial data in Saudi Arabia is a relatively recent advancement that reflects the Kingdom's efforts to transform its financial system, in-

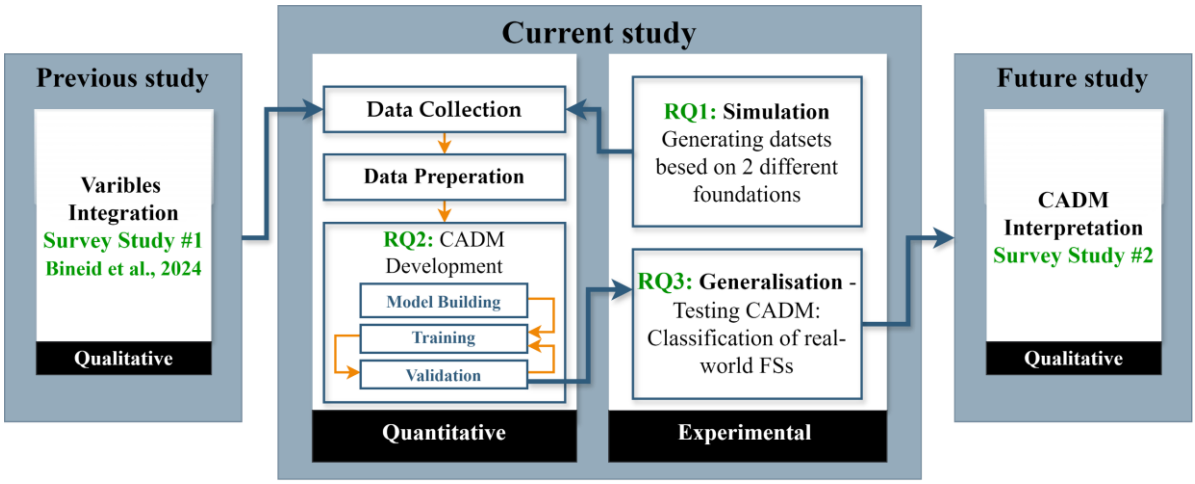


Fig. 1: CADM Research Framework

crease transparency, and attract domestic and foreign investment. These efforts are rooted in institutional reforms that began in the early 2000s, particularly with the creation of the CMA and have accelerated under the Vision 2030 economic transformation plan[22].

Saudi Arabia's enforcement of financial disclosure standards and IFRS compliance since 2017 has significantly improved the reliability and usability of its financial datasets. Real-world financial data can be accessed from the Saudi Exchange online platform (Tadawul), which provides official financial disclosures, including downloadable PDFs and Excel data. CMA and SOCPA platforms also offer regulatory filings, enforcement actions, historical announcements, standards, conversion guidelines, and auditor regulations. Third-party aggregators, such as Argaam.com and Mubasher.com, offer processed versions of the raw data; nevertheless, these platforms are commercially operated entities, not government-affiliated sources.

The Saudi market has 22 different sectors, each representing a distinct market segment with a varying number of companies in each sector. The largest sector by market capitalisation is the Energy sector, and the smallest is the Entertainment which has a limited number of companies with a relatively lower market capitalisation compared to others. Sectors and the exact number of companies can vary due to new listings and reclassifications in the market. However, FSs of Saudi-listed companies are publicly available in several formats. Tadawul provides financial data for all stakeholders, available in several formats, either online or offline. They also provide

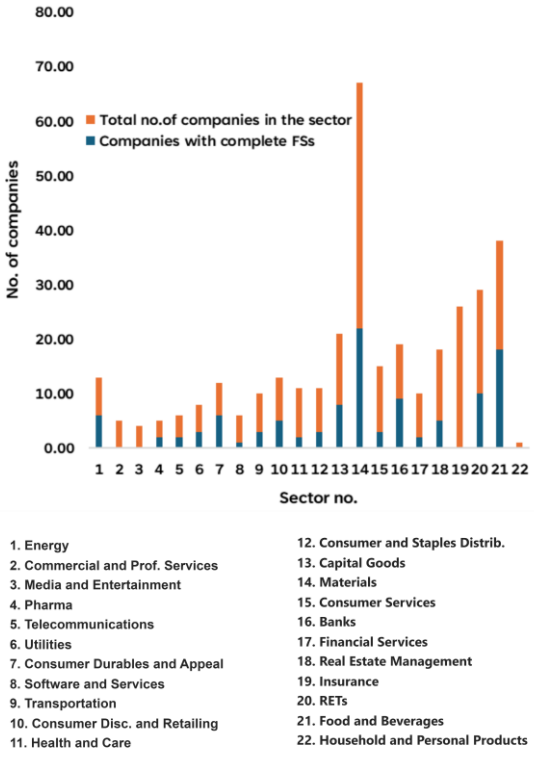


Fig. 2: Availability of FSs in the Saudi Market

a useful analytical tool for paid subscription users. But still, some companies fail to publish regularly in Tadawul, as shown in Fig. 2.

The Energy sector, having 7 companies, was chosen for this study. Although the Materials sector has the highest number of companies and is academically recognised for its high consistency across firms, making data preparation relatively easier, there was a concern that such uniformity might restrict the model's ability to learn diverse patterns and ultimately reduce training effectiveness. To better challenge the model and assess its generalisation capabilities, the Energy sector was selected due to the greater diversity among its companies.

C. *Simulation Strategy*

Although there is an increase in data availability, real-world examples of CA are challenging to find, particularly in the Saudi context [7], as companies do not usually admit publicly to using CA techniques. Without a real dataset, the best alternative would be to employ proxies, such as accrual-based analysis [26], on publicly available financial datasets to infer potential earnings management. However, accrual-based models like the Beneish M-Score [9] rely on assumptions and may not provide a definitive ground truth for CA, as they may include false positives and negatives, admitted by Beneish himself [10].

Moreover, AI strategies that can effectively address the lack of data in training deep learning models, such as Semi-supervised Learning, cannot be used because they lack the level of customisation needed in the dataset, and their results will be difficult to generalise [23]. It can also be helpful to use pre-trained models on a broader dataset (e.g., fraud detection models) as in Transfer Learning. Still, they are more likely to perform poorly on the domain-specific (Saudi-listed companies), and controlled variables (selected CA patterns) used in this research.

Consequently, simulating financial data appeared to be the optimal strategy that fulfils the need to have a sector-specific dataset with labelled data and controlled variables. Simulating financial statements is a well-established practice in the literature, employed for various analytical purposes. For instance, Leitch and Chen [24] simulated monthly FSs to evaluate key financial indicators and explore broader organisational dynamics.

Building upon their methodology, this study simulates two datasets: one based on general accounting measures and another tailored to the characteristics of the Saudi-listed companies, including financial (FIN) and non-financial (N-FIN) variables. Each dataset is prepared to train different versions of the CADM model. Generating a simulated dataset can explicitly encode CA techniques relevant to the Saudi business environment, providing flexibility for training scenarios and control over manipulation patterns, specifically when the base of the simulation is the real-world dataset. In other words, collecting the real-world dataset is essential to start the simulation process.

A structured pattern-based financial simulation approach has been utilised to generate the training dataset. First, two

primary goals were established: (1) to simulate a dataset closely resembling real-world data to ensure the model is effectively trained and (2) to account for four different scenarios of account manipulations, given the lack of clarity about the exact manipulation scenarios present in real datasets, if any. The dataset consists of two groups of companies: labelled as CA and labelled as N-CA (Non-Creative Accounting). The N-CA group was generated without applying any manipulations, while the CA group underwent an additional process where CA patterns were embedded.

To achieve the first goal, FIN features such as raw accounting data and financial ratios, and NFIN features such as CG metrics and auditor's status, were incorporated into the simulation process. To address the second objective of this study, two datasets were simulated using two different bases: one based on general accounting measures reported in the literature and the other grounded in real-world data analysis. The first simulated dataset (DST_DSL) was generated by adopting the simulation methodology proposed by Leitch and Chen [24]. For the second simulated dataset (DST_DSR), initial statistical analyses of the real-world dataset were conducted and used as the baseline for determining measurements and ratios. This dual approach enabled the training of two separate models on distinct simulated datasets, thereby enhancing the robustness and adaptability of the proposed detection framework. A summary of the datasets used in the processes is provided in Table I.

Both simulated datasets are validated before being used in training; they are statistically analysed to compare key financial metrics with the real-world dataset, ensuring that trends, outliers, and energy sector characteristics are aligned. This was performed iteratively; when the validation results were not acceptable, the simulation was redesigned, and the generation phase was restarted.

TABLE I: DATASET NAMING AND SPECIFICATIONS USING MATLAB

Dataset	Dataset type	Base	Purpose
DST_DSL.mat	Simulated	[20]	Training CADM1
DSV_DSL.mat	Simulated	[20]	Validation CADM1
DST_DSR.mat	Simulated	DSR.mat	Training CADM2
DSV_DSR.mat	Simulated	DSR.mat	Validation CADM2
DSR.mat	Real-world	N/A	Testing CADM1&2

D. Feature Selection and Engineering

This study aims to simulate a set of widely recognised and frequently cited CA techniques, as in [25], [26], within a controlled dataset, representing them as manipulation patterns for CADM to learn from and detect. Four key CA patterns were selected for inclusion in the simulated dataset: revenue inflation, underreporting of COGS, assets overstatement, and underreporting of liabilities, as detailed in Table II. To make sure these patterns are relevant to the Saudi context, an additional layer of selection criteria was applied. Specifically, these techniques were also chosen based on their prevalence in the local literature [4], [19], [27], their documentation in real-world regulatory reports [28], and their alignment with the findings of the integration study [7].

To simulate a realistic and varied dataset, not all four CA techniques were applied to all companies in the CA label. Instead, each manipulated company was randomly assigned one of the four techniques, which was then consistently applied across all five-year financial periods. This approach avoids compound distributions that may arise from stacking multiple manipulation techniques and allows the model to learn distinct behavioural patterns associated with each manipulation type. It also better reflects real-world conditions, where companies typically engage in certain earnings management strategies over others based on their internal practices, industry norms, and regulatory pressure.

While other CA techniques have been identified in the literature, the selected CA patterns affect quantifiable financial ratios that make them suitable for detection by deep learning models based on time-series patterns. They are also compatible with the simulation process, as each technique can be implemented in a consistent and replicable way across multiple years. In addition, these patterns evolve, which is a key feature that aligns well with the capabilities of LSTM-based architectures. Finally, it is important to limit the number of manipulation types to control the model's complexity and reduce the risk of introducing noise or overlapping effects that could compromise the model's interpretability and reliability

TABLE II: SIMULATED TECHNIQUES OF CA

CA technique	**Behaviour in FSs**	**Simulation**
Revenue inflation	Revenue overstatement is one of the commonly used earnings management strategies and directly impacts key ratios such as net profit margin, EPS, and ROA [1], [2], [3].	A 15% inflation is applied to the revenue variable for each year for the generated companies.
Underreporting of COGS	This manipulation affects profitability without inflating top-line revenue and is less obvious than revenue inflation, thus posing a greater challenge for detection models[3], [4].	Reported COGS are decreased 10% (manipulated by 0.9) for each year while keeping the revenue constant.
Assets overstatement	This technique can distort key financial ratios such as ROA, current ratio, and debt-to-equity ratio. This tactic is commonly observed in cases where firms attempt to hide their losses or signal strength to creditors and investors [5].	Asset value is increased by 10% for each year across all years.
Liabilities underreporting	This practice is used to present a healthier financial position. It improves leverage ratios (debt-to-equity) and conceals financial risk. This pattern is particularly relevant in the credit evaluation context [3], [6].	Liabilities were reduced by 10%for each year, thereby relatively overstating equity.

Both financial (FIN) and non-financial (NFIN) indicators associated with the selected patterns were incorporated into the model. These indicators are typically variables integrated from financial reports and FSs. The selection of FIN variables was based on their relationship with the accounting pattern under examination. In contrast, non-financial variables were incorporated based on their demonstrated significance in identifying accounting manipulation, as evidenced by the findings of the integration study [7]. The model included 14 FIN (FIN1-FIN14) and 16 NFIN variables (NFIN1-NFIN20) as in Table III.

E. Model Architecture

The next step after data preparation and the identification of CA patterns is to select an appropriate deep learning technique for developing the predictive model CADM. As previously discussed, LSTM neural networks are widely regarded as one of the most effective methods for detecting unexpected anomalies in sequential unstructured data spanning multiple years. As an advanced type of RNN, LSTM is specifically designed to learn long-term dependencies in time series through a gated architecture. These gates, responsible for retaining, discarding, or updating information, enable the model to capture complex temporal relationships in historical financial statements.

Unlike traditional models, LSTM is robust against the need for extensive manual feature engineering, as it can automatically learn features from raw data. This allows for the inclusion of diverse FIN and N-FIN variables. Moreover, while CADM is designed to be trained using FIN and Non-FIN data, LSTM excels at modelling this non-linearity in data, specifically when it is required to capture temporal and contextual variables such as CGS measures, regulations, and multi-year data. This makes it possible to build a powerful deep learning model tailored to the Saudi business context. To effectively leverage these advantages, the CADM was implemented using multiple layers, where each layer contributes to progressively learning higher-level temporal features from the FIN and N-FIN input data. All LSTM-related mathematical formulations used in this study are adapted from Hochreiter and Schmidhuber [29], unless otherwise noted. The layered architecture (illustrated in Table III) is composed as follows:

Input Layer

The input layer receives sequential financial data spanning multiple years, structured as a time-series input as in Fig. 3. Each input sequence represents a company's FIN and N-FIN variables over consecutive reporting periods, allowing the

TABLE III: FIN AND N-FIN VARIABLES

FIN			NFIN			
Variable		Code	Variable		Code	Encoding
Accounting raw data	Bank balance and cash	FIN1	Basic	Year	NFIN1	2019-2023
	Inventory	FIN2		Company name	NFIN2	Table 1 Appendix
	Total assets	FIN3		Auditor Firm	NFIN3	Table 2 Appendix
	Total liabilities	FIN4		Audit opinion[1]	NFIN4	0 clean (unqualified), 1 otherwise
	Total equity	FIN5	CG metrics	Newcomer	NFIN6	1 yes, 0 otherwise
	Revenue	FIN6		Board Size	NFIN7	Count
	COGS	FIN7		CEO Duality[2]	NFIN8	0 no, 1 yes
	Gross profit	FIN8		Board Dependency[3]	NFIN9	0 independent,1 dependent
Financial ratios	Liquidity	FIN9		Board meetings	NFIN10	Count
	Profitability	FIN10		Audit committee size	NFIN11	Count
	Efficiency	FIN11		Ownership concentration	NFIN12	0 compliant, 1 non
	Leverage	FIN12		CEO tenure	NFIN13	0 compliant, 1 non
	Market ratio	FIN13		Adopted standards	NFIN14	0 SOCPA, 1 IFRS
			Auditor	Auditor status[4]	NFIN15	0 currently unauthorised, 1 authorised
				Audit Big 4[5]	NFIN16	0 no, 1 yes
				Auditor allegations[6]	NFIN17	0 no, 1 yes

1. There are four types of **auditor opinions**. The most common opinion is the Unqualified (Clean) Opinion, where the auditor believes the FSs are accurate and comply with accounting standards. This value is represented by 0. The second opinion is issued when the FSs are mostly accurate, but there is a specific issue that does not comply with standards, often explained in detail. This is called the Qualified Opinion. The remaining two types of opinions represent two levels of FS's misrepresentation [28]. To simplify the model's input, any opinion other than the clean opinion is represented as 1.
2. **CEO duality** indicates if the Chief Executive Officer holds the position of chairman of the board of directors. This feature is one of the main CG influencers, as holding both roles can lead to a concentration of power and potentially affect the board's ability to oversee management independently. This CEO duality takes 1 if the CEO is also the Chairman of the Board (indicating CEO duality exists) and 0 if the CEO and Chairman roles are held by separate individuals (no CEO duality).
3. According to the Saudi regulations by the CMA [13], CG regulations demand that at least one-third of the board members must be independent. Independent directors are non-executives. All independents are non-executives, but not all non-executives are independents (they might have other relationships). Therefore, the threshold in our design is 33%, so at least 33% must be independent. If less than 33%, then 1 (dependent). If 33% or more, then 0 (independent). **Board dependency** is measured using the following formula: *BD = (independent directors/ directors) *100*
4. Because the **status of the auditing firm** in the present indicates some flags about its integrity and professionalism, the performance history by allegations and by the status of the auditing firm by the authorities is highly effective in the analysis. If the audit firm (for a specific year) is authorised, this value takes 0, otherwise, the value is 1 if the audit firm is suspended in the current year.
5. The **size of the auditing firm** is considered as an important auditor's reputation feature as indicated by the findings of [2]. This variable asks whether the auditing firm is considered one of the Big 4 firms in Saudi Arabia (KPMG, PWC, EY, and Deloitte). It takes 1 if one of the big 4s, 0 otherwise.
6. It leverages the audit firm if it has no **previous allegations** or lawsuits either with companies or the authorities. The variable takes 1 if the auditing firm has no previous or current allegations, otherwise, it is 0.

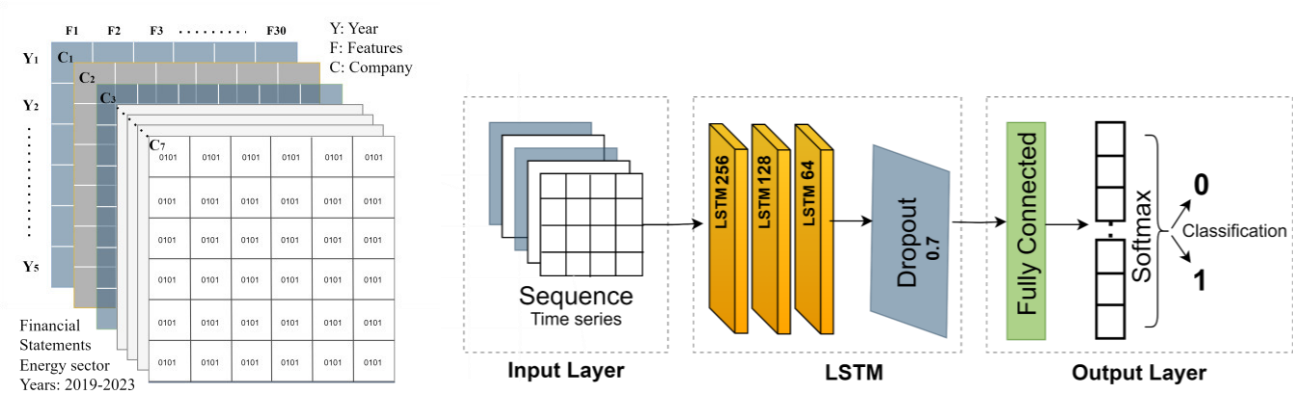


Fig. 3: Data Frame and Model Structure

model to capture temporal trends and patterns relevant to detecting CA. Time-series input can be mathematically expressed as follows:

$$X = \{x^1, x^2, x^3, \dots, x_T\}, x_t \in R^n \quad (1)$$

Where:
T number of time steps (years).
n number of features (financial + non-financial variables).
$\boldsymbol{x_t}$ feature vector at time step.

LSTM Layers

Multiple layers learn long-term dependencies between years of FSs, treating them as time series data. By leveraging memory cells and gated mechanisms, LSTM layers enable the model to identify changes in financial behaviour, such as trends in earnings manipulation or shifts in governance-related variables, that span across reporting periods. This is critical for CA detection, which may evolve gradually over time rather than appearing as sudden anomalies in a single year. The model consists of three LSTM layers with 256, 128, and 64 units, respectively. The first layer outputs the hidden states for all time steps to capture full temporal dependencies with 'sequence' output mode. The second layer with the 'last' output mode was used to summarise the input sequence into a single context vector, and the last layer with the same mode was used to further condense the sequential representation. The following equations describe the standard LSTM operation [29].

$$f_t = \sigma(Wfxt + Ufht - 1 + bf) \quad (2\text{-a})$$
$$i_t = \sigma(Wixt + Uiht - 1 + bi) \quad (2\text{-b})$$
$$o_t = \sigma(Woxt + Uoht - 1 + bo) \quad (2\text{-c})$$
$$\acute{c}_t = \tanh(Wcxt + Ucht - 1 + bc) \quad (2\text{-d})$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \acute{c}_t \quad (2\text{-e})$$
$$h_t = o_t \odot tanh(c_t) \quad (2\text{-f})$$

Where:
σ: sigmoid activation $\boldsymbol{h_t}$: hidden state (output)
tanh: hyperbolic tangent activation $\boldsymbol{c_t}$: cell state
⊙: element-wise multiplication
W,U,b : trainable parameters

Dropout Layer

Since the model is trained on simulated data that may not fully reflect the variability of real-world financial statements, a dropout layer with a rate of 0.7 is applied to prevent data overfitting by randomly deactivating 70% of neurons during training. This regularises the LSTM model and reduces the risk of memorising synthetic patterns that do not generalise well to actual data.

Fully Connected Layer

This layer, with 2 neurons, aggregates the learned temporal features extracted by the preceding LSTM layers and maps them to the target output classes. By combining information across all time steps and input variables, the fully connected layer enables the model to produce a binary prediction that reflects the likelihood of CA behaviour in a company's financial statements.

$$z = W_{fc} h_T + b_{fc} \quad (3)$$

Where:
$\mathbf{W_{fc}}$: weight matrix for the dense layer
$\mathbf{b_{fc}}$: bias vector

SoftMax Layer

Performs binary classification (two classes, CA and N-CA). It consists of a single neuron with a sigmoid activation function, which outputs a probability score between 0 and 1, indicating the model's confidence that the financial statement exhibits CA behaviour. The threshold applied to convert this probability into a final class label is 0.5.

$$\bar{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

Where:
$\bar{y} \in (\mathbf{0}, \mathbf{1})$ is the predicted probability of CA
A threshold $\boldsymbol{\theta = 0.5}$ is used to convert probabilities to class labels

$$\textbf{Predictions} = \begin{cases} 1 \; if \; \bar{y} > 0.5 & CA \\ 0 \; otherwise & N - CA \end{cases}$$

Training configuration

The model was trained using the Adam Optimiser for up to 100 epochs with a small mini-batch size of 10 to suit the limited number of training samples. Validation was performed every 5 epochs using a separate validation set, and early stopping was applied with a patience of 20 validation checks. Conversely, evaluation is performed once while training the model and another time during the testing phase, while comparing results. This phase is performed only once to assess the simulation efficiency.

F. Evaluation Metrics

An appropriate evaluation criterion must be established to assess the generalisation phase of both models and compare their performances. In the long-term plan of this research, the assessment criterion is to employ qualitative assessment, namely expert evaluation, to interpret testing results and validate the model's utility. This current stage relies on quantitative distributional analysis for the model outputs. Specifically, *SoftMax* confidence distributions are examined alongside the predicted labels to evaluate each model's behaviour and how biased it is towards each class. This approach allows for an initial assessment of how confidently the model classifies real-world FSs.

III. Implementation - The Experiment

This section details the implementation of CADM, focusing on the experimental process used to train and evaluate its performance. The implementation was carried out in the following phases: simulation (explained earlier), model training using the simulated datasets, model testing on real-world financial statements, and model evaluation as illustrated in Fig. 4 Training, Testing, and Evaluation phases are presented in the following subsections, outlining the experimental setup, hyperparameter choices, and evaluation outcomes.[22]

A. Training

As shown in Fig. 5, CADM1 achieved 100% accuracy in training and validation. In contrast, CADM2 achieved 92% accuracy in training and 80% accuracy in validation. All predicted labels were true by CADM1 and only 3 were false by CADM2 as shown in Fig. 6 These results were achieved after several iterations involving multiple rounds of dataset re-generation. The training datasets were carefully re-simulated to introduce variability and reduce the model's familiarity with recurring patterns, thereby promoting better generalisation and minimising overfitting.

To optimise model accuracy, various hyperparameters were fine-tuned, including the number of LSTM Units. They experimented with different neuron counts to determine the optimal balance between model complexity and learning capacity. While smaller configurations, such as 128, 64, and 32 units, were initially considered to reduce overfitting, the final architecture adopted a larger structure —256, 128, and 64 units —demonstrating improved learning performance without compromising generalisation. This architecture, applied to both CADM1 and CADM2, retained sufficient capacity to capture complex temporal dependencies in the data while maintaining robust performance on unseen samples, despite the relatively small dataset size.

Moreover, the Learning Rate was adjusted to prevent convergence issues, and the batch size and epochs were optimised for computational efficiency and model performance. Additionally, L2 regularisation was applied and a dropout layer to mitigate overfitting. Finally, the Optimiser Selection compared Adam, RMSprop, and SGD to identify the most effective optimisation strategy.

B. Testing

Following the training of both models, CADM1 and CADM2, the next step is to test them to assess their generalisation capabilities. Testing is designed to be implemented on

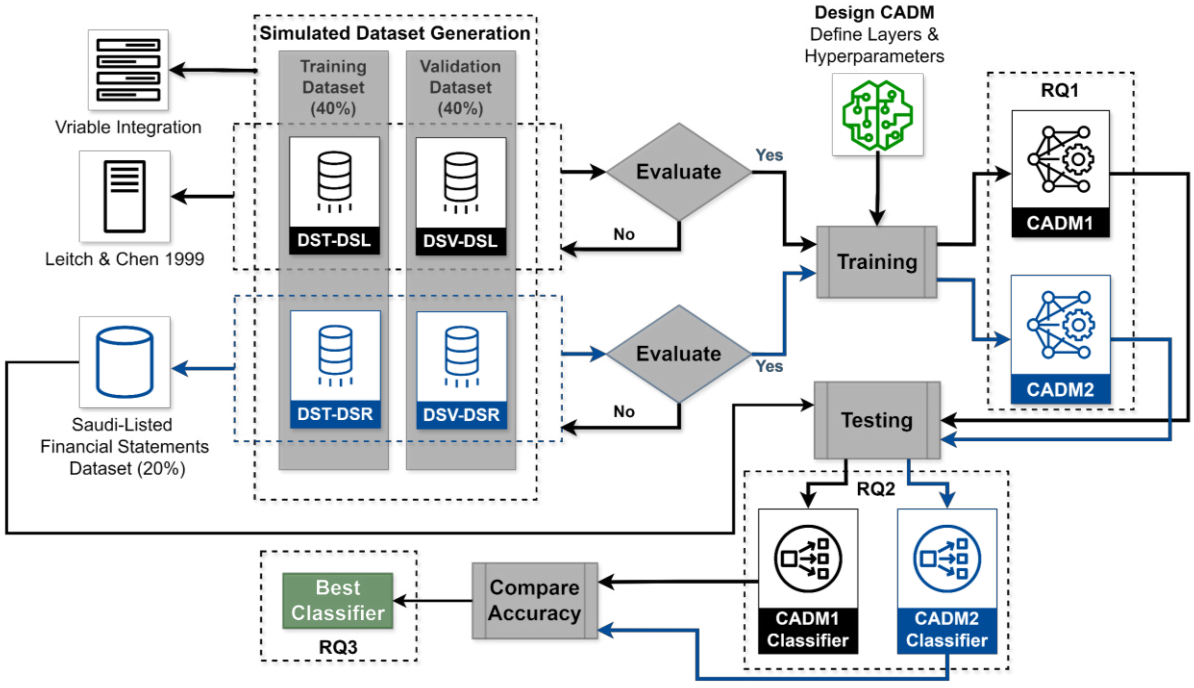


Fig. 4: Workflow of CADM Simulation, Training, Testing, and Evaluation

the real-world dataset; the collected and pre-prepared FSs, which have become DSR after preparation. Due to having two models, testing was done twice on the same dataset. However, DSR is prepared to be in the same shape as the simulated dataset to have correct and accurate testing and be ready to be used by CADM.

However, CADM1 predicted five out of six companies as class 1 (CA) Fig. 7, but with SoftMax probabilities clustered very close to the classification threshold (ranging between 0.5001 and 0.5042) Fig. 8. The single prediction of class 0 (N-CA) was also marginal, with a class 0 probability of just 0.5004. This narrow range of values indicates that CADM1 was uncertain in its predictions, which reflects the model's weak classification capability and lack of confidence when applied to real-world financial statements.

In contrast, CADM2 predicted all companies class 0 (N-CA) Fig. 7, with SoftMax probabilities ranging from 0.5021 to 0.5063. Being close to the decision threshold as well does not indicate any better performance than CADM1, although they have more consistency (all above 0.5). Still, CADM2 demonstrates modest and more stable confidence in labelling the data.

IV. Discussion

This study successfully developed and evaluated two LSTM-based models (CADM1 and CADM2) to classify FSs

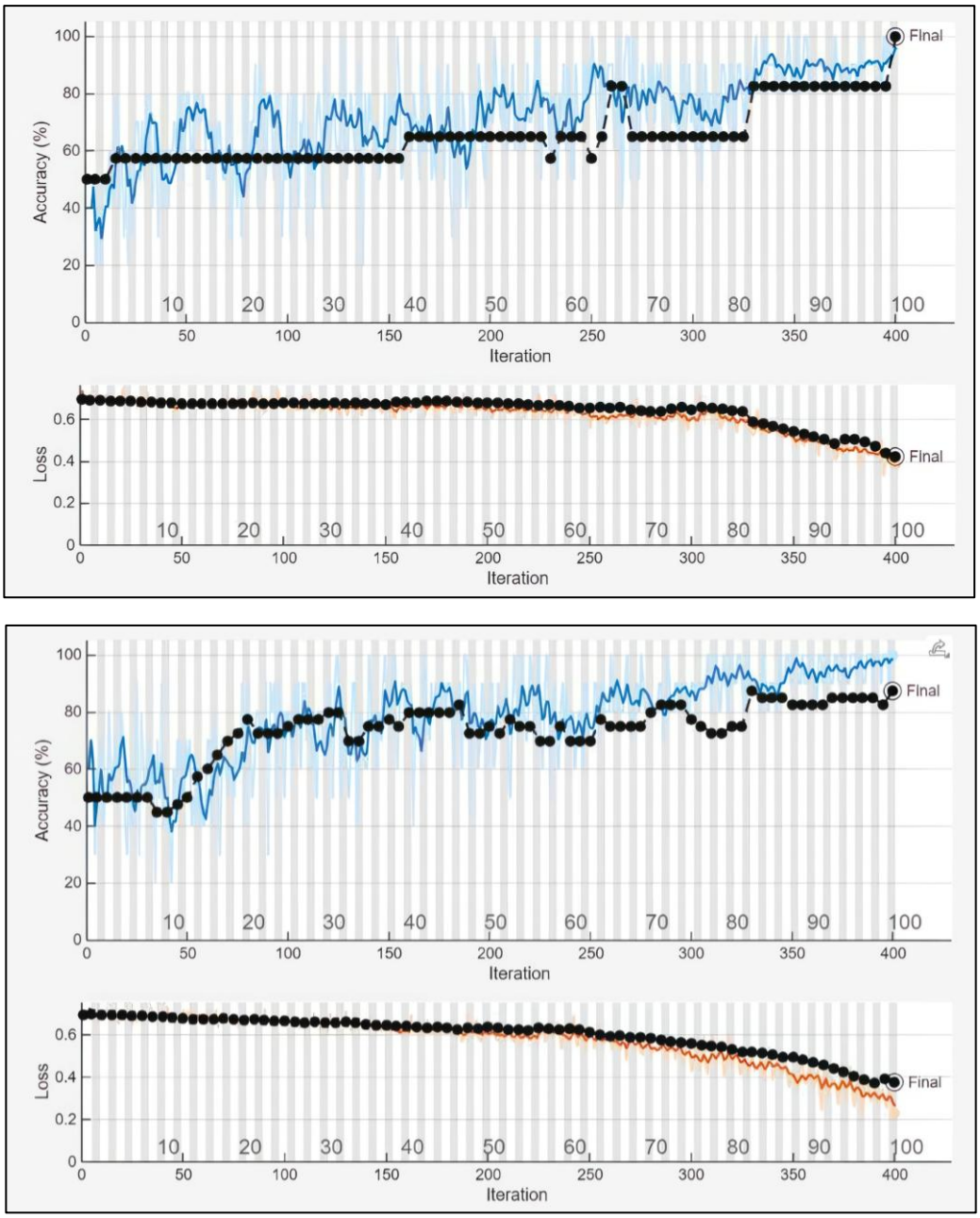


Fig. 5: Training accuracy and loss curves for CADM1 (top) and CADM2 (bottom), showing the models' learning performance over 100 epochs

based on the likelihood of CA practices being used in their preparation. The models were trained and validated to assess their effectiveness in controlled simulations and real-world Saudi market data. Additionally, the study compared the models' performance regarding the underlying simulation approach used to generate their training dataset. This section discusses the research objectives and demonstrates meaningful contributions to DL-driven financial analysis.

RQ1: Training CADM: Learning from Simulated Data

The results of training CADM reflect a key contrast between real-world context and theoretical modelling. CADM1 achieved perfect accuracy (100%) in both training and validation phases. This technical success confirms that the model architecture is responsive to clean data with well-defined patterns. However, this result may also indicate overfitting to the ideal dataset that lacks real-world variability.

In contrast, training CADM2 was more challenging. Despite extensive hyperparameter tuning, CADM2's accuracy was lower than CADM1's, reflecting the inherited complexity of real-world Saudi financial behaviour. This result is not a flaw but a strength: it underscores CADM2's exposure to more realistic noise, non-linear interactions, and imperfect relationships characterising actual FSs.

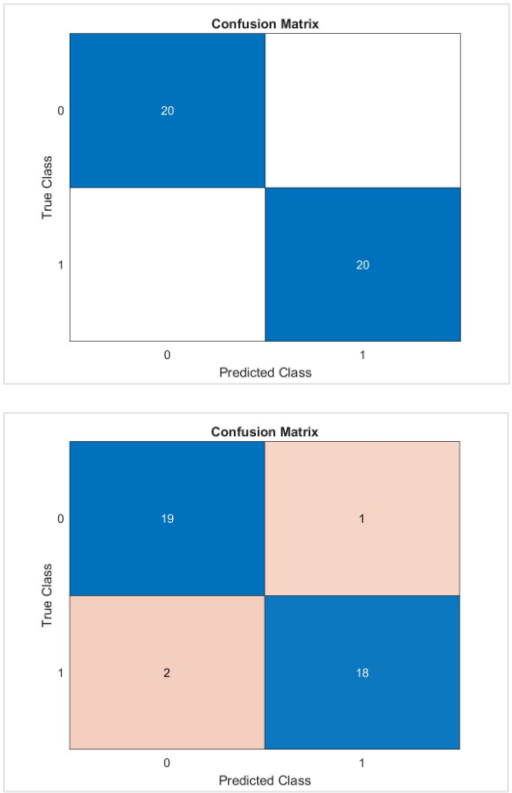


Fig. 6: Classification Results of CADM1 Training (top) and CADM2 Training (Bottom)

In addition, the training dynamics promote valuable reflections on the boundaries of CADM's optimal performance, raising questions about model robustness and the interpretability of results when using overly cleaned data, a significant debate in DL-based financial models. CADM1's perfect performance might not perfectly generalise to real-world data, while CADM2, despite lower learning scores, likely learned patterns closer to reality. This observation supports the argument in the literature about the risk of training financial AI models on idealised datasets.

RQ2: Generalisation: Performance on Real-World Data

As ground truth for CA is not available, innovative assessment methods become necessary. At this stage of the research, the generalisation capacity of both models was evaluated using class distribution patterns and feature-based differentiation. Although theoretically accurate, CADM1 was more liable to misclassification when confronted with real-world variance. In contrast, the final version of CADM2 testing showed that all companies in the Energy sector are CA-free. This can be interpreted as the model trained on a dataset that already sees these scenarios as non-CA, which reflects stronger generalisation capacity from CADM2.

From the perspective of CA theory, CA is intentionally designed to evade detection mechanisms. Consequently, a deep contextual understanding and the incorporation of non-financial signals are essential for effective detection. This may explain the comparatively better classification performance by CADM2 when tested on real-world FSs, as it appears to have

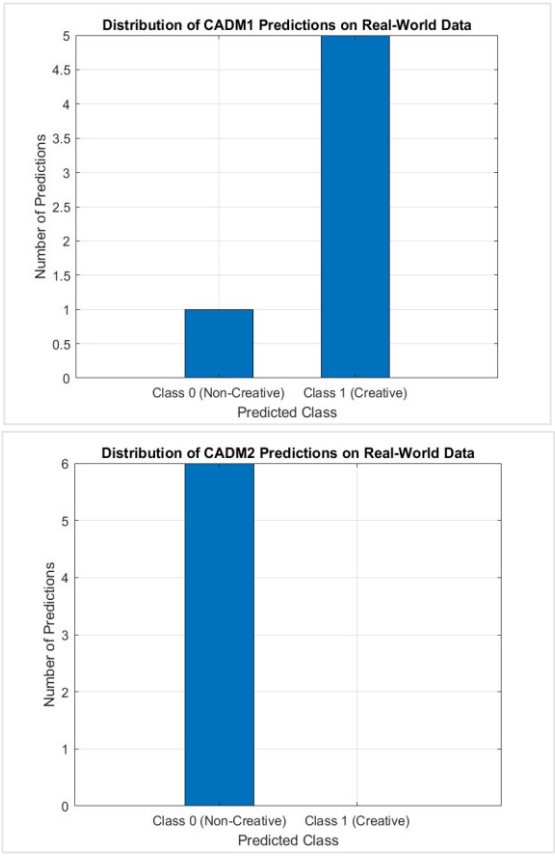


Fig. 7: Classification results of CADM1 (top) and CADM2 (bottom) on the real-world dataset, showing predicted counts for Class 0 (N-CA) and Class 1 (CA).

captured some attributes of the complexity inherent in real-world data. Nevertheless, generalisation was naturally challenged by sector-specific anomalies. Companies in this sector have diverse accounting scenarios and include a dominant outlier (ARAMCO), with significantly different scales and metrics compared to its peers. These factors influenced the model's sensitivity and underscore the importance of sector-aware tuning in future iterations.

RQ3: Evaluating The Simulation Base: Theory vs Reality

The comparison between CADM1 and CADM2 provides critical insight into how simulation strategy affects model behaviour. Evaluating the simulation base was done by comparing the results of CADM1 and CADM2 classifiers. The SoftMax probability distributions further illustrate these tendencies. While CADM1 demonstrated theoretical strength, CADM2 has more consistent confidence values, particularly when they were consistently above 0.5 and demonstrated more stable confidence in labelling companies as N-CA. The more consistent confidence profile of CADM2 is attributed to its training on a dataset simulated using real-world statistical analysis, which enabled it to better mirror the characteristics of actual financial reporting in the Saudi market. This confirms the validity of using real-world statistical patterns in training simulations.

The contrasted results between CADM1 and CADM2 illustrate that models trained on idealised literature-driven simulations achieve higher accuracy yet lower generalisation ability to real-world FSs compared to models trained on real-world-based simulations. This supports the theoretical argument in the literature that training models on clean data may not adequately prepare them for the complexities and imperfections found in real-world data [30], [31]. In addition, the experiment in this study suggests that hybrid approaches, such as blending literature-based metrics with real-world features, may offer the most promising path forward.

V. Conclusion

This study represents a significant milestone in developing a data-driven framework for detecting CA within Saudi-listed companies. By training two variants of the Creative Accounting Detection Model (CADM) using different simulation criteria, this research establishes a credible foundation for AI-enabled financial monitoring. The model was then tested on the real-world dataset collected from the Saudi market portal to check if the model is capable of looking at how financial and non-financial metrics evolve, CADM and classifying Saudi-listed companies in the Energy sector as creative accounting or non-creative accounting practitioners.

Both SoftMax confidence distributions of CADM predictions on the real-world dataset models offer a significant insight: the importance of strong theory and the need to reflect real-world scenarios when detecting financial anomalies. The implication of this study extends beyond model performance. It sets a new direction for applying deep learning to complicated grey areas like creative accounting. Unlike fraud detection, this model identifies forms of financial misrepresentation that fall within regulatory bounds but still distort finan-

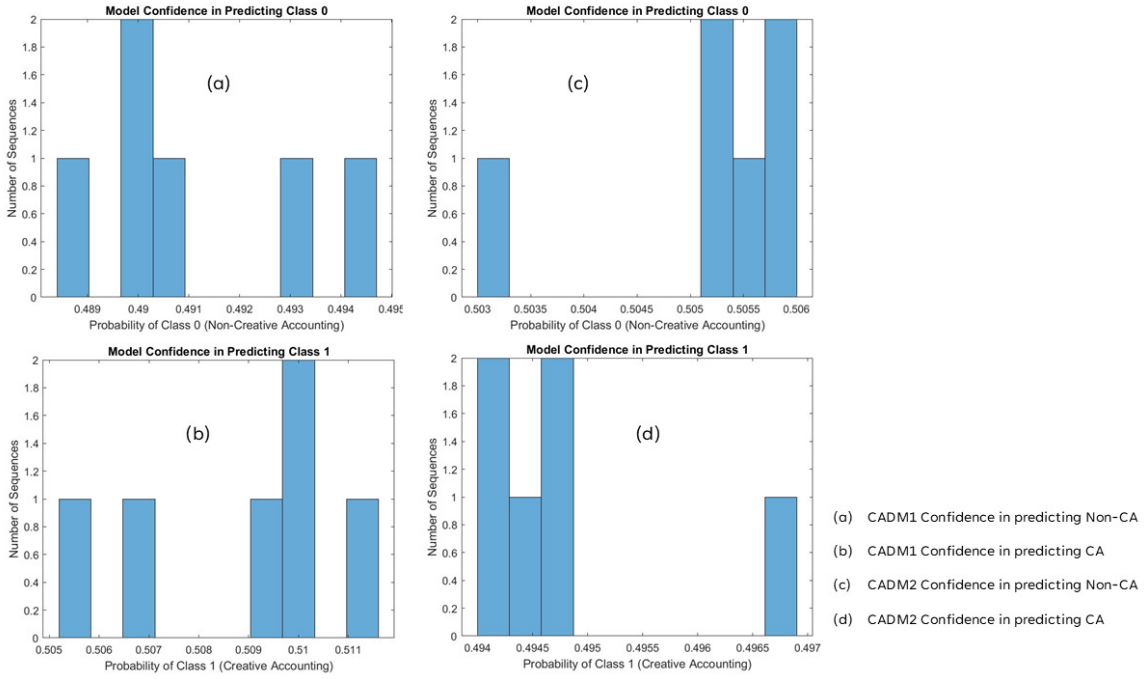


Fig. 8: SoftMax confidence distributions of CADM predictions on the real-world dataset

cial reality. Further research may involve integrating additional financial indicators such as the revenue recognition, misclassification of accounting elements, and off-balance sheet financing, and governance indicators, such as management compensations and narrative disclosure. It could also add explainability mechanisms to the model, such as applying custom attention layer to the model to enhance model transparency and make CADM a valuable tool for regulators and auditors. In addition, the simulation scope could be expanded to have semi-creative accounting data, and the testing could include other sectors to improve sensitivity and real-world relevance. Future work is intended to integrate expert interpretation of classification results through interviews to validate the model's output and illuminate practical constraints.

This research serves as a foundation for the next generation of intelligent accounting analytics. Besides considered a proof of the feasibility of AI in financial transparency, it promotes meaningful integration of simulation, sector knowledge, and regulatory insight.

References

[1] L. Revsine, 'Enron: Sad but inevitable', *J. Account. Public Policy*, vol. 21, no. 2, pp. 137–145, 2002, doi: 10.1016/S0278-4254(02)00044-3.

[2] Hoje Jo, Annie Hsu, Rosamaria Llanos-Popolizio, and Jorge Vergara-Vega, 'Corporate Governance and Financial Fraud of Wirecard', *Eur. J. Bus. Manag. Res.*, Mar. 2021, doi: 10.24018/ejbmr.2021.6.2.708.

[3] O. Mesioye and I. Bakare, 'Evaluating Financial Reporting Quality: Metrics, Challenges, and Impact on Decision-Making', *Int. J. Res. Publ. Rev.*, vol. 5, pp. 1144–1156, Oct. 2024, doi: 10.55248/gengpi.5.1024.2735.

[4] M. Bineid and A. Asiri, 'Creative Accounting Incentives and Techniques in Saudi Public Companies', *KAU J. Adm. Econ.*, vol. 2, no. 27, 2013, doi: 10.4197/Eco.

[5] Jones Michael, *Creative accounting, fraud, and international accounting scandals*. Chichester, West Sussex, England; John Wiley and Sons, 2011.

[6] E. E. Akpanuko and N. J. Umoren, 'The influence of creative accounting on the credibility of accounting reports', *J. Financ. Report. Account.*, vol. 16, no. 2, pp. 292–310, 2018, doi: 10.1108/JFRA-08-2016-0064.

[7] M. Bineid, A. Khanina, N. Beloff, and M. White, *Integrating Non-financial Data into a Creative Accounting Detection Model: A Study in the Saudi Arabian Context*, vol. 504. in Lecture Notes in Business Information Processing, vol. 504. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-61657-0.

[8] M. Bineid, N. Beloff, M. White, and A. Khanina, 'CADM: Big Data to Limit Creative Accounting in Saudi-Listed Companies', presented at the 18th Conference on Computer Science and Intelligence Systems, Sep. 2023, pp. 103–110. doi: 10.15439/2023F3888.

[9] M. D. Beneish, 'The Detection of Earnings Manipulation', *Financ. Anal. J.*, vol. 55, no. 5, pp. 24–36, 1999, doi: 10.2469/faj.v55.n5.2296.

[10] M. D. Beneish and P. Vorst, 'The Cost of Fraud Prediction Errors', *Account. Rev.*, vol. 97, no. 6, pp. 91–121, Oct. 2022, doi: 10.2308/TAR-2020-0068.

[11] P. M. Dechow, A. P. Hutton, J. H. Kim, and R. G. Sloan, 'Detecting Earnings Management: A New Approach', *J. Account. Res.*, vol. 50, no. 2, pp. 275–334, 2012, doi: 10.1111/j.1475-679X.2012.00449.x.

[12] J. Wang, S. Hong, Y. Dong, Z. Li, and J. Hu, 'Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting', *J. Comput. Sci. Softw. Appl.*, vol. 4, no. 3, pp. 1–7, 2024.

[13] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, 'METAFRAUD: A META-LEARNING FRAMEWORK FOR DETECTING FINANCIAL FRAUD', *MIS Q.*, vol. 36, no. 4, pp. 1293–1327, 2012, [Online]. Available: http://www.misq.org

[14] P. Craja, A. Kim, and S. Lessmann, 'Deep learning for detecting financial statement fraud', *Decis. Support Syst.*, vol. 139, Dec. 2020, doi: 10.1016/j.dss.2020.113421.

[15] M. Schreyer, T. Sattarov, C. Schulze, B. Reimer, and D. Borth, 'Detection of Accounting Anomalies in the Latent Space using Adversarial Autoencoder Neural Networks', Aug. 02, 2019, *arXiv*: arXiv:1908.00734. doi: 10.48550/arXiv.1908.00734.

[16] F. Alam and S. Ahmad, 'Intelligent Fraud Detection Framework for PFMS Using HGRO Feature Selection and OC-LSTM Fraud Detection Technique', *SN Comput. Sci.*, vol. 4, no. 4, p. 400, May 2023, doi: 10.1007/s42979-023-01855-5.

[17] Yanash Azwin Mohmad, 'Credit Card Fraud Detection Using LSTM Algorithm', vol. 1, no. 3, 2022, doi: https://doi.org/10.31185/wjcm.60.

[18] *Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability*. [Online]. Available: https://www.mdpi.com/2071-1050/9/6/899

[19] H. A. Almustawfiy, 'Creative Accounting Applications, Oppotunistic Behavior, and Integrity of Accounting Information System: The Case of Iraq', vol. 24, no. 6, pp. 1–12, 2021.

[20] M. Al Shetwi, 'Earnings Management in Saudi Nonfinancial Listed Companies', *Int. J. Bus. Soc. Sci.*, vol. 11, no. 1, 2020, doi: 10.30845/ijbss.v11n1p3.

[21] A. F. Al-Hassan, 'Earnings Management using Accruals: Empirical Study on Saudi Companies', *Arab. J. Adm.*, 2018, [Online]. Available: http://search.mandumah.com/record/940867

[22] R. Mubeen, D. Han, J. Abbas, S. Álvarez-Otero, and M. S. Sial, 'The Relationship Between CEO Duality and Business Firms' Performance: The Moderating Role of Firm Size and Corporate Social Responsibility', *Front. Psychol.*, vol. 12, p. 669715, Dec. 2021, doi: 10.3389/fpsyg.2021.669715.

[23] Saul Calderon-Ramirez, Shengxiang Yang, David Elizondo, 'Semi-supervised Deep Learning for Image Classification with Distribution Mismatch: A Survey', *IEEE Trans. Artif. Intell.*, 2022.

[24] R. A. Leitch and Y. Chen, 'Simulation of Controlled Financial Statements', *Rev. Quant. Finance Account.*, vol. 13, no. 2, pp. 189–207, Sep. 1999, doi: 10.1023/A:1008304127780.

[25] I. D. L. TORRE, *Creative Accounting Exposed*. Palgrave Macmillan London, 2008.

[26] C. W. Mulford and E. E. Comiskey, *The Financial Numbers Game: Detecting Creative Accounting Practices*. John Wiley & Sons; 1st Edition, 2002.

[27] F. Tassadaq and Q. A. Malik, 'Creative accounting and financial reporting: Model development and empirical testing', *Int. J. Econ. Financ. Issues*, vol. 5, no. 2, pp. 544–551, 2015.

[28] CMA, 'The Most Prominent Observations On The Disclosures Of The Financial Statements Of Listed Companies', 2021 2020.

[29] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[30] V. Kamath, R. A., V. G. Kini, and S. Prabhu, 'Exploratory Data Preparation and Model Training Process for Raspberry Pi-Based Object Detection Model Deployments', *IEEE Access*, vol. 12, pp. 45423–45441, 2024, doi: 10.1109/ACCESS.2024.3381798.

[31] Y. Bengio, Y. Lecun, and G. Hinton, 'Deep learning for AI', *Commun. ACM*, vol. 64, no. 7, pp. 58–65, Jul. 2021, doi: 10.1145/3448250.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 111–119
DOI: 10.15439/2025F6165
ISSN 2300-5963 ACSIS, Vol. 43

Comparison of Large Language Models Supporting the Polish Language in Terms of Faithfulness in Retrieval-Augmented Generation Applications

Marcin Blachnik, Jakub Chmielewski
0000-0003-3336-4962, –
Silesian University of Technology
ul. Akademicka 2A, 44-100 Gliwice, Poland
Email: marcin.blachnik@polsl.pl

***Abstract*—This article presents an evaluation of Large Language Models with support for the Polish language, focusing on their ability to accurately extract detailed information embedded within input text called Faithfulness. This scenario reflects a typical use case in Retrieval-Augmented Generation systems, where precise factual recall is critical. For this purpose, a modified needle-in-a-haystack test was conducted, in which all queries targeted numerical values concealed within extended textual contexts. The evaluation was based on recent reports from Poland's Central Statistical Office (GUS), ensuring that the content was not included in the training data of the evaluated models.**

The results demonstrate that the best-performing model was NeuralDaredevil-8B-Abliterated, followed by PLLuM-12B-instruct and Bielik-11B-v2.3-Instruct. Notably, the error rate of NeuralDaredevil-8B-Abliterated was approximately half that of the second- and third-ranking models, marking a significant performance gap. The article also explores potential explanation for these discrepancy.

I. Introduction

ONE of the prominent and increasingly common applications of large language models (LLMs) is the integration with external information retrieval systems through a method known as Retrieval-Augmented Generation (RAG). RAG bridges the capabilities of LLMs with classical retrieval-based techniques, forming a hybrid system that combines the flexibility of generative models with the precision of structured data access. In this approach, a user query submitted to the RAG system is first processed by an information retrieval component, which selects the most relevant source documents from a predefined knowledge base. These retrieved documents are then appended to the original query and collectively passed to the language model. As a result, the model's response is informed not only by the user's input but also by the curated external content, effectively enriching the context provided to the LLM.

This architecture addresses several key challenges associated with the practical deployment of LLMs in business and industrial settings. One significant advantage of RAG is its ability to provide access to up-to-date and domain-specific information without requiring time-consuming and resource-intensive model retraining. Furthermore, RAG supports fine-grained permission control, allowing systems to retrieve only those documents that a user is authorized to access—thus aligning with enterprise data governance requirements. Perhaps most critically, RAG contributes to reducing the incidence of hallucinations—a common issue in generative models wherein outputs are plausible-sounding but factually incorrect. By grounding responses in retrieved, authoritative documents, RAG enhances the factual accuracy and reliability of generated outputs, making it a valuable framework for knowledge-intensive tasks across domains.

A significant challenge in the deployment of large language models (LLMs) lies in the limited availability of models that support the Polish language. This issue is particularly acute in business and institutional contexts, where data privacy and security requirements often necessitate on-premise deployment of LLMs rather than reliance on commercial cloud-based solutions. Despite the growing ecosystem of open-weight LLMs developed by leading research consortia—such as LLAMA, MISTRAL or QWEN—these models typically lack adequate support for Polish, making them less suitable for direct application in Polish-language tasks without additional fine-tuning or adaptation.

Although the Hugging Face platform offers a diverse collection of LLMs fine-tuned for the Polish language, preliminary investigations reveal that only a small subset of these models demonstrate the reliability and performance required for real-world applications. Many available models exhibit functional limitations, including issues such as repetitive text generation or outright failure to produce coherent outputs. These limitations underscore the need for a systematic evaluation of Polish-capable LLMs, particularly those with open weights suitable for secure, local deployment.

To address this gap, a curated selection of ten models was identified for detailed evaluation. These include four models with a context window of 4096 tokens: Bielik-11B-v2.3-Instruct, Bielik-7B-v0.1, trurl-2-13b, and gpt-3.5-turbo-instruct; five models with an extended context window of 8192 tokens: NeuralDaredevil-8B-Abliterated, Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT, Kruk-7B-SP-001, Starling-

Topical area: Advanced Artificial Intelligence in Applications

LM-7B-alpha, and OpenChat-3.5-0106-Gemma; and one model, PLLuM-12B-instruct, supporting a 128k token context window. This selection serves as the foundation for comparative analysis in terms of stability and accuracy of identifying the detailed knowledge expressed in Polish-language.

Various methods for evaluating large language models (LLMs) have been proposed in recent literature [1]. However, when it comes to Retrieval-Augmented Generation (RAG) applications, the evaluation task often becomes more nuanced. In such scenarios, the primary objective is to assess the model's ability to extract specific and precise information embedded in a larger body of text—particularly when the required detail may be a minor element, potentially buried within an extensive document. To verify performance under these conditions, a modified version of the "needle in a haystack" test [2] was employed in this study. That allows for verification of the quality of the detailed knowledge extraction mechanism implemented in the LLMs measuring the Faithfulness of the model.

The article is structured as follows. The next section provides a short overview of LLM's evaluation techniques, then in section III the modified "needle in a haystack" test is described. Next in section IV, the details on the performed experiments are provided, and in the following section, the obtained results are discussed. The last section discusses the obtained results, identifying possible sources of differences between these models.

II. Large Language Models Evaluation Methods

Evaluating the performance of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems is essential for ensuring their reliability, robustness, and alignment with real-world applications [3], [4]. Typically for LLM evaluations metrics such as BLEU, ROUGE, METEOR, and LEPOR are used [5]. They evaluate the n-gram overlap between generated and reference texts. These metrics are primarily used in tasks like machine translation and summarization. However, they may fall short in capturing semantic correctness and contextual relevance. In terms of semantic evaluation the common approach is based on embedding-based metrics leverage vector representations (e.g., Sentence-BERT) [6] to quantify the semantic similarity between outputs and reference texts. Additionally, LLM-as-a-Judge methodologies employ a language model to assess generation quality based on coherence, relevance, and factual correctness, offering a more nuanced and human-like evaluation [7]. Another approach for LLM's evaluatino is based on human evaluation. Despite being resource-intensive, human evaluation remains the gold standard. Annotators rate generated outputs based on fluency, factual accuracy, and relevance, although results may be subject to variability across raters. There are also several automated evaluation frameworks among which DeepEval [8] and ARES [9] have streamlined the model evaluation process. The DeepEval integrates various metrics and supports LLM-as-a-Judge assessments, while ARES reduces the dependency on human annotations by training lightweight evaluators for tasks such as context relevance and answer faithfulness.

In the context of RAG, the evaluation task becomes more complex due to the multi-component architecture of the system. A typical RAG pipeline includes a document corpus, an information retrieval module, and a large language model (LLM) that processes the retrieved document chunks to generate a final response. Each of these components can be evaluated individually, as well as in terms of their overall contribution to the system's performance.

The retrieval module is commonly assessed using several key metrics. Relevance measures how well the retrieved documents correspond to the user query. Comprehensiveness evaluates the diversity and coverage of retrieved content, ensuring that different aspects of the query are captured. Correctness refers to the accuracy of retrieved documents compared to all possible relevant candidates. Context Relevance assesses whether the retrieved context is sufficient to support a correct and complete response to the query. Fore more details see [10].

The LLM component can also be evaluated using distinct criteria. Faithfulness captures the degree to which the generated response accurately reflects the information found in the retrieved documents. Relevance, in this context, refers to the alignment of the generated response with the user's query intent. Key Point Recall measures how well the response incorporates essential information from the retrieved content. Response Completeness evaluates whether the answer fully addresses the user's query, while Response Conciseness assesses the amount of extraneous or irrelevant content present in the response. (see [11])

In this work, we focus only on the Faithfulness measure of the models supporting the Polish language.

III. The Modiffied "needle in a haystack" test

The classical formulation of the needle in a haystack test involves embedding a sentence containing a specific piece of factual information into a longer textual passage, followed by querying the model with a prompt that indirectly or directly references this content. While this method provides a foundational framework for assessing factual recall in language models, it exhibits several limitations. In particular, the artificial insertion of a standalone sentence often lacks semantic coherence with the surrounding context. This disjunction can disrupt the natural flow of the passage and interfere with the attention mechanisms of transformer-based architectures, potentially introducing artifacts that obscure the model's true retrieval capabilities. Consequently, such tests may not reliably reflect performance under realistic usage scenarios.

To overcome these shortcomings, the evaluation methodology was refined by embedding the target information in a more contextually integrated manner. Instead of inserting isolated factual statements, passages were selected or constructed to ensure that critical details were naturally embedded within a coherent narrative structure. Queries were then formulated to require comprehension, synthesis, and accurate retrieval of these embedded facts. In particular, we focused on numerical

facts that can be easily and precisely evaluated, allowing for direct measurement of the quality. In particular, we measured the relative error of the returned numerical value. This approach better reflects real-world RAG conditions, where relevant information is interwoven with broader context, and ensures a more robust and valid assessment of model behavior.

Beyond these refinements, the needle in a haystack framework was extended to investigate two additional factors critical to LLM evaluation: context length and needle position. The first factor concerns the hypothesis that shorter textual contexts simplify retrieval by reducing the model's search space, whereas longer contexts pose greater challenges due to increased sequence length and potential dilution of attention. The second factor addresses the position of the target information within the passage. Intuitively, content located at the beginning or end of a prompt may be more salient and thus more easily retrieved, whereas information embedded in the middle of the context may be less accessible. Evaluating model performance across these dimensions provides deeper insight into their retrieval fidelity and helps characterize their suitability for deployment in practical, high-recall RAG systems.

IV. The Experiment Setup

For the purpose of model evaluation, a test corpus was constructed using recently released reports from Poland's Central Statistical Office (GUS). The use of this data ensured that the language models under evaluation had not encountered the content during their pretraining, thereby minimizing the risk of data leakage and enabling a more rigorous assessment of generalization and retrieval capabilities. Notably, the information queried from the models consisted primarily of numerical values, which served as the "needles" in the evaluation framework. The use of numeric data allowed for an objective and precise evaluation of retrieval accuracy, as the model outputs could be directly compared against ground truth values.

A detailed list of the source documents used in the study, along with the corresponding needle information, is provided in Table I.

The query used in the evaluation is presented in Table II

The evaluation method for scoring the quality of the model is shown in Figure 1. It searches for a part of the text containing numbers and then calculates the relative error between the extracted value and the true value. If the answer doesn't contain numerical values, the returned error is 1.

An additional critical factor examined in this study was the influence of input text length on retrieval performance. The evaluation encompassed models with three different context window sizes: 4096 tokens, 8192 tokens, and one model supporting an extended context length of 128k tokens. Accordingly, two main families of experiments were conducted: one for models limited to a 4096-token context, and another for those capable of handling 8192-token contexts. To ensure consistency in the comparative analysis, the 128k-token model was evaluated using inputs constrained to 8192 tokens.

Table I: Documents, needles and the query used in the experiments. The documents were obtained from GUS website.

id	**Document title**	**Needle**	**Query**
1	Efektywność wykorzystania energii w latach 2012–2022	Według scenariusza rekomendowanego do 2050 r. ponad 66% budynków zostanie doprowadzonych do standardu pasywnego	Ile procent budynków zostanie doprowadzonych do standardu pasywnego do 2050 r.?
2	Efektywność wykorzystania energii w latach 2012–2022	W podziale na sektory wskaźnik ODEX brutto wykazywał poprawę efektywności energetycznej w przemyśle (o 57,2% w porównaniu do 2000 r.)	Jaką poprawę efektywności energetycznej w przemyśle wykazał wskaźnik ODEX brutto w porównaniu do 2000 r.?
3	Powierzchnia i ludność w przekroju terytorialnym w 2024 r.	Tereny wiejskie obejmujące gminy wiejskie i obszary wiejskie w gminach miejsko -wiejskich zajmują łącznie powierzchnię 29 012 600 ha, co stanowi 92, 42% obszaru Polski.	Jaki procent obszaru Polski zajmują łącznie tereny wiejskie?
4	Uczenie się osób dorosłych 2022	Na wsi udział uczniów szkół o profilu zawodowym wyniósł 23,8%	Jaki był udział szkół o profilu zawodowym na wsi?
5	Uczenie się osób dorosłych 2022	Znajomość więcej niż jednego języka obcego deklarowało 27,0% badanych.	Jaki procent badanych deklarowało znajomość więcej niż jednego języka obcego?
6	Wybrane wskaźniki przedsiębiorczości w latach 2018–2022	W 2022 r. podmioty małe (o liczbie pracujących od 10 do 49 osób) stanowiły 53% przedsiębiorstw szybkiego wzrostu oraz 79% szybkiego spadku	Jaki procent przedsiębiorstw szybkiego spadku stanowiły podmioty małe w 2022r.?
7	Wybrane wskaźniki przedsiębiorczości w latach 2018–2022	W 2022 r. wśród przedsiębiorstw stabilnych 64% stanowiły jednostki małe, a 29% podmioty średnie.	Jaki procent przedsiębiorstw stabilnych stanowiły podmioty średnie w 2022r.?
8	Powszechny Spis Rolny 2020 Charakterystyka gospodarstw domowych rolników na podstawie połączonych danych z PSR 2020 i NSP 2021	W okresie dziesięciolecia 2010–2020 wzrosła liczba gospodarstw najmniejszych o powierzchni do 1 ha UR włącznie (o 1,6%)	O jaki procent wzrosła liczba gospodarstw o powierzchni do 1 ha UR w okresie dziesięciolecia 2010- 2020?
9	Powszechny Spis Rolny 2020 Charakterystyka gospodarstw domowych rolników na podstawie połączonych danych z PSR 2020 i NSP 2021	Blisko połowa (48,6%) ludności wiejskiej tworzącej gospodarstwa domowe z użytkownikiem posiada wykształcenie zasadnicze zawodowe lub podstawowe.	Jaki procent ludności wiejskiej tworzącej gospodarstwa domowe z użytkownikiem posiada wykształcenie zasadnicze zawodowe lub podstawowe?
10	Koniunktura w przetwórstwie przemysłowym, budownictwie, handlu i usługach 2000-2024	W porównaniu z lipcem ub.r. wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników (z 17,1% do 21,9%)	Do ilu procent wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników w porównaniu z lipcem ub. r.?

Table II: Prompt used for model evaluation. The *query* and *needle* are provided in Table I

```
Kontekst: {context including needle}
Pytanie: {query}
Jako odpowiedź możesz podać tylko liczbę.
Jeżeli nie znajdziesz wyniku napisz BRAK.
Odpowiedz po polsku.
```

```python
def evaluate_response(response, true_answer):
    pattern = re.compile(r"[0-9]+,[0-9]+|[0-9]+\.[0-9]+|[0-9]+"re.IGNORECASE)
    match = pattern.findall(response.replace(" ", ""))
    if not match:
        return evaluate_response_body(response, true_answer)

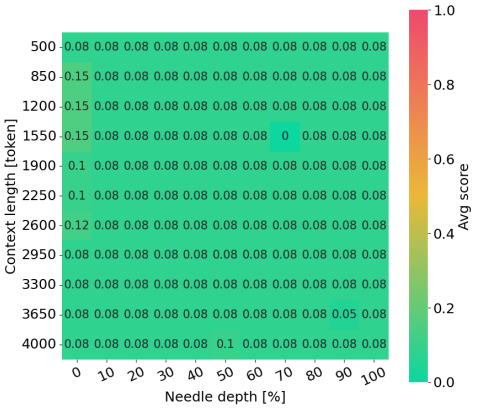
    best = 1
    for m in match:
        score = evaluate_response_body(m, true_answer)
        if best > score:
            best = score
    return best
```

(a) Evaluate_response function

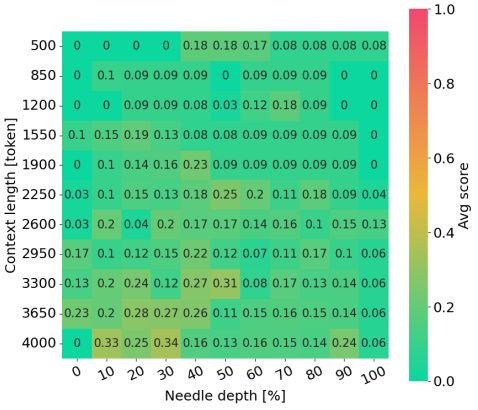
```python
def evaluate_response_body(response, true_answer):
    response = response.replace("%", "").replace(",", ".")
    try:
        response = float(response)
        return abs(true_answer - response) / true_answer
    except ValueError:
        return 1
```

(b) Evaluate_response_body function

Figure 1: Functions used for evaluation the response returned by the LLMs



(a) Bielik-11B-v2.3-Instruct



(b) Bielik-7B-v0.1

Figure 2: The results obtained for models with 4k context length. Part A. X-axis represent the depth of the needle, and Y-axis represent context length.

To systematically assess the effect of input size, each model was tested on 11 different text lengths. Additionally, the influence of the needle's position within the input was evaluated by placing the target information at 11 distinct locations throughout the text, including the beginning, middle, and end. This design enabled a comprehensive investigation into how retrieval performance varies with respect to both input length and the positional salience of the information.

The following models were selected for evaluation using a 4096-token context window:

- Bielik-11B-v2.3-Instruct,
- Bielik-7B-v0.1,
- trurl-2-13b,
- gpt-3.5-turbo-instruct,

and for the context length of 8196 tokens the following models were evaluated

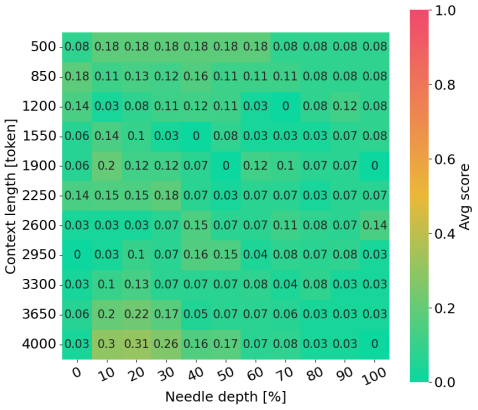
- NeuralDaredevil-8B-Abliterated,
- Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT,
- Kruk-7B-SP-001,
- Starling-LM-7B-alpha,
- OpenChat-3.5-0106-Gemma,
- PLLuM-12B-instruct.

V. Results and Discussion

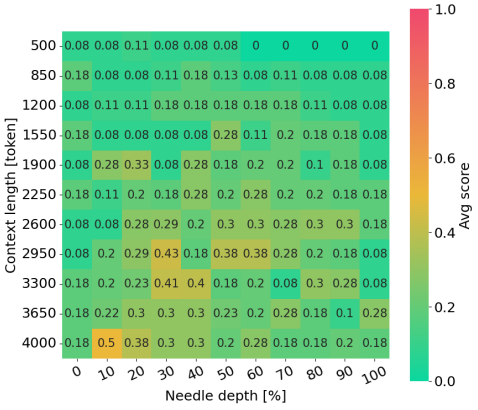
As described in the previous section, the experiments were conducted separately for models with two different maximum context lengths. The results are presented using heatmap visualizations on a two-dimensional grid, where the X-axis represents the relative depth (position) of the needle within the input text, and the Y-axis denotes the length of the input context. Each cell in the heatmap reflects the probability of an error, with values ranging from 0 (indicating no errors across all test samples) to 1 (indicating that the model consistently failed to retrieve the correct information). Accordingly, green indicates perfect accuracy, while red denotes complete failure in retrieval. To facilitate consistent visual interpretation, the same colormap scale was used across all heatmaps, allowing for direct comparison of prediction performance across different models and experimental conditions.

A. 4k Context Length

- **Bielik-11B-v2.3-Instruct** The results are presented in Figure 2a. The findings indicate that the model exhibits



(a) Trurl-2-13b



(b) GPT3.5-turbo-instruct

Figure 3: The results obtained for models with 4k context length. Part B. X-axis represent the depth of the needle, and Y-axis represent context length.

a consistent error probability across the evaluated range, suggesting that the likelihood of an error is not significantly influenced by either the needle depth or the context length. In all cases, the observed errors corresponded to a single test case involving the needle *W porównaniu z lipcem ub.r. wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników (z 17,1% do 21,9%)* and the query *Do ilu procent wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników w porównaniu z lipcem ub. r.?*, where the model consistently returned the first numerical value instead of the second one, despite the query clearly referencing the latter.

- **Bielik-7B-v0.1** The results are shown in Figure 2b. These indicate that the model's performance is highly dependent on the position of the needle. Specifically, the model is particularly sensitive when the needle is located early in the context—around 20% of the total context length. Additionally, its performance deteriorates with longer context lengths. For the longest context length, the model failed in approximately one-third of the cases when the needle was situated at about 20% depth.

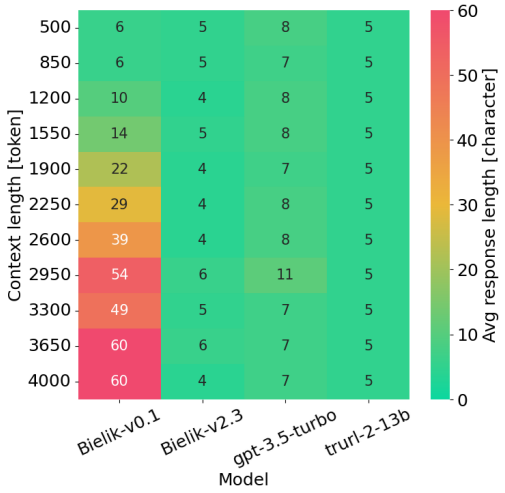


Figure 4: Relation between context length and the length of returned output in tokens for 4k models.

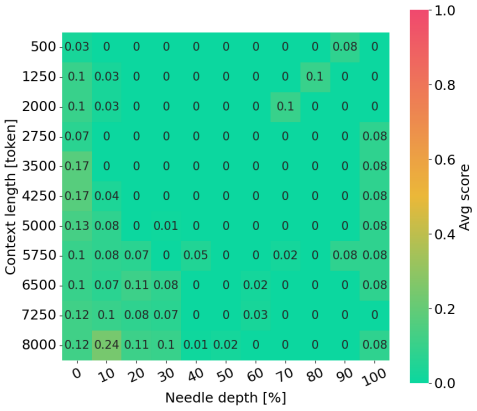
- **Trurl-2-13B** The corresponding results are presented in Figure 3a. The model exhibits a performance pattern similar to Bielik-7B-v0.1, with the most significant errors occurring when the needle is positioned at around 20% of the context depth and when the context length approaches its maximum. However, unlike Bielik-7B-v0.1, Trurl-2-13B shows improved performance when the context length is reduced to 3300 tokens or fewer.
- **GPT-3.5-turbo-instruct** The results are shown in Figure 3b. Among all evaluated models with a 4k token context window, this model demonstrated the weakest performance. In the worst-case scenario, it exhibited an error rate of up to 50% when the needle was located near the beginning of the context (approximately 10%) and at the maximum context length. It also frequently produced incorrect answers for shorter context lengths around 2950 to 3300 tokens, where Trurl-2-13B performed comparatively well. GPT-3.5-turbo-instruct only achieved reliable performance when the context length was limited to 1550 tokens or less.

In summary, the best-performing model with a 4k context length was Bielik-v2, followed by Trurl, while GPT-3.5 surprisingly demonstrated the weakest performance.

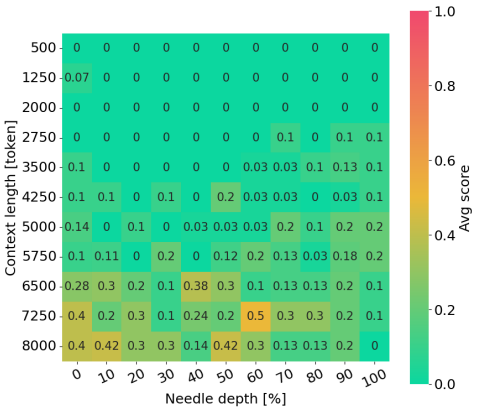
Additional insight is gained by analyzing the length of the output stream generated by the models, as shown in Figure 4. The results indicate that the Bielik-v1 model tends to produce significantly longer outputs, averaging up to 60 tokens when the context length approaches 4k. This is in stark contrast to the expected output, which typically consisted of only a few tokens representing a numerical value. This excessive verbosity contributed to a substantial increase in error rates.

B. 8k Context Length

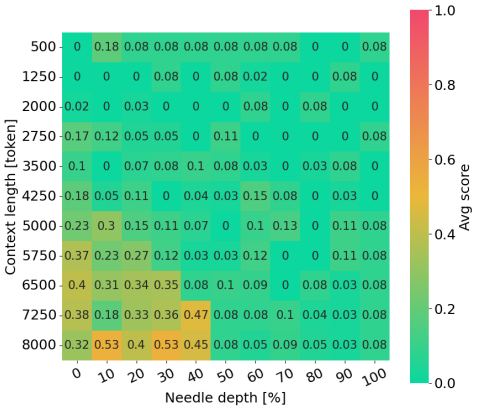
- **NeuralDaredevil-8B-Abliterated** The obtained results are shown in figure 5a. The model for a vary large part of



(a) [NeuralDaredevil-8B-Abliterated

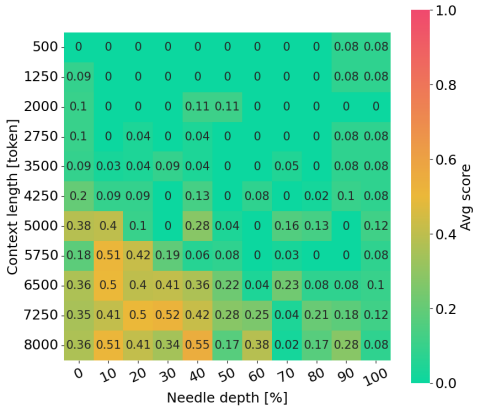


(b) Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT

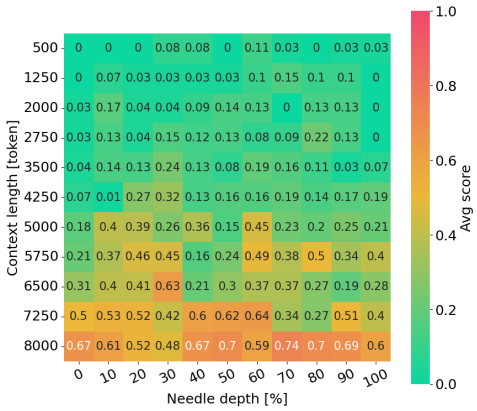


(c) Kruk-7B-SP-001

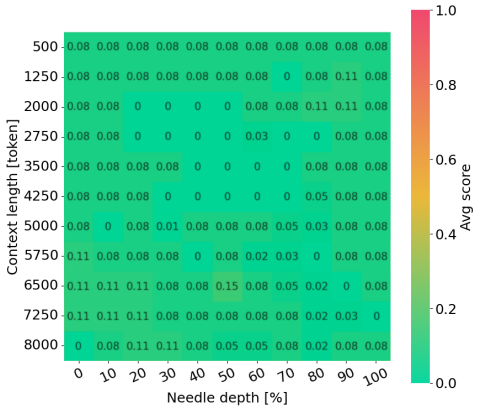
Figure 5: The results obtained for models with 8k context length. Part A. X-axis represent the depth of the needle, and Y-axis represent context length.



(a) Starling-LM-7B-alpha



(b) OpenChat-3.5-0106-Gemma



(c) PLLuM-12B-instruct

Figure 6: The results obtained for models with 8k context length. Part B. X-axis represent the depth of the needle, and Y-axis represent context length.

the experiment space didn't achieve any error. The errors appeared when the needle was located at the beginning of the context as the context length was growing. But the error rate was even though relatively low.

- **Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT** The obtained results are shown in figure 5b. It achieves perfect predictions for short context length up to 4k, but when the context length starts to grow it starts to fail. The worst results are obtained for very long context at its begining.
- **Kruk-7B-SP-001** The obtained results are shown in figure 5c. It has similar behavior to all other models. Its error rate starts to grow when context is getting long and the needle is located at the begining. When compared to the omnibus model it can be observed a significant higher error rates for context length below 4k.
- **Starling-LM-7B-alpha** The obtained results are shown in figure 6a. Starling behaves similarly to Kruk except it has better performance for shorter context. Athough, it is warse for larger context, and achives slighly higher error rates.
- **OpenChat-3.5-0106-Gemma** The obtained results are shown in figure 6b. It is the worst of evaluated 8k models. It got larger error rates for the short context, and significantly worse results for the long context, where the error rates reaches 70% when the context length is close to 8k. But when compared to the 4k models it achieves similar performance reaching 0.32 error rates for context close to 4k.
- **PLLuM-12B-instruct** The obtained results are shown in figure 6c. Again it has very good performance for short context, and similarly when the context grow it achieves larger error rates but at most reaching 34%. This allows to get the second place behind NeuralDerdevil model.

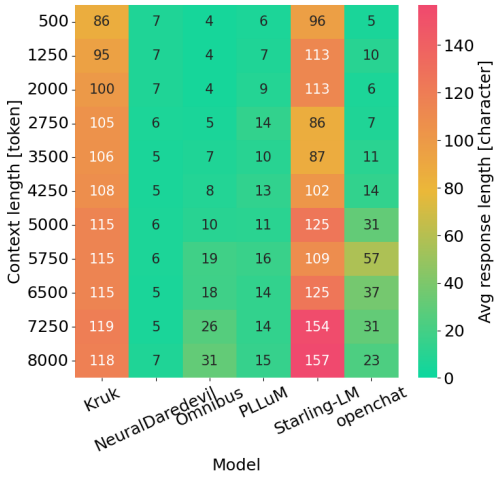


Figure 7: Relation between context length and the length of returned output in tokens for 8k models.

- **NeuralDaredevil-8B-Abliterated**
  The results are presented in Figure 5a. Across a large portion of the experimental space, the model achieved near-perfect performance. Errors occurred primarily when the needle was positioned at the beginning of the context as the overall context length increased. However, even in these cases, the error rate remained relatively low.
- **Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT**
  As shown in Figure 5b, this model achieved perfect predictions for context lengths up to 4k tokens. However, performance degraded as the context length increased, with the most significant errors observed when the needle was located near the beginning of long contexts.
- **Kruk-7B-SP-001**
  The results, depicted in Figure 5c, show a pattern similar to other models. Error rates increase with longer contexts, particularly when the needle is located near the beginning. Compared to the Omnibus model, Kruk exhibits significantly higher error rates for contexts shorter than 4k tokens.
- **Starling-LM-7B-alpha**
  Figure 6a illustrates that Starling's behavior is comparable to Kruk's. It performs better for shorter contexts but shows slightly worse performance as context length increases, with marginally higher error rates overall.
- **OpenChat-3.5-0106-Gemma**
  The evaluation results, shown in Figure 6b, indicate that this model is the weakest among the evaluated 8k-context models. It exhibits higher error rates even for short contexts and significantly poorer performance for long contexts—reaching up to 70% error when the context is close to 8k tokens. However, when limited to 4k contexts, its performance aligns with that of other 4k-context models, reaching an error rate of approximately 32%.
- **PLLuM-12B-instruct**
  The results in Figure 6c show that the model performs very well for short contexts. As with other models, error rates increase with longer contexts but remain relatively low, peaking at around 34%. This strong performance places it second only to the NeuralDaredevil model.

In summary the best performing model among 8k models was NeuralDaredevil followed by PLLuM, and the worst one is OpenChat. Similarly to the 4k models, worth deeper investigation is the output text token length. Such relation is shown in figure 7. It shows that Starling and Kruk has significant longer output length. These models insted of returning precise output value that was queried, returned full sentence, often setence containing the output text. Similarly Omnibus and OpeChat when couldn't find the answer in the text returned full sentence insted of simple and short answer.

C. Summary and Discussion

In summary, among the evaluated models, NeuralDaredevil-8B-Abliterated achieved the best performance, followed by PLLuM and Bielik. The weakest performance was observed for OpenChat - see Figure 8.

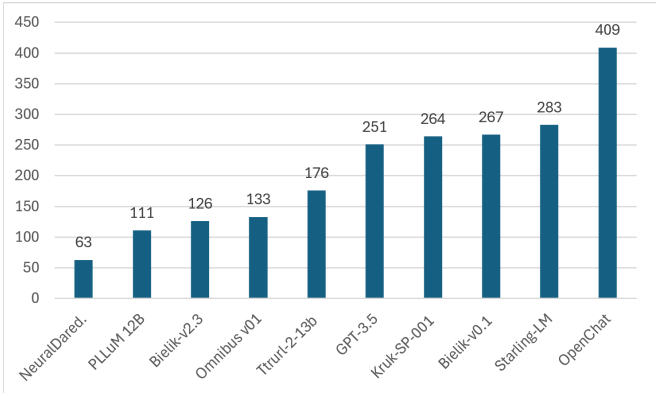


Figure 8: Number of factual errors made by each evaluated model.

It is particularly notable that the NeuralDaredevil-8B-Abliterated model significantly outperformed all other models, including PLLuM and Bielik-v2.3. The error rates for PLLuM and Bielik-v2 were nearly twice as high as those observed for NeuralDaredevil. A closer inspection revealed that the majority of errors across models were linked to a single query: *W porównaniu z lipcem ub.r. wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników (z 17.1% do 21.9%)* and the corresponding question *Do ilu procent wzrosło znaczenie barier niedoboru wykwalifikowanych pracowników w porównaniu z lipcem ub. r.?* In this case, many models incorrectly returned the initial value (17.1%) rather than the correct final value (21.9%). This suggests difficulties in interpreting comparative constructions in Polish, particularly the meaning of prepositions such as "z" (from) and "do" (to).

One possible explanation for NeuralDaredevil's strong performance lies in the abliteration process, which may enhance the model's generalization capabilities. Previous studies have shown that excessive safety alignment or over-optimization can negatively impact a model's reasoning and factual recall abilities [12], [13], [14]. NeuralDaredevil-8B-Abliterated is a fine-tuned variant of the Daredevil-8B model based on the LLaMA-3 architecture. Fine-tuning was conducted with a single pass over the `mlabonne/orpo-dpo-mix-40k` dataset using the Direct Preference Optimization (DPO) method [15]. The abliteration technique, as described by Arditi et al. [16], involves removing the activation direction associated with refusal behaviors in the transformer's residual stream. This modification may allow the model to respond more freely to prompts that might otherwise trigger refusal responses, thereby improving reasoning in neutral tasks such as factual retrieval.

Furthermore, a consistent pattern across all evaluated models was observed: retrieval performance degraded as the input text approached the maximum context length and the target (needle) appeared near the beginning of the document. This is visually evident in all the figures (2,3,5 and 6), where the bottom-left corners (representing early-position needles in long contexts) are more yellow, indicating higher error rates. In contrast, queries located in the first half of the context window often resulted in near-perfect accuracy, as seen with models such as NeuralDaredevil-8B-Abliterated, Llama-3-8B-Omnibus-1-PL-v01-INSTRUCT, and also in Starling-LM-7B-alpha. This suggests that attention limitations in transformer-based models still substantially impact retrieval success in long-context scenarios.

Interestingly, some models—namely Bielik-11B-v2.3-Instruct and PLLuM-12B-Instruct—exhibited relatively stable error rates across the entire context window. This may be a result of their fine-tuning strategies, which could help mitigate the performance degradation typically caused by long input sequences.

Summarizing the length of the returned output shown in Figure 7 and Figure 4 some models tends to return significantly longer responses. In particular Starling and Kruk consistently produced significantly longer output sequences. Instead of returning a concise value in response to the query, these models often generated full sentences that included or paraphrased the expected output. Similarly, Omnibus and OpenChat, when unable to locate the exact answer in the context, tended to return verbose responses rather than the brief, precise values requested. This behavior may contribute to higher error rates and reduced response faithfulness in scenarios requiring exact factual retrieval.

VI. Conclusions

The primary goal of this research was to evaluate the ability of language models to accurately extract factual information from Polish input texts, rather than focusing on the linguistic quality of their output. This problem was formulated to provide a rough estimate of the models' susceptibility to hallucinations when used in Retrieval-Augmented Generation (RAG) systems, particularly in the context of less commonly supported languages such as Polish. Among the evaluated models, NeuralDaredevil-8B-Abliterated clearly outperformed both PLLuM and Bielik-v2. Surprisingly, the commercial GPT-3.5 model performed poorly, exhibiting twice as many errors as PLLuM or Bielik, and four times as many as NeuralDaredevil.

Additionally, a recurring type of error was observed across models: the misidentification of numerical values in comparative statements. Specifically, models often selected the initial value instead of the final one in scenarios describing change over time. This indicates a difficulty in precisely understanding certain linguistic constructs, that is a challenge in correctly interpreting comparative expressions involving "from" (Polish: *od*) and "to" (Polish: *do*).

Finally, our results revealed that all models experienced a decline in retrieval accuracy when the input text approached the maximum context window and the relevant information (needle) was located near the beginning of the document. This highlights a continuing limitation of transformer-based architectures in processing long documents, due to the reduced effectiveness of attention mechanisms over extended contexts. These findings underscore that, despite recent advancements,

long-range dependency handling remains a significant challenge for contemporary language models.

ACKNOWLEDGMENT

This research was supported by the Silesian University of Technology, grants No. BK-227/RM4/2025

REFERENCES

[1] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma, "Leveraging large language models for NLG evaluation: Advances and challenges," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024. doi: 10.18653/v1/2024.emnlp-main.896 pp. 16 028–16 045. [Online]. Available: https://aclanthology.org/2024.emnlp-main.896/

[2] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev, "Babilong: Testing the limits of llms with long context reasoning-in-a-haystack," *Advances in Neural Information Processing Systems*, vol. 37, pp. 106 519–106 554, 2024.

[3] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736*, 2023.

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[5] T. Hu and X.-H. Zhou, "Unveiling llm evaluation focused on metrics: Challenges and solutions," *arXiv preprint arXiv:2404.09135*, 2024.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[7] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.

[8] "Deepeval," https://docs.confident-ai.com/, 2024.

[9] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "Ares: An automated evaluation framework for retrieval-augmented generation systems," *arXiv preprint arXiv:2311.09476*, 2023.

[10] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*. Springer, 2024, pp. 102–120.

[11] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[12] J.-C. Gu, H.-X. Xu, J.-Y. Ma, P. Lu, Z.-H. Ling, K.-W. Chang, and N. Peng, "Model editing harms general abilities of large language models: Regularization to the rescue," *arXiv preprint arXiv:2401.04700*, 2024.

[13] W. Yang, F. Sun, X. Ma, X. Liu, D. Yin, and X. Cheng, "The butterfly effect of model editing: Few edits can trigger large language models collapse," *arXiv preprint arXiv:2402.09656*, 2024.

[14] S. Sonkar, N. Liu, and R. G. Baraniuk, "Regressive side effects of training language models to mimic student misconceptions," *arXiv e-prints*, pp. arXiv–2404, 2024.

[15] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.

[16] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, "Refusal in language models is mediated by a single direction," *arXiv preprint arXiv:2406.11717*, 2024.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 121–132
DOI: 10.15439/2025F8736
ISSN 2300-5963 ACSIS, Vol. 43

A Framework for Model-Driven AI-Assisted Generation of IT Project Management Plan and Scope Documents

Jānis Rihards Blaževičs
Riga Technical University,
Institute of Information Technologies,
Riga, Latvia
Email: janis-rihards.blazevics@edu.rtu.lv

Oksana Nikiforova
0000-0001-7983-3088
Riga Technical University,
Institute of Information Technologies,
Riga, Latvia
Email: oksana.nikiforova@rtu.lv

Oscar Pastor
0000-0002-1320-8471
Universitat Politècnica de València, Valencia, Spain
Email: opastor@dsic.upv.es

***Abstract*—This paper presents a model-driven AI-assisted approach for generating IT project management plan and scope documents, aiming to improve efficiency and quality in software development projects. Effective documentation in the early project phase is critical, yet often resource-intensive. The proposed solution consolidates best practices from widely used project management methodologies and standards to create a dynamic, adaptable framework for document generation. The study identifies key components and input data required for generating high-quality plans using model transformations and generative AI. A prototype supporting the solution is developed featuring a local processing engine, integrated with a large language model, a vector database, and an embedded model.**

***Index Terms*—IT project management, project documentation, model-driven development, generative AI, prompt engineering, software development automation, large language models, AI-assisted tools.**

I. Introduction

IN MODERN IT project development, the demand for accelerated delivery cycles continues to increase, driven by competitive pressure, agile practices, and customer expectations [1]. One of the most critical processes in the early phases of the project lifecycle is the development of project documentation [2]. Well-structured documentation, such as project management plans and scope statements, forms the basis for mutual understanding between clients and delivery teams, and contributes directly to project success and client satisfaction [3].

To address the need for speed without compromising quality, automation tools have been introduced to support documentation processes. These tools can generate content based on reusable templates, domain knowledge, and recognized best practices [4]. However, most existing solutions act as passive assistants, requiring significant manual effort to ensure completeness, consistency, and compliance with professional standards [5].

Since the emergence of generative artificial intelligence (AI) technologies around 2020 [6], there has been growing interest in using AI to automate the creation of project documentation. Generative models such as large language models (LLMs) [7] offer promising capabilities for producing structured, domain-specific texts [8]. However, researchers should overcome challenges related to the need for precise prompt formulation, ensuring contextual relevance, managing data quality, and aligning with industry standards [9]. Addressing these challenges requires not only technical solutions but also methodological guidance.

This raises a fundamental research question: Is it possible to integrate generative AI into documentation development without compromising quality and compliance? This paper investigates this question by proposing an AI-assisted solution grounded in software engineering principles, aiming to maintain documentation integrity while leveraging the potential of generative AI. The research goal of the results presented in this paper is to develop an AI-assisted solution for generating IT project management plan and scope documents applicable in software development projects, by consolidating best practices in project documentation as defined across various development methodologies and industry standards.

The paper is organized as follows. The next section discusses the research background and summarizes the related literature. Section 3 presents the foundations used in the solution and Section 4 explains the essence of the solution. Section 5 demonstrates the practical application of the solution and comments its validation and improvements. The final section summarizes the main research findings and outlines directions for future work.

This research has been supported by Research and Development grant No RTU-PA-2024/1-0015 under the EU Recovery and Resilience Facility funded project No. 5.2.1.1.i.0/2/24/I/CFLA/003 “Implementation of consolidation and management changes at Riga Technical University, Liepaja University, Rezekne Academy of Technology, Latvian Maritime Academy and Liepaja Maritime College for the progress towards excellence in higher education, science, and innovation”.

II. Background and Related work

Since the emergence of generative artificial intelligence (AI) technologies around 2020, such as OpenAI's GPT-3 [6], there has been growing optimism that AI could dramatically streamline project documentation by enabling the automatic generation of high-quality documents in just a few clicks. The rapid advancement of LLMs has fueled expectations that these tools could reduce the manual effort required for creating project specifications, requirement documents, and other critical deliverables. However, several significant challenges remain in realizing this vision fully:

1. Generative AI often struggles to capture the specific organizational context, domain terminology, and stakeholder expectations that are essential for meaningful documentation [9].
2. Effective AI-generated documentation depends heavily on the availability of structured, high-quality input data. In many real-world projects, this data may be incomplete, ambiguous, or inconsistent [10].
3. Automatically generated content can raise concerns about traceability, authorship, and the responsibility for errors or omissions in critical documents [11].
4. Industry-specific methodologies and compliance requirements (e.g., ISO, IEEE) can be difficult to encode into generic AI systems without significant customization [12].
5. Even advanced generative tools require skilled human supervision to ensure accuracy, relevance, and alignment with project goals [13].

These limitations suggest that while generative AI has strong potential to support project documentation, its role is currently best viewed as augmentative rather than fully autonomous. Future developments may narrow these gaps, but effective human-AI collaboration remains essential for reliable outcomes in IT project documentation.

Recent studies have demonstrated the growing potential of generative AI in enhancing project management processes, particularly in the areas of documentation, planning, and efficiency. One of the key applications of generative AI is the automation of project documentation, including project plans and scope definitions. AI models can consolidate and structure content from various sources to generate standardized and coherent documents, ensuring compliance with project management standards and improving overall consistency [14]. Additionally, generative AI is increasingly used to generate meeting summaries, translate technical content, and update project texts, contributing to more accurate and timely documentation [15].

In project planning, AI tools can support a wide range of functions such as scope definition, scheduling, cost estimation, resource allocation, stakeholder analysis, and risk assessment. Some studies suggest that generative AI can match or even outperform human planners in certain structured planning tasks, while still requiring expert oversight for contextual decisions [16]. Moreover, the integration of AI in Agile methodologies has shown promise in optimizing team workflows, enhancing collaboration, and automating repetitive tasks [17]-[18].

The collaboration between human project managers and AI systems has been identified as a critical factor for success. Key performance indicators for such collaboration include efficiency gains in time, cost, and resource usage, as well as improved quality, clarity, and accuracy of project outputs [19]. Research also highlights the benefits of using generative AI in the early stages of innovation, such as digital prototyping and ideation, where rapid iteration and content generation are crucial [20].

However, several challenges remain. Effective integration of AI into project management environments must address issues such as input data quality, compatibility with legacy systems, and ethical concerns regarding algorithmic transparency and accountability [21]-[22]. Furthermore, while generative AI can produce initial drafts of project documentation, human expertise is still essential to validate and refine the output to ensure relevance, accuracy, and compliance with specific organizational contexts [16].

These findings underscore the dual role of generative AI as both an accelerator and a collaborator in project management. The literature suggests that while generative AI tools provide substantial benefits, their successful implementation relies on structured prompts, methodological support, and continuous human oversight.

To address the limitations of generative AI in project documentation, Nikiforova, *et al.* [23] proposed a solution based on the integration of model-driven development (MDD) principles. By embedding structured models into the AI-assisted documentation process, the approach aims to constrain and guide AI output according to predefined rules, templates, and domain-specific semantics. This model-driven approach helps overcome common issues such as contextual inconsistency, structural errors, and lack of standard compliance. It also enhances traceability and maintainability by linking generated content to formal models that represent project requirements and design logic.

In response to the limitations of generative AI tools, researchers and practitioners have increasingly focused on the field of prompt engineering, which is the practice of crafting and structuring input prompts in a way that guides AI systems to produce higher-quality, more relevant outputs [24]. Rather than relying on generic instructions, prompt engineering emphasizes precise formulations, contextual cues, and example-based guidance to improve the performance of language models in complex tasks such as project documentation. The emerging field of prompt engineering aims to optimize prompt formulation in order to elicit high-quality, contextually accurate responses from generative models [25].

This emerging discipline is particularly relevant in domains like IT project management, where document quality is influenced by terminology accuracy, logical structure, and compliance with standards. By experimenting with different prompt

patterns, researchers aim to understand how to systematically elicit responses that align with user expectations and domain-specific requirements.

Early studies and practical experiments indicate that carefully designed prompts can significantly enhance the clarity, completeness, and consistency of AI-generated documents. As a result, prompt engineering is becoming a critical skill for those seeking to integrate generative AI into documentation workflows, bridging the gap between AI capabilities and real-world needs.

The research presented in this paper builds upon the idea of constraining generative AI to reduce hallucinations and uncontrolled variability in generated content. The proposed approach is oriented toward leveraging the strengths of model-driven engineering by generating as much documentation content as possible through formal model transformations [26]. Where model transformation is insufficient, such as in the generation of narrative or descriptive elements, prompt engineering is applied, guided by artifacts from the target model [27]. This ensures that even AI-generated free-text sections remain contextually relevant, structurally aligned, and traceable to defined project components. The overall goal is to balance AI creativity with formal control, reducing risks while maintaining efficiency and adaptability.

III. Research Methodology and Conceptual Framework

This research follows the Design Science Research (DSR) methodology as proposed by Hevner et al. [28], which is widely used for the development and evaluation of IT artifacts. The goal is to create a novel, AI-assisted solution for generating IT project management plan and scope documentation. This artifact integrates model-driven engineering principles with generative AI capabilities to support early-phase documentation in software development projects.

In accordance with Hevner's seven guidelines for DSR, this study:

1. Designs an innovative artifact – a model-driven framework and software prototype for document generation;
2. Addresses a relevant business problem – namely, the inefficiency and inconsistency in early-phase project documentation;
3. Is evaluated through practical application and iterative refinement;
4. Contributes both a utility-focused solution and theoretical insights into the integration of generative AI with structured project management models;
5. Applies rigorous methods for modeling, implementation, and validation;
6. Communicates the research effectively to both academic and practitioner audiences;
7. Clearly situates the research within the context of information systems and project management disciplines.

The proposed approach balances the creative potential of generative AI with the structural rigor of model-driven development. The main objective is to constrain AI-generated content within well-defined semantic and syntactic boundaries, ensuring compliance with established project management standards described in the next subsections.

A. Model-Driven Development Principles

Model-Driven Development (MDD) is a software engineering approach that emphasizes the use of formal models as the primary artifacts throughout the development lifecycle [29]. Instead of writing low-level implementation code directly, developers create abstract models that define system structure, behavior, and logic [30]. These models are then transformed into executable code or other technical artifacts through automated tools. The main goal of MDD is to increase productivity, improve consistency, and reduce human error by shifting development effort to higher levels of abstraction.

MDD was originally developed to address complexity in large-scale systems and to enable better alignment between business requirements and technical implementation. It has proven particularly effective in domains requiring precision, standardization, and traceability, such as embedded systems, enterprise architectures, and critical software solutions.

Based on systematic literature mapping of solutions offered under the MDD idea, Nikiforova *et al.* [31] conducted an in-depth analysis of transformations between key artifacts used in the early phases of IT projects, that is, before the implementation stage. The study focused on identifying how different project artifacts, such as goal models, process models, use case models, domain models, and architectural overviews, can be interrelated through formal transformation rules.

The research highlights that certain artifacts can be automatically derived from others using well-defined transformation methods, significantly reducing manual effort and increasing consistency across documentation. For instance, structured requirements can be transformed into initial system models or project plans with minimal human intervention. However, not all relationships are fully automatable (shown as dotted lines in Fig. 1), some transformations still require manual refinement or validation, especially in cases involving complex contextual interpretation or stakeholder input. Exactly, these transformations can be assisted with the usage of generative AI.

An overview of all these artifact transformations and their interdependencies is illustrated in Fig. 1 [27]. This schema serves as a foundation for the proposed AI-assisted generation process, where transformation rules help define the scope of content that can be automatically constructed and guide the prompt generation strategy for generative AI components.

Although MDD adoption has declined in recent years, partly due to the rise of more lightweight, iterative approaches like Agile [32], the core idea of structuring development around formalized models remains powerful. In the context of generative AI, MDD principles offer a unique advantage: they can provide a well-defined, controlled input structure for guiding AI outputs.

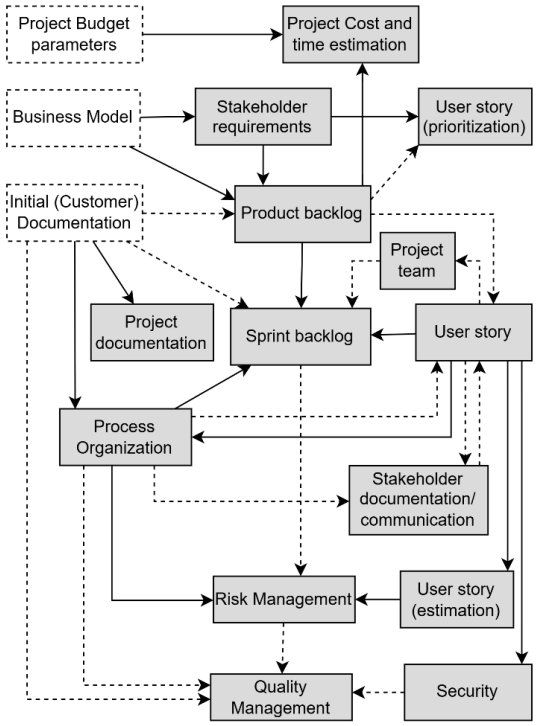


Fig 1. Transformation among IT project artifacts

This makes it possible to constrain generative models within meaningful boundaries, improving the reliability and quality of generated content. By reintroducing MDD concepts into AI-assisted documentation generation, this research aims to combine the rigor of model-driven processes with the flexibility of modern language models, achieving better control, relevance, and integration in project management documentation workflows.

B. Software Engineering Standards

The implementation of software engineering standards plays a vital role in ensuring consistency, quality, and traceability in IT project documentation. Given the increasing complexity of software development projects and the integration of AI-assisted tools, aligning documentation processes with internationally recognized standards becomes not only beneficial but necessary. The proposed solution leverages the following software engineering standards:

1. ISO/IEC/IEEE 29148:2018 Systems and software engineering — Life cycle processes — Requirement's engineering [33];
2. ISO/IEC/IEEE 16326:2019 Systems and software engineering — Life cycle processes — Project management, 2019 [34];
3. ISO 21502:2020, Project, programme and portfolio management — Guidance on project management, 2020 [35];
4. Corresponding elements of SWEBOK [36];
5. Corresponding elements of PMBOK [37].

Software engineering standards are used in two ways. The first, they define the macrostructure of the document (like in the way presented in Fig. 2) [31].

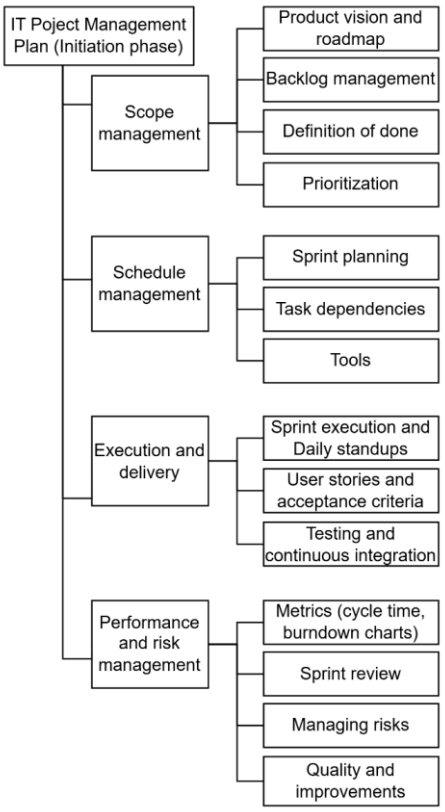


Fig 2. IT project documentation structure and content

The standards' instructions are used to guide the structure and content of generated project documents. These standards provide clear expectations regarding the format, terminology, and required content of artifacts such as project charters, requirement specifications, and management plans.

And, the second, the standards shape the micro-level instructions that guide the AI in generating specific paragraphs or sections. Rather than embedding static prompts into the generative AI component, this solution dynamically generates prompts based on structured input, specifically utilizing content derived from model transformations. The AI model is thus directed by prompts that are themselves informed by software engineering standards and documentation guidelines. This ensures that generated outputs are not only linguistically fluent but also semantically aligned with professional project management practices.

For example, prompts for requirement specification generation integrate instructional content drawn directly from ISO/IEC/IEEE 29148:2018, such as the need for completeness, verifiability, traceability, and feasibility.

Additionally, content expected from model transformations, such as stakeholder roles, system goals, and component descriptions, is automatically fed into the prompt generation process. This staged approach enables the AI to act within clearly defined semantic boundaries while maintaining flexibility for natural language output. This mechanism represents a hybrid strategy that combines the strength of model-

driven development with the adaptive capabilities of generative AI. By grounding prompt generation in formal standards and structured content, the solution significantly reduces the risk of hallucinations and irrelevant output, ensuring higher reliability and usability of project documentation.

In summary, the integration of software engineering standards serves as a scaffolding for both the structure and semantics of AI-generated project artifacts. It bridges the gap between free-form generative models and formal documentation requirements, resulting in outputs that are both technically valid and aligned with industry expectations.

IV. Solution Design and Approach

This section presents the formal framework and implementation of the proposed approach for AI-assisted generation of project documentation. The solution is built on a hybrid methodology, combining model-driven development (MDD), prompt engineering, and alignment with software engineering standards listed in previous sections.

The proposed solution is based on the principle that structured artifacts (i.e., models, templates, and metadata) should guide document generation. When natural language output is required, prompt engineering is applied, combining structured artifacts and guidance to direct the language model. This hybrid approach reduces hallucinations and improves consistency. The solution operates through several stages:

- **Input Stage:** Project-specific structured artifacts, such as project artifacts, are provided as input. Wherever possible, content is derived from model transformations, using approaches and tools like [38] [39].
- **Instruction Stage:** Documentation structure and content expectations are sourced from industry standards. These sources serve as a foundation for structuring both the overall layout and the content of the generated documentation. Excerpts from standards, such as definitions, guidelines, and section requirements, are used as instructional cues within prompts to ensure that the generated text aligns with professional expectations.
- **Transformation Stage:** Input artifacts and instruction templates are combined through structured transformation logic (defined in JSON), which dynamically generates task-specific prompts.
- **Generation Stage:** These dynamic prompts are used to control a language model (LLM) that generates natural language output. Furthermore, in the proposed solution, the prompts used by the generative AI component are not static or hardcoded. Instead, they are dynamically generated based on software engineering standards and best-practice "instructions," as well as content defined through model transformations. This means that the system first interprets structured input, such as elements derived from project models, and then constructs a tailored prompt to guide the language model.
- **Validation Stage:** Output is validated against structural schemas and evaluated on semantic alignment, completeness, and relevance.

A key innovation is the dynamic generation of prompts based on model-driven and SE standards-based inputs. Instead of using static instructions, the approach uses structured data to generate a prompt, which in turn is used to generate the final documentation text. This “generative AI powering generative AI” approach ensures that outputs are both contextually relevant and semantically accurate, thus allowing precise alignment between structured models and natural language output and reducing the likelihood of irrelevant or hallucinated content. It is like a controlled pipeline that leverages automation while maintaining contextual fidelity and compliance with documentation standards.

The system is implemented in Python using the Ollama framework. A locally hosted LLaMA 3.2:3b model handles language generation, while ChromaDB stores embedded vector representations of source content. The conceptual schema of the solutions architecture is shown in Fig. 3. Following the outlined stages, a generative AI-based solution was developed, structured around three key modules: the Input Module, the Processing Module, and the Output Module.

The **Input Module** validates and processes user-provided files, including contextual documents and control JSONs. Uploaded PDF content is split into fragments using token limits and overlap parameters to preserve context. Each fragment is embedded as a vector and stored in ChromaDB, along with metadata such as file type and source ID. Instruction files are parsed into stepwise JSON objects for controlled processing.

The **Processing Module** executes semantic queries based on the instruction steps. Each step retrieves relevant fragments using vector similarity and uses them as input for the next step. This enables incremental generation of documentation components, where each output builds on previous results. The final output is compiled into structured sections based on a predefined JSON schema.

The **Output Module** formats the completed documentation into a standards-compliant structure, ready for export or integration.

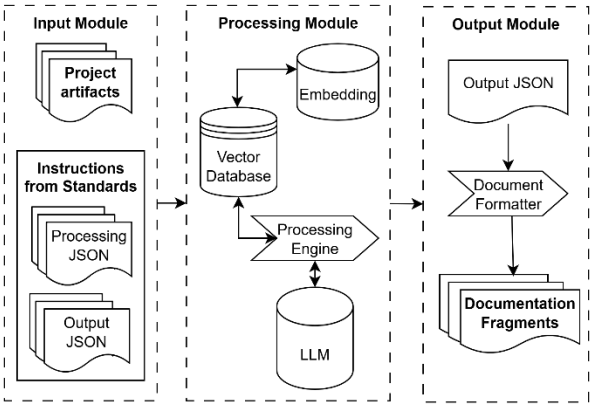


Fig 3. IT project documentation structure and content

The prototype screenshot is shown in Fig. 4. The user interface enables uploading project artifacts and control JSON files based on engineering standards. It supports “Process” and “Generate” actions to trigger analysis and document creation via a lightweight, Python-based framework.

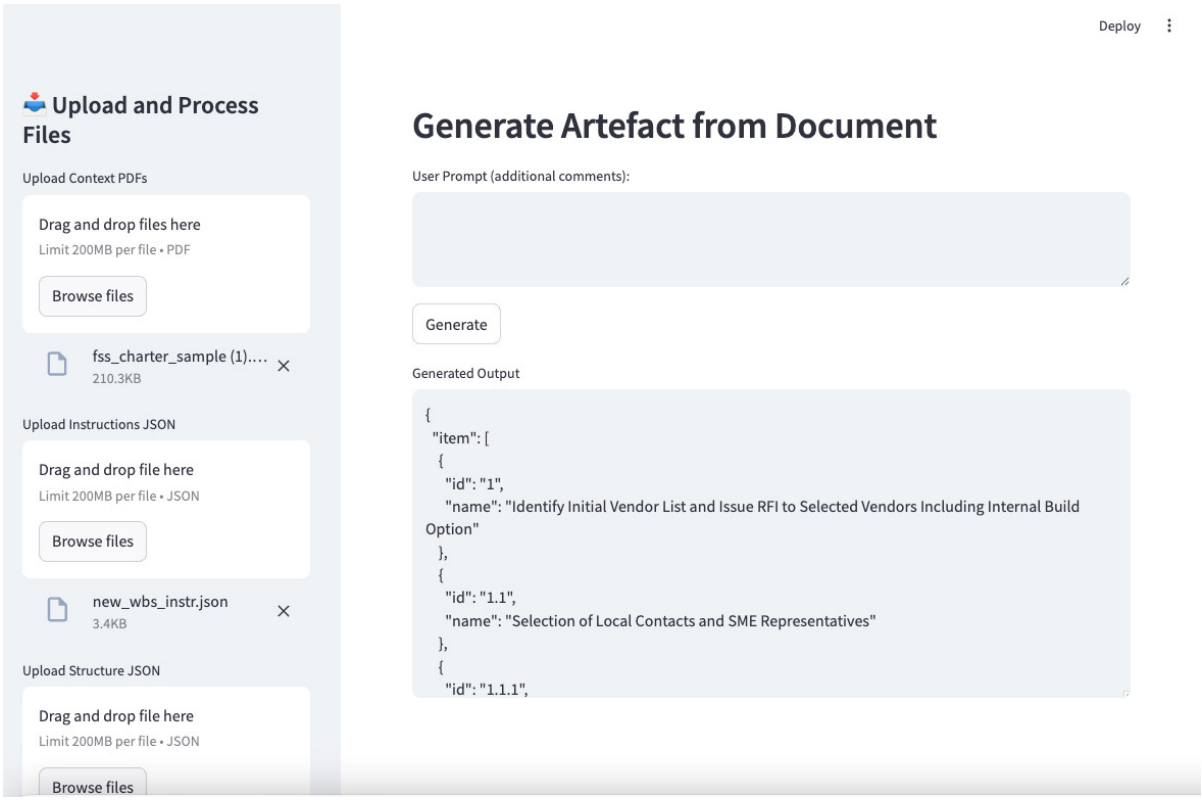


Fig 4. Solutions supporting prototype UI

V. Solution Demonstration and Validation

The developed solution focuses on the generation of documentation for the project initiation phase, which primarily are the project scope, requirements, and stakeholder register. However, the overall architecture and methodology are inherently adaptable to other types of software project artifacts (e.g., test strategies, change logs), provided that the step-based instruction approach is followed.

The design of the approach was informed by a critical analysis of limitations in existing AI-based documentation methods [41]. A key technical constraint was the limited token processing capacity of large language models (LLMs). Supplying the model with excessive or poorly structured context, even within token limits, can degrade output quality. This challenge was especially relevant due to the decision to implement the entire solution locally, without access to commercial API-based models.

To overcome these constraints, the solution relies on a hybrid generation process that combines model-driven input with structured prompt engineering techniques. The system employs dynamic prompt construction, where instructions are derived from formal templates and merged with input artifacts. This enables semantic precision and output consistency, while minimizing hallucinations.

A. Prompt Construction and Input Engineering

Prompt engineering is a critical set of techniques aimed at improving the precision, relevance, and structural integrity of instructions provided to large language models (LLMs). Given the inherent token limits and sensitivity to vague or ambiguous inputs, effective prompt design plays a key role in achieving high-quality, context-aware results.

While most modern LLMs can process both natural language and semi-structured formats, generating well-structured outputs, such as formal documentation artifacts, remains challenging. This is especially true when using general-purpose models that are not fine-tuned for technical or procedural content. To address this, researchers and practitioners have explored strategies such as zero-shot prompting, few-shot prompting, chain-of-thought prompting, and input injection, where contextual cues are embedded into the prompt to guide the generation process more effectively.

An increasingly popular and effective technique, particularly in software engineering and academic contexts, is schema-based prompting. In this method, the input is organized into clearly defined data segments or templates, enabling the model to recognize expected content categories and follow predefined output structures. This improves both the model's contextual understanding and its ability to maintain the desired format over longer outputs [41].

In the proposed solution, prompt and input engineering form the core of the generation workflow. Instruction files are parsed into stepwise key-value pairs and stored in session memory. During generation, the system performs iterative prompting, combining each instruction step with previously generated content.

Fig. 5 presents an example of the JSON-based prompt structure, and Figure 6 shows the resulting generated output.

<Context>
1. I am creating a local generative AI solution using Llama 3.2:3b model together with embedding model and local vector database ChromDB. The envrionment will not have access to the internet.
2. Model purpose is to be a local generative text assistant that handles input data based on provided instructions.
</Context>
<Instructions>
1. The model will receive following inputs interatively as prompt:
   a. fragments from the main document,
   b. instruction describing what to do with the fragments. This instruction changes for each iteration.
   c. Result from previous iterations
2. The model should only use information received as context fragments or previous results as source. Instructions are the task what model should achieve.
3. Take into account that the prompt cannot be too long due to token limitations of the model if used for every interaction.
4. Write me a coprehensive prompt to use as system prompt for Llama 3.2:3b model to set it's main purpose.
</Instructions>
<Examples>
1. You are a document generator llama 3.2:3b that generates answers based on context.
2. Main rule : You only can use context and previous outputs as source of information
</Examples>

Fig 5. A prompt for prompt generation on GPT-4-mini model

Sistēmas ievade
system_prompt = """
You are a document generation assistant local model llama3.2:3b.
You will receive:
- <Context>: You will receive text fragments for context input
- <Instructions>: JSON payload with processing rules under 'llm_prompt_instructions'.
- <Output structure>: JSON payload defining 'output_structure'.
- <User input>: optional additional information.
Tasks:
1. Process <Context> according to <Instructions> JSON.
2. Generate output strictly following <Output structure> JSON keys and hierarchy.
3. If <Context> lacks any required field, insert a placeholder in the form <MISSING_field_name>.
4. Do not add information not present in <Context>:.

Master rule: If you do not find information based on specific <Instructions>:. Leave output empty.
"""

Fig 6. The generation result: a prompt for Llama 3.2:3b model

Additionally, semantic retrieval is supported through a local vector database queried via Retrieval-Augmented Generation (RAG). Relevant context fragments are injected into each prompt, ensuring that the model's output is both semantically grounded and contextually accurate.

So far, the instruction is processed as part of the iterative prompting loop, where the system performs the following steps:

1. Retrieves relevant fragments from the vector database using tags such as "stakeholder_goals" and "deliverables".
2. Constructs a prompt that integrates retrieved content, the instruction goal, and global generation rules (e.g., tone, structure).
3. Executes the prompt using the LLM, producing a scoped and semantically grounded section.

The resulting output typically begins with a structured summary of project boundaries, highlights included and excluded deliverables, and clarifies key stakeholder expectations, which all are grounded in the source context provided by the uploaded artifacts.

This process ensures that generated text is not only grammatically correct and coherent, but also aligned with input data, methodologically consistent, and easy to validate. If any relevant data is missing (e.g., no deliverables listed), the model leaves the section incomplete or explicitly notes the absence, avoiding hallucinated content.

The core task defined for the model was text generation based on structured input instructions. An example of the instructions given for the Project Scope is shown in Fig. 7. Crucially, each iteration updates the prompt dynamically, preserving prior results and refining the context. This strategy has shown to be highly effective in reducing hallucinations and ensuring continuity across output sections. The base system prompt (or system message) is used to define global context and behavior for all interactions with the LLM.

To optimize prompt formulation, OpenAI's GPT-4-turbo model (May 2024 version) was used to assist in designing the system-level prompts for the local LLaMA 3.2:3b model. These instructions define the system's role, generation boundaries, and output expectations based on the runtime environment and technologies used in this project [41]-[42].

These instructions clarified the expected format and explicitly emphasized that the context PDF file is the sole source of truth for content generation. The model was also guided to respect the limitations of the local environment and its own processing capacity, discouraging overly long or ambiguous prompts.

```
{
  "instruction_id": "project_scope_statement",
  "document_type": "Project Scope Statement",
  "description": "Using the project charter as input, build a Project Scope Statement.",
  "input_type": "project_charter",
  "llm_prompt_instructions": {
    "goal": "Create a valid Project Scope Statement JSON, containing only the elements defined below, drawn solely from the input charter.",
    "steps": [
      "1. Project Overview: Extract from the charter a concise statement of project purpose, objectives, and expected outcomes. Insert data into ProjectName, high-level Objectives, SummaryDescription fields.",
      "2. Product Scope Description: Detail what will be built or delivered. List the key features, functions, and performance characteristics of the final product or service as described in the charter.Insert data into ProductScopeDescription fields ",
      "3. Deliverable Breakdown: Enumerate each major deliverable the project will produce. For each deliverable, assign a unique identifier and a brief description. Do not include any tasks, only end-products or results. Insert data into id, description fields ",
      "4. Scope Boundaries: Define what is in scope (specific inclusions) and what is out of scope (specific exclusions). Use two separate lists: one for inclusions and one for exclusions. Insert data into ScopeInclusions, ScopeExclusions fields",
      "5. Acceptance Criteria: For each deliverable, specify the measurable conditions or tests that must be satisfied for formal acceptance. Insert data into deliverableId, criteria fields if data exists. If missing, use `<MISSING_AcceptaneCriteria> `.",
      "6. Assumptions: List all assumptions from the charter (conditions believed to be true). Each assumption must have a unique ID and description. Insert data into id, description fields if data exists. If missing, use `<MISSING_Assumptions>`.",
      "7. Constraints: List all project constraints from the charter (limitations on time, budget, resources, or technology). Each constraint must have a unique ID and description. Insert data into id, description fields if data exists. If missing, use `<MISSING_Constraints>`.",
      "8. Output: Combine all the previous output_result into a single JSON object matching the template structure (fields: ProjectOverview, ProductScopeDescription, Deliverables, ScopeInclusions, ScopeExclusions, AceptanceCriteria, Assumptions, Constraints). Ensure valid JSON and no additional fields."
    ]
  }
}
```

Fig 7. The instructions for Project Scope Statement

As a part of the development process, a system prompt was iteratively refined and formatted using OpenAI tools, resulting in a stable, high-level instruction set. This ensures more predictable and consistent model behavior across varying input scenarios. To reduce hallucinations, a key rule was introduced: if essential information is missing from the context, the model must leave the corresponding field blank rather than fabricate content.

This strict constraint not only improves trust in the AI-generated output but also allows for easier human validation. It encourages transparent failure handling rather than misleading completions, which is critical for professional documentation processes. As a result, the JSON files of the required documentation fragments are generated.

The examples of outputs for project scope, requirements un stakeholder register are shown in Fig. 8-10.

The solution supports structured evaluation of generated documents based on two perspectives. First, document-level criteria assess completeness, terminology, and traceability to source inputs. Second, project-specific context examines content relevance and adaptability based on the detail of provided artifacts. Testing with varied project samples confirmed that output quality improves with richer inputs. Detailed project charters produced precise and tailored documentation, while minimal inputs resulted in more generic, though structurally valid, content. This demonstrates that the system adapts to input complexity while maintaining consistent structure, making it suitable for diverse documentation scenarios across different project environments.

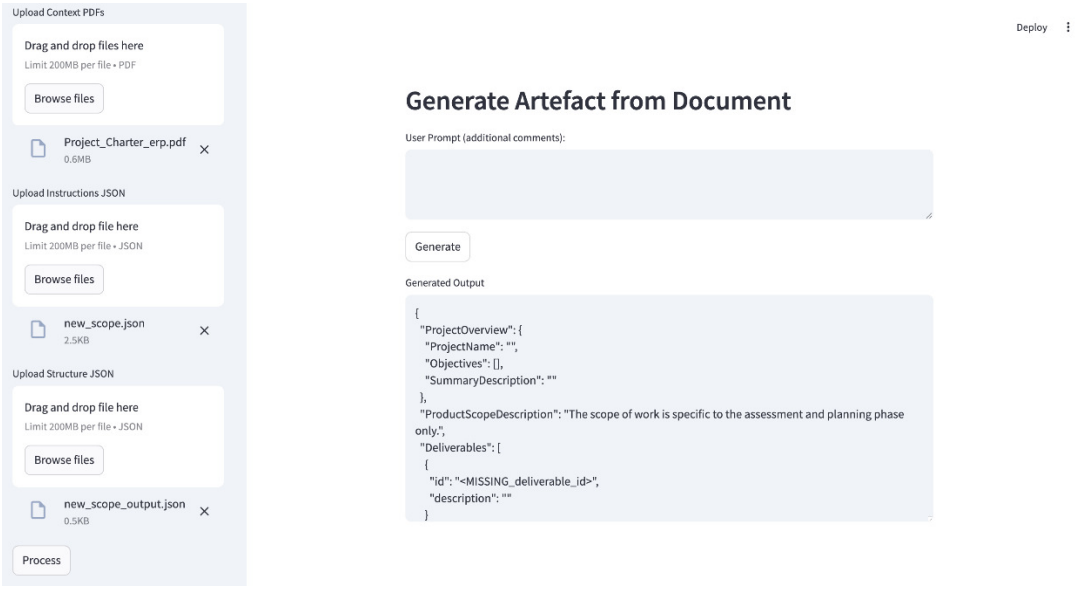


Fig 8.A fragment of output JSON for project scope

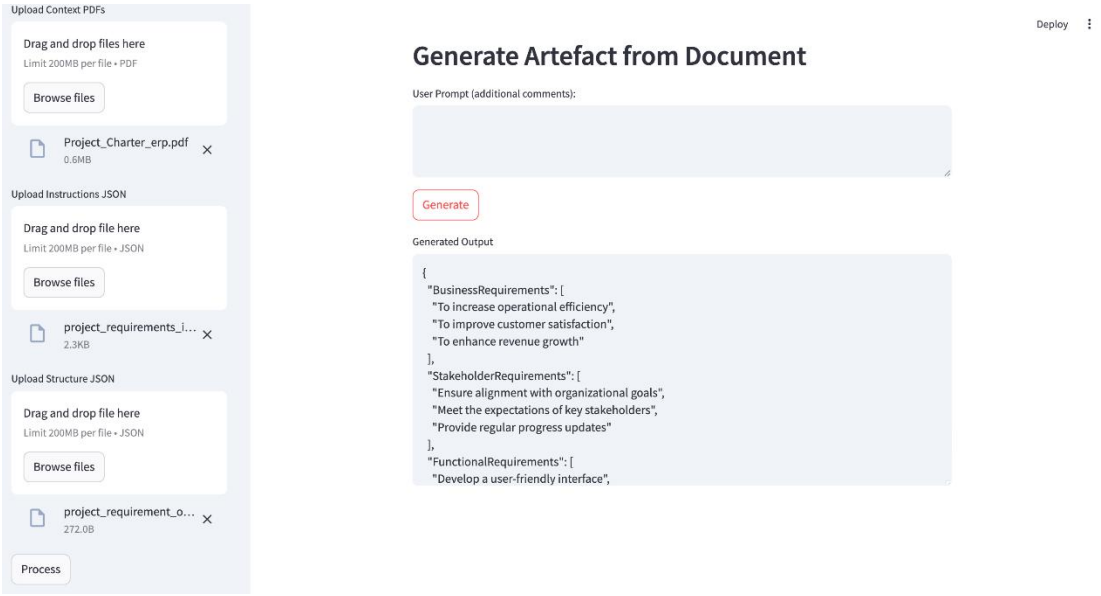


Fig 9. A fragment of output JSON for requirements

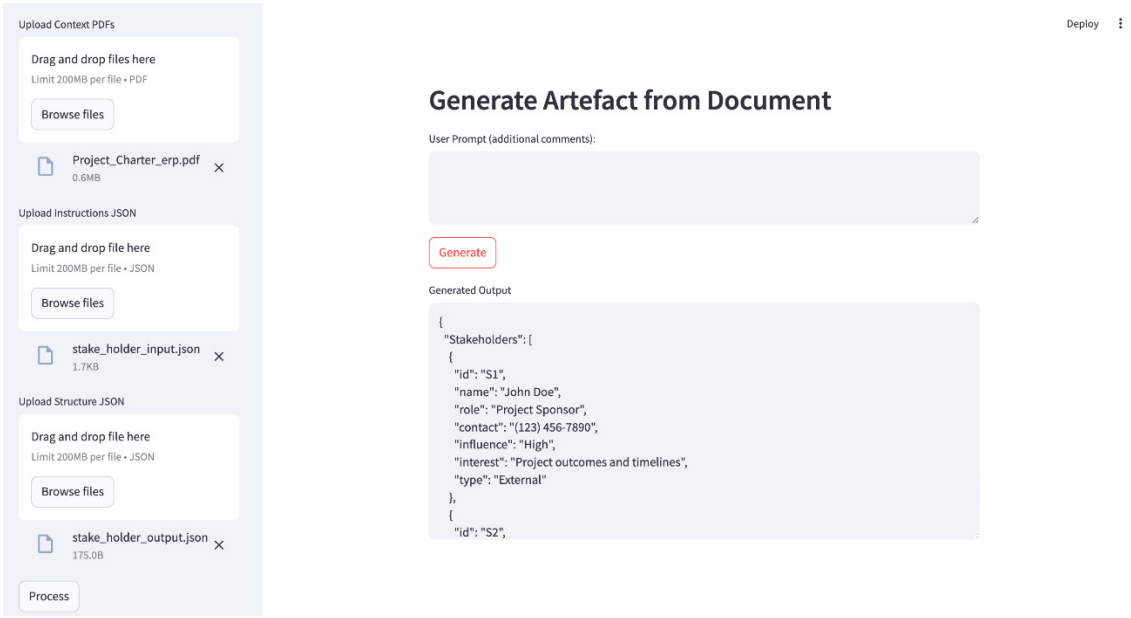


Fig 10. A fragment of output JSON for stakeholders

B. Solution Testing

The proposed approach was evaluated using two large language models (LLMs): LLaMA 3.2:3b and OpenAI GPT-4-mini. Both models received a unified instruction template and identical context limits. To ensure fairness, GPT-4-mini was restricted from using any external tools or accessing the web, relying solely on the provided input. To avoid priming effects, methodologies behind the instruction design were deliberately excluded from the prompt.

Each model processed 200-token segments with 50-token overlaps. For each evaluation step, five context fragments were drawn from the source document to ensure relevant yet limited context. Every instruction tied to an artifact was executed ten times, and outputs were stored for later analysis. To preserve neutrality, LLaMA's vector database was refreshed after each run, while GPT-4-mini was instructed to disregard prior outputs and treat each prompt as new.

Model performance was assessed using a rubric-based method, evaluating outputs across five dimensions. Each dimension was rated on a scale from 0.00 to 1.00, as detailed in Table 1.

TABLE I.
RUBRIC SCORE SCALE

Score	Rubric Criteria
0.00	Output contains incorrect or fabricated data
0.25	Output is mostly empty or contains only fragments
0.50	Output contains at least one correct element per section
0.75	Output aligns with dimensional expectations but lacks details
1.00	Output fully meets dimensional expectations

Each dimension was further weighted according to its significance in the evaluation, with the highest weights assigned to hallucination prevention and instruction compliance, given their criticality in constrained local environments [43]. Output clarity and consistency were assigned lower weights due to their less critical impact. The dimension weights are shown in Table 2.

TABLE II.
EVALUATION DIMENSIONS AND WEIGHTS

Evaluation Dimension	Weight
Data Hallucinations	0.30
Contextual Relevance	0.20
Output Consistency	0.10
Instruction Compliance	0.30
Output Clarity	0.10

A similar dimensional evaluation approach has been observed in empirical research investigating hallucinations across popular large language models. One such study assessed the hallucination rates of various LLMs by processing 1,000 documents, with results reported by [44]. Notably, the LLaMA 3.2:3b model demonstrated a hallucination rate of 7.9% in this evaluation.

In this study, simulation results were aggregated by assigning weights to each evaluation dimension, calculating average scores for each output fragment. The outcomes of the ten repeated iterations per instruction were averaged, and the dimension-specific results are presented in Table 3. Despite architectural and operational differences, that is, LLaMA 3.2:3b running locally and GPT-4-mini hosted via cloud services, both models produced comparable top results.

TABLE III.
EVALUATION DIMENSIONS AND WEIGHTS (1- HALLUCINATIONS, 2 - CONTEXT USE, 3 - CONSISTENCY , 4 - INSTRUCTION COMPLIANCE, 5 – CLARITY)

Artifact	Model	1	2	3	4	5	Overall	Scenario
Scope Statement	LLaMA 3.2:3b	0.7	0.5	0.45	0.6	0.75	0.61	One Washington
	GPT-4-mini	0.9	0.75	0.8	0.8	1.0	0.84	
Requirements Document	LLaMA 3.2:3b	0.6	0.45	0.35	0.5	0.75	0.53	
	GPT-4-mini	0.8	0.75	0.8	0.8	1.0	0.81	
Work Breakdown Structure	LLaMA 3.2:3b	0.5	0.4	0.3	0.35	0.5	0.42	
	GPT-4-mini	0.8	0.75	0.65	0.7	1.0	0.77	
Stakeholder Register	LLaMA 3.2:3b	0.5	0.5	0.3	0.4	0.5	0.45	
	GPT-4-mini	0.7	0.7	0.65	0.7	1.0	0.725	
Scope Statement	LLaMA 3.2:3b	0.5	0.5	0.45	0.4	0.75	0.49	Work Reports
	GPT-4-mini	0.8	0.75	0.8	0.6	1.0	0.75	
Requirements Document	LLaMA 3.2:3b	0.5	0.25	0.35	0.3	0.7	0.4	
	GPT-4-mini	0.7	0.75	0.8	0.6	1.0	0.71	
Work Breakdown Structure	LLaMA 3.2:3b	0.4	0.2	0.3	0.25	0.3	0.3	
	GPT-4-mini	0.7	0.6	0.65	0.5	1.0	0.64	
Stakeholder Register	LLaMA 3.2:3b	0.35	0.35	0.3	0.3	0.4	0.34	
	GPT-4-mini	0.6	0.5	0.65	0.5	1.0	0.59	
Scope Statement	LLaMA 3.2:3b	0.6	0.5	0.45	0.5	0.75	0.55	Average
	GPT-4-mini	0.85	0.75	0.8	0.7	1.0	0.79	
Requirements Document	LLaMA 3.2:3b	0.55	0.35	0.35	0.4	0.725	0.46	
	GPT-4-mini	0.75	0.75	0.8	0.7	1.0	0.76	
Work Breakdown Structure	LLaMA 3.2:3b	0.45	0.3	0.3	0.3	0.4	0.36	
	GPT-4-mini	0.75	0.68	0.65	0.6	1.0	0.7	
Stakeholder Register	LLaMA 3.2:3b	0.43	0.43	0.3	0.35	0.45	0.39	
	GPT-4-mini	0.65	0.6	0.65	0.6	1.0	0.66	

The local model performed well when processing fragments closely matching the prompt, but it struggled with semantically diverse expressions. For instance, in requirement analysis or stakeholder registers, the instructions required extracting detailed information implied in the text. Here, GPT-4-mini showed superior performance in delivering more consistent and contextually aligned outputs. A notable gap was observed between the models' ability to interpret instructions, particularly in the LLaMA model. For example, when processing a scope statement, LLaMA inferred a resource planning system implementation based solely on the vague reference to a "system."

An analysis of the results revealed that both models performed better on the One Washington sample. This document was more structured, contained instruction-aligned headings, and used less academic language. These factors are beneficial especially for the local model.

In contrast, higher hallucination rates were observed when processing the productivity project charter. For instance, the scope description included a false claim, whereas the delivery items correctly referenced the convening of a steering committee. Similarly, the "Out of Scope" section was accurate. However, in the assumptions section expected to list six elements the model included only two, of which only one was correct. This may be attributed to the token limitation, which possibly resulted in incomplete sentence inputs that the model extrapolated upon.

Evaluation of GPT-4-mini's results showed tendencies toward verbosity, as the model often expanded text to produce grammatically complete phrases. While linguistically correct, these expansions altered the intended meaning, thus reducing factual accuracy. For example, among the three output goals, only the first two were valid; the third, referencing real-time reporting, was not present in the source text. Similar inconsistencies and hallucinations were noted across other sections, including deliverables and assumptions.

C. Solution Drawback and Improvements

Another key observation during evaluation was the difference in how the models responded to insufficient or ambiguous information. LLaMA 3.2:3b tended to return empty or partial outputs when it could not confidently locate the requested data. In contrast, GPT-4-mini frequently attempted to complete outputs regardless, even when context was lacking. Often generating fabricated content in direct violation of the provided instructions.

In the context of this study, where data integrity is of high importance, LLaMA's behavior was actually preferable. Empty outputs are easier to verify, as they indicate uncertainty rather than assumed correctness, which reduces the risk

of unintentionally integrating hallucinated data into downstream processes.

The results suggest that smaller, locally hosted models like LLaMA 3.2:3b, when given clearly defined instructions and constrained environments, are capable of producing viable outputs, particularly when a human reviewer is involved in validation. Although GPT-4-mini demonstrated superior generalization abilities, its tendency to confidently fabricate data limited its advantage in this controlled use case.

The test design used small input samples and short text fragments. In several cases, both models failed to retrieve relevant content and instead included false or unrelated claims. LLaMA also demonstrated high variability between runs: in multiple cycles, it returned empty or low-information responses, likely due to how fragments were selected via the nearest-neighbor search over the vector database using Ollama embeddings.

Each processing step queried only five top-matching fragments, assuming semantic accuracy in the embedding model. However, no additional validation of the relevance of retrieved fragments was performed. Furthermore, the model's iterative process continuously appended data to key/value stores without restriction, potentially leading to overwriting or compounding errors in later stages.

In summary, while GPT-4-mini showed superior fluency and semantic approximation, it introduced more hallucinations than expected. LLaMA 3.2:3b, though less consistent, adhered more strictly to constraints and demonstrated a conservative generation strategy more suitable for high-integrity documentation workflows in local AI-assisted systems.

VI. Conclusion

This study demonstrates that the integration of generative AI into IT project documentation processes is not only feasible but also valuable, which is provided by the frame within a model-driven and standards-aligned architecture. By combining the expressive capabilities of large language models with the structural rigor of model-driven development, the proposed solution supports the generation of high-quality, semantically consistent documentation artifacts during the early phases of IT projects.

From a theoretical perspective, this work contributes to the design science research body of knowledge by introducing a novel artifact: a model-based framework that constrains and guides generative AI outputs. It builds upon the principles of semantic control in AI-assisted software engineering, showing that MDD structures can serve as effective scaffolds for LLM behavior. Additionally, it expands on prior work in documentation automation by demonstrating how prompt engineering, semantic traceability, and formal process alignment can mitigate common challenges such as hallucinations and incoherence.

In terms of practical implications, the solution enables project teams, especially in resource-constrained environments, to accelerate and standardize documentation workflows without compromising methodological integrity. The prototype shows that even with local infrastructure, it is possible to harness generative AI tools securely and effectively. Reusable, cross-method instruction sets and alignment with common standards further enhance the applicability and scalability of the approach in varied project contexts.

Several key findings of the research are as follows:

1. Structured inputs and well-defined transformation rules significantly improve the quality and reliability of AI-generated documents.
2. Embedding documentation within a model-driven framework allows for semantic validation, traceability, and reduced ambiguity.
3. Local deployment of LLM-based tools is technically feasible and aligns with data governance requirements in many organizations.
4. Standard elements (e.g., objectives, risks, assumptions) can be reused across methods, enabling cross-framework consistency in documentation practices.

This research opens new avenues for extending the framework to other documentation types and domains (e.g., test plans, architecture overviews, compliance checklists), and for integrating dynamic feedback loops to iteratively refine generated content based on stakeholder input or project evolution.

Future work could explore as follows:

1. Comparative evaluation across multiple LLMs and domain contexts;
2. Deeper integration with enterprise modeling tools or agile project platforms;
3. Automated validation mechanisms for AI-generated artifacts.

In conclusion, the results affirm that generative AI, when properly constrained and guided, can enhance formal documentation practices in IT project management. Rather than replacing human expertise, such tools augment human capabilities, i.e., improving efficiency, consistency, and knowledge transfer in complex project environments.

References

[1] X. Zhang and B. Dorn, "Accelerating software development through agile practices: A case study of a small-scale, time-intensive web development project at a college-level IT competition," *J. Inf. Technol. Educ.*, vol. 11, pp. 1–12, 2012, doi: 10.28945/1545.

[2] H. Kerzner, "Project Management: A Systems Approach to Planning, Scheduling, and Controlling", 13th ed. Hoboken, NJ, USA: Wiley, Mar. 2022, ISBN 978-1119805373.

[3] B. Habib and R. Romli, "A Systematic Mapping Study on Issues and Importance of Documentation in Agile," 2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2021, pp. 198-202, doi: 10.1109/ICSESS52187.2021.9522254

[4] W. Behutiye, P. Seppänen, P. Rodríguez, and M. Oivo, "Documentation of quality requirements in agile software development," in Evaluation and Assessment in Software Engineering (EASE' 20), Trondheim, Norway, Apr. 15–17, 2020, ACM, New York, NY, USA, pp. 1–10, doi: 10.1145/3383219.3383245

[5] A. Ataman, "Data quality in AI: Challenges, importance & best practices," AIMultiple. [Online]. Available: https://research.aimultiple.com/data-quality-ai/

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020, doi: 10.48550/arXiv.2005.14165

[1] P. Liang, R. Bommasani, D. Tsipras, et al., “Holistic evaluation of language models,” arXiv preprint arXiv:2211.09110, 2022, doi: 10.48550/arXiv.2211.09110.

[2] OpenAI, “GPT-4 Technical Report,” OpenAI, Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4

[3] F. Nafz, M. Krajinovic, and M. Ley, “Artificial intelligence in software documentation: Embracing the documentation as code paradigm,” in “Go Where the Bugs Are”, G. Ernst et al., Eds., “Lecture Notes in Computer Science”, vol. 15765, Cham, Switzerland: Springer, 2025, doi: 10.1007/978-3-031-92196-4_14.

[4] K. J. Ajeigbe and O. Emma, “Dynamic Documentation Generation with AI,” 2024. Available at: https://www.researchgate.net/publication/390265865_Dynamic_Documentation_Generation_with_AI

[5] K. Dearstyne, A. Rodriguez, and J. Cleland-Huang, “Supporting Software Maintenance with Dynamically Generated Document Hierarchies,” arXiv preprint arXiv:2408.05829, Aug. 2024, doi: 10.48550/arXiv.2408.05829

[6] S. Mehta, A. Rogers, and T. Gilbert, “Dynamic Documentation for AI Systems,” arXiv preprint arXiv:2303.10854, Mar. 2023, doi: 10.48550/arXiv.2303.10854.

[7] C. Leyh, A. Lorenz, M. J. Faruga, and L. Koller, “Critical Success Factors for ERP Projects Revisited: An Update of Literature Reviews,” in Proc. 19th Conf. on Computer Science and Intelligence Systems (FedCSIS), vol. 39, Annals of Computer Science and Information Systems, pp. 131–140, 2024, doi: 10.15439/2024F6271

[8] T. Vetriselvi, M. Mathur, and M. Bhuvaneswari, “Applying Generative AI to Create SOP, Reducing API Costs Through Prompt Compression and Evaluating LLM Responses with Tonic Validate RAG Metrics,” in Proc. 4th Int. Conf. Ubiquitous Comput. Intell. Inf. Syst. (ICUIS), 2024, doi: 10.1109/ICUIS64676.2024.10867024

[9] S. Tsuchiwata, Y. Tanabe, Y. Hosoya, and K. Aoyama, “Generative AI Driven Clinical Drug Development,” Japanese Journal of Clinical Pharmacology and Therapeutics, 2025, doi: 10.3999/jscpt.56.2_109

[10] A. Barcaui and A. Monat, “Who is better in project planning? Generative artificial intelligence or project managers?,” Project Leadership and Society, vol. 4, 2023. 10.1016/j.plas.2023.100101

[11] A. Bahi, J. Gharib, and Y. Gahi, “Integrating Generative AI for Advancing Agile Software Development and Mitigating Project Management Challenges,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150306

[12] Z. Alliata, T. Singhal, and A.-M. Bozagiu, “The AI Scrum Master: Using Large Language Models (LLMs) to Automate Agile Project Management Tasks,” in Lect. Notes Bus. Inf. Process., 2025, doi: 10.1007/978-3-031-72781-8_12

[13] M. AbuAlfateh, M. Ali, and M. Messaadia, “Establishing Key Performance Indicators for Human-Generative AI Collaboration,” Studies in Big Data, vol. 110, 2025, doi: 10.1007/978-3-031-83911-5_42

[14] V. Bilgram and F. Laarmann, “Accelerating Innovation With Generative AI: AI-Augmented Digital Prototyping and Innovation Methods,” IEEE Eng. Manag. Rev., vol. 51, no. 2, pp. 40–47, 2023, doi: 10.1109/EMR.2023.3272799

[15] N. Ibadildin, Z. Kenzhin, G. Yeshenkulova, and A. Kadyrova, “Artificial Intelligence in Project Management: A Bibliometric Analysis,” Problems and Perspectives in Management, vol. 23, no. 1, pp. 130–142, 2025, doi: 10.21511/ppm.23(2).2025.17

[16] S. Salimimoghadam, A. N. Ghanbaripour, R. J. Tumpa, and M. Skitmore, “The Rise of Artificial Intelligence in Project Management: A Systematic Literature Review of Current Opportunities, Enablers, and Barriers,” Buildings, vol. 15, no. 2, 2025. 10.3390/buildings15071130

[17] O. Nikiforova, J. Grabis, O. Pastor, K. Babris, M. K. Miļūne, and R. Bobkovs, “Model-Based Methodology for Development of IT Project Management Plan and Scope Using Artificial Intelligence: Project in Progress,” in Proc. 19th Int. Conf. on Research Challenges in Information Science (RCIS), CEUR Workshop Proc., 2025 (in press).

[18] J. White, C. Kirchner, T. Paschal, S. Hays, A. Kazerouni, T. Mytkowicz, and M. Monperrus, “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” arXiv preprint arXiv:2302.11382, 2023, doi: 10.48550/arXiv.2302.11382

[19] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” ACM Comput. Surv., vol. 55, no. 9, Art. no. 195, pp. 1–35, Sep. 2023, doi: 10.1145/3560815

[20] O. Nikiforova, K. Babris, U. Karlovs-Karlovskis, M. Narigina, A. Romanovs, A. Jansone, J. Grabis, and O. Pastor, “Model Transformations Used in IT Project Initial Phases: Systematic Literature Review,” Computers, vol. 14, no. 2, p. 40, 2025, doi: 10.3390/computers14020040

[21] O. Nikiforova, K. Babris, M. K. Miļūne, N. Tanguturi, and Ó. Pastor, “Key Artefacts in the Initial Phases of IT Project Management: Systematic Mapping Study,” in Proc. 20th Int. Conf. Evaluation of Novel Approaches to Software Engineering (ENASE), SciTePress, 2025, pp. 773–781, doi: 10.5220/0013471000003928

[22] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[23] M. Brambilla, J. Cabot, and M. Wimmer, Model-Driven Software Engineering in Practice, 2nd ed. Cham, Switzerland: Springer, 2017.

[24] O. Pastor and J. Molina, Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling. Berlin/Heidelberg, Germany: Springer, 2007. [Online]. Available: https://doi.org/10.1007/978-3-540-71868-0

[25] O. Nikiforova, M. K. Miļūne, K. Babris, and O. Pastor, “Generation of IT Project Documentation Elements from a Model Transformation Chain,” in Proc. Int. Conf. on Software Technologies (ICSOFT), 2025, pp. 1–10, DOI: 10.5220/0013568300003964

[26] Agile Manifesto (2001) https://agilemanifesto.org/

[27] ISO/IEC/IEEE, Systems and software engineering — Life cycle processes — Requirements engineering, ISO/IEC/IEEE 29148:2018, International Organization for Standardization, Geneva, Switzerland, 2018

[28] ISO/IEC/IEEE, Systems and software engineering — Life cycle processes — Project management, ISO/IEC/IEEE 16326:2019, International Organization for Standardization, Geneva, Switzerland, 2019

[29] ISO, Project, programme and portfolio management — Guidance on project management, ISO 21502:2020, International Organization for Standardization, Geneva, Switzerland, 2020.

[30] P. Bourque and R. E. Fairley, Eds., Guide to the Software Engineering Body of Knowledge (SWEBOK Guide), Version 4.0, IEEE Computer Society, 2024. [Online]. Available: https://www.computer.org/education/bodies-of-knowledge/software-engineering

[31] Project Management Institute, A Guide to the Project Management Body of Knowledge (PMBOK® Guide), 7th ed., Newtown Square, PA, USA: PMI, 2021

[32] O. Nikiforova, M. Kirikova, and N. Pavlova, “Two-hemisphere driven approach: Application for knowledge modeling,” in Proc. 7th Int. Baltic Conf. Databases and Information Systems, 2006, pp. 244–250, art. no. 1678503. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-34250753483

[33] O. Nikiforova and K. Gusarovs, “Comparison of BrainTool to other UML modeling and model transformation tools”, AIP Conf. Proc., vol. 1863, art. no. 330005, 2017. doi: 10.1063/1.4992503

[34] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan, “Does Prompt Formatting Have Any Impact on LLM Performance?,” arXiv preprint arXiv:2411.10541, 2024, doi: 10.48550/arXiv.2411.10541

[35] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al., “Self-Refine: Iterative Refinement with Self-Feedback,” arXiv preprint arXiv:2303.17651, 2023, doi: 10.48550/arXiv.2303.17651

[36] OpenAI, ChatGPT (May 2024 version) [Large Language Model]. Available: https://chat.openai.com/

[37] T. Bao, J. Yang, Y. Yang, and Y. Yin, "RM2Doc: A tool for automatic generation of requirements documents from requirements models," in Proc. ACM/IEEE 44th Int. Conf. Softw. Eng.: Companion Proc. (ICSE ’22), New York, NY, USA, 2022, pp. 188–192, doi: 10.1145/3510454.3516850.

[38] Vectara. (2024). Hallucination leaderboard [Computer software]. GitHub. https://github.com/vectara/hallucination-leaderboard

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 133–143
DOI: 10.15439/2025F6266
ISSN 2300-5963 ACSIS, Vol. 43

A New Perspective of Associative Memories for Partial Patterns

Soma Dutta
0000-0002-7670-3154
University of Warmia and Mazury in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: soma.dutta@matman.uwm.edu.pl

Jaewon Shin
0009-0005-5284-295X
University of Warmia and Mazury in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
Email: jaewon.shin@student.uwm.edu.pl

***Abstract*—This study aims to present a new perspective for creating associative memories using incomplete data and project them into artificial neural networks to retrieve complete data from incomplete data. In previous studies, various approaches to construct associative memories and retrieval of stored data have been proposed. This paper attempts to pin point some of the limitations observed in the existing approaches and propose a way to get rid of that. In particular, a different perspective of comparing two incomplete patterns is proposed and based on that a flexible way of constructing associative patterns of a given input (partial) pattern is developed. Finally, the respective neural network architecture is proposed following similar construction reported in the existing research.**

I. Introduction

PATTERN recognition is a natural cognitive process that attempts to match an encountered instance of a pattern with the already stored patterns in the memory and thereby retrieve (plausibly) relevant information about the new pattern. Learning from patterns, extracting association among patterns and then applying that knowledge in recognizing useful patterns in different contexts of decision making characterize different important features of human intelligence [1]. The emergence of artificial neural network (ANN) [2] as a prototypical model of human brain, or more specifically biological neural network, is already well known. From its inception ANN has been considered as an effective tool in pattern recognition, signal processing, symbol processing etc [3], [4], [5]. The task of pattern association generally aims at storing different patterns similar to a given input pattern so that a concept related to the given input pattern can be recalled when a new pattern appears to be similar or a member of the stored patterns. Variations of a given input pattern is a common phenomenon, often caused by noise incorporated while collecting data or partial specification of data due to incomplete information. Content-Addressable Memory (CAM) or Associative memory (AM) [5], [6], [7] is the research domain of ANN which concerns about associating similar patterns of a given input data to create a memory and recalling the concepts associated to the input data for those stored patterns and implement the idea in the context of ANN.

In [3] the authors described that "The term associative memory (AM) or content-addressed memory refers to a memory system where recall of a stored pattern is accomplished by providing a noisy or partially specified input pattern.". A mathematical representation of the problem is as follows.

Given a data containing a set of input-output pairs $(\vec{u_1}, \vec{v_1})$, $(\vec{u_2}, \vec{v_2}), \ldots, (\vec{u_k}, \vec{v_k})$ where each input vector $\vec{u_i}$ of dimension n is connected to an output vector $\vec{v_i}$ of dimension m, the task of learning the relation between input-output patterns is realized through standard perceptron model with different activation functions. However, the additional point of associating any pattern from the associative memory of a given input pattern $\vec{u_i}$ with its output $\vec{v_i}$ is done by creating the notion of *associative memory* of each input vector $\vec{u_i}$, denoted as $AM(\vec{u_i})$, and then projecting the association of $(AM(\vec{u_i}), \vec{v_i})$ through a binary mapping module. The $AM(\vec{u_i})$ for a given input pattern $\vec{u_i}$ is created based on the standard notion of Hamming distance by counting number of bits that differ between two binary vectors. Further setting a desired precision level ρ_i, a limit for allowing number of mismatch of bits with $\vec{u_i}$, the $AM_{\rho_i}(\vec{u_i})$ is created by accepting vectors differing from $\vec{u_i}$ within the range of the pre-fixed precision level ρ_i.

On the other hand, a binary mapping is simply a function $f_I : U \longrightarrow V$ from the domain $U = \{\vec{u_i} : 1 \leq i \leq k\}$ to the range set $V = \{\vec{v_i} : 1 \leq i \leq k\}$ such that $f_I(\vec{u_i}) = \vec{v_i}$. In order to make f_I be able to recall $\vec{v_i}$ for any pattern from $AM_{\rho_i}(\vec{u_i})$ it is extended to f_A where the domain $U_A = \cup_{i=1}^{k} AM_{\rho_i}(\vec{u_i})$. So, once through the binary mapping module f_I is extended to f_A by defining $f_A(\vec{u}) = \vec{v_i}$ if $\vec{u} \in AM_{\rho_i}(\vec{u_i})$ the next task leads towards tuning the perceptron model recognizing f_I in a way that it can recognize the extended function f_A. Here, the precision level ρ_i chosen for each input vector $\vec{u_i}$ in creating $AM_{\rho_i}(\vec{u_i})$ plays a crucial role as in order to ensure unambiguous recall of a particular output $\vec{v_i}$ for a noisy input pattern $\vec{u}$ the condition $AM_{\rho_i}(\vec{u_i}) \cap AM_{\rho_j}(\vec{u_j}) = \emptyset$ for any $i \neq j$, $1 \leq i, j \leq k$ is needed [3], [8].

Here a few points to be noticed. One is, by allowing a mismatch in the bits of $\vec{u_i}$ within a specified precision level ρ_i the point of noisy data, as mentioned in the above quote, is addressed. However, the aspect of recalling associative memory in the context of incomplete patterns of a given pattern is not paid a due attention to as according to their proposal two partial vectors only can be comparable using

Topical area: Advanced Artificial Intelligence in Applications

standard Hamming distance when the positions of missing data of both the vectors are the same. More specifically, two partial binary vectors $\langle 1, ?, 1, 0, 1\rangle$ and $\langle 1, 1, ?, 0, 1\rangle$, where ? denotes missing value, cannot be compared following [3]. However, as on the known bits positions $\{1, 4, 5\}$ both the vectors match, potentially they could represent the same pattern. The second issue is related to the condition $AM_{\rho_i}(\vec{u_i}) \cap AM_{\rho_j}(\vec{u_j}) = \emptyset$ for any $i \neq j$, $1 \leq i, j \leq k$ which is imposed in [3] in order to avoid multiple associative recall possibility for a noisy vector. For example following [3], both the above vectors are considered to be partial patterns of $\langle 1, 1, 1, 0, 1\rangle$ which means $\langle 1, ?, 1, 0, 1\rangle \odot \langle 1, 1, 1, 0, 1\rangle$ and $\langle 1, 1, ?, 0, 1\rangle \odot \langle 1, 1, 1, 0, 1\rangle$ hold. So, both can belong to $AM(\langle 1, 1, 1, 0, 1\rangle)$ with precision level 1 although they are not comparable with each other following standard Hamming distance. For reference the readers are referred to Figure 1 considered in [3] where the Hamming distance, denoted as HD, could not be computed for half-pattern. This seems a bit counter-intuitive as if we consider two partial patterns are equally potential to represent some fragments of a full image they might have some common fragments matching with each other. Thus, without having a possibility of comparing how two such partial patterns match to each other may lead to significant information gap. Moreover, the same partial pattern $\langle 1, ?, 1, 0, 1\rangle$ could be considered as a partial pattern of a full image $\langle 1, 0, 1, 0, 1\rangle$ and thus considering the lowest precision level 1, we would not manage to avoid the conflict $\langle 1, ?, 1, 0, 1\rangle \in AM(\langle 1, 1, 1, 0, 1\rangle) \cap AM(\langle 1, 0, 1, 0, 1\rangle)$. In [3], the non-overlap of associative memories of two input vectors is achieved by imposing specific precision levels ρ_i and ρ_j of two input vectors $\vec{u_i}$ and $\vec{u_j}$ in a way that $d_H(\vec{u_i}, \vec{u_j}) > \rho_i + \rho_j$. However, as d_H itself has some limitation in calculating distance between two partial vectors we opt for departure from the idea presented in [3]. The third point is concerning the method of preprocessing incomplete patterns before projecting them to a perceptron model. In [3] the authors did not account for incomplete patterns in the training sample. However, while extending the relation f_I from the input-output pairs $(\vec{u_1}, \vec{v_1})$ …, $(\vec{u_k}, \vec{v_k})$ of the training sample over the sets of associative patterns for the inputs $\vec{u_1}, \ldots, \vec{u_k}$, associative patterns are allowed to be incomplete. But before creating the perceptron model corresponding to f_A, the extension of f_I on incomplete patterns, the missing parts are imputed just by setting them to 0's. As mentioned in several papers [9], [10] setting missing parts of a partial pattern by 0's indicates ignoring the missing part while analyzing the data and thus leads to flaws in classification accuracy.

So, in this paper our aim is to address all these above mentioned conditions by some more general perspectives. Specifically, our first proposal is to define a distance function which can compare any two partial (possibly full) binary vectors resolving the first problem. The second consideration is to generalize the notion of 'partial pattern of a vector', denoted by $\odot$ by a more general notion and thus allowing the precision level for the associative memory of an input vector $\vec{u_i}$ to be selected keeping an eye on the whole data of input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$. In regard to the third point, in our approach, firstly we allow incomplete patterns in the input training sample. Moreover, considering that data collected in a particular context might have a reconciliation possibility within themselves, we propose a way of matching partial input patterns within themselves in order to visualize if they reflect any possibility of reconciling full pattern(s) (see Section VI). The idea comes from the way of reconciling content of a full image from several of its parts as is done in jigsaw puzzles. Projection of the extended function f_A, in our context, in a perceptron model can be realized in the same way as in [3], [11].

The content of the paper is organized as follows. Section II reviews the basic details and mathematical model of associative memory proposed in [3], [8] and discusses about some limitations of the considered approach. Section III introduces a new perspective of measuring distance between two partial vectors following Hamming distance as well as presents a new notion of potential extension of partial Patterns. Moreover, it points out the departure from the similar notions proposed in [3], [8]. Section IV presents the essential conditions and results that a dataset needs to be satisfied for creating associative memory. Section V describes the data processing conditions, methods and iterations for creating associative memories of a given input set of partial vectors keeping the standard condition of avoiding multiple recall [3] in mind. Section VI sketches an idea for retrieving associative memories in the context of perceptron model. Section VII presents some concluding remarks.

II. Review of Related Research

To make the paper self-contained in this section let us present a summary of basic relevant details. As mentioned in Introduction the starting point is a data set containing a set of input-output pairs such as $(\vec{u_1}, \vec{v_1})$, $(\vec{u_2}, \vec{v_2})$, …, $(\vec{u_k}, \vec{v_k})$. Neural memories are generally of two kinds, namely *auto-associative memories* and *hetero-associative memories*. The aim of the auto-associative memories is to reconstruct a pattern from a noisy or partially specified pattern; that is here output vector corresponding to each input vector $\vec{u_i}$ is the same and target is to reconstruct $\vec{u_i}$ from $AM(\vec{u_i})$. On the other hand, hetero-associative memory aims to retrieve information concerning an input vector $\vec{u_i}$ given any instance from its $AM(\vec{u_i})$. The relation between k pairs of input-output vectors of dimension n and m respectively is realized through a 2-layer perceptron having the input layer comprising of n neurons, each of which connected to each of the k neurons lying in the hidden layer, and each of these k neurons are then connected to m output neurons. Clearly, n neurons at the input layer represent n-components of an input vector, say $\vec{u_i}$, m neurons at the output layer represent m components of an output vector, and each of the k neurons at the hidden layer represents k different input-output pairs. For more details the readers are referred to [3], [11].

A. Associative Memory

Now as described in Introduction to extend f_I from U to the sets of associative memories for each input vector of U, first $AM(\vec{u_i})$ is defined in the following manner.

For any $\vec{u_i} \in U$ and respective precision level ρ_i, $AM_{\rho_i}(\vec{u_i})$ $=\{\vec{u}|\vec{u} \in \mathbf{B}^n \& d_H(\vec{u}, \vec{u_i}) \leq \rho_i\}$[1] where $\mathbf{B}^n$ represents the set of all n-dimensional binary vectors and d_H represents the Hamming distance. Then f_I is extended over $\cup_{i=1}^{k} AM_{\rho_i}(\vec{u_i})$ in the following manner.

$f_A : \cup_{i=1}^{k} AM_{\rho_i}(\vec{u_i}) \rightarrow V$, $f_A(\vec{u}) = \vec{v_i}$; if $\vec{u} \in AM_{\rho_i}(\vec{u_i})$, $1 \leq i \leq k$. As mentioned before, the precision level ρ_i for each $\vec{u_i}$ is chosen such that $AM_{\rho_i}(\vec{u_i}) \cap AM_{\rho_j}(\vec{u_j}) = \emptyset$ for any $i \neq j$, $1 \leq i, j \leq k$.

As clearly visible that the above context of associative memory does not represent the possibility for including partial vectors with missing values at some of the components. In [3], [8] the idea of constructing associative memories for partially specified patterns is developed in the context of bipolar vectors where components of a vector may assume values from $\{-1, 0, 1\}$. However, 0 is used only as the marker for missing positions and any partial vector is supposed to assume values from only $\{1, -1\}$. Given any partial input vector $\vec{u_i}$ the missing positions of the vector are first padded with 0 so that the Hamming distance between two vectors can be computed. Then for two $\vec{u}$ and $\vec{v}$ the notion, viz., *$\vec{u}$ is a partial pattern of $\vec{v}$* is defined in the following way.

Definition II.1 ([3], [8])**.** *Given a partial bipolar vector $\vec{u}$, $\vec{u}$ is said to be a partial pattern of $\vec{v}$, denoted as $\vec{u} \odot \vec{v}$, if the available components of $\vec{u}$ are the same as that of the corresponding components of $\vec{v}$.*

The Hamming distance between two partial bipolar patterns is computed using the standard definition as $d_H(\vec{u}, \vec{v})$ is defined with the constraint that the unknown positions of $\vec{u}$ and $\vec{v}$ are the same. As a consequence after padding with 0's only the available positions with 1 and -1 would matter in counting mismatches between $pad0(\vec{u})$ and $pad0(\vec{v})$. Then the notion of associative memory in the context of partial vectors is defined in the following way.

Definition II.2 ([3], [8])**.** *(i) Given any input pattern $\vec{u_i}$, the set of all partial patterns of $\vec{u_i}$ with j-bits is given by* $U_i^j = \{\vec{u} : bits(\vec{u}) = j \ \& \ \vec{u} \odot \vec{u_i}\}$.
(ii) Further given a specified precision level ρ_i for $\vec{u_i}$, the associative memory for $\vec{u_i}$ with precision level ρ_i is defined as $AM_{\rho_i}^j = \{pad0(\vec{u'}) : \exists \vec{u} \in U_i^j \ \& \ d_H(\vec{u'}, \vec{u}) \leq \lfloor j/n \rfloor \times \rho_i\}$.

However, here a few points to be noted. The restrictions imposed above regarding only counting 1, -1 for the bit positions indicates that the purpose could have been served by binary vectors as well. Moreover, two vectors $\langle 1, -1, ?, -1, 1\rangle$ and $\langle 1, ?, 1, -1, 1\rangle$ intuitively may represent the same content as on the known positions they have the same values. As presented in [3], it seems that to claim $\vec{u} \odot \vec{u_i}$ the vector $\vec{u_i}$ must be a full vector. Thus, firstly for an input pattern $\vec{u_i}$ with missing values we cannot create $AM_{\rho_i}^j(\vec{u_i})$. Secondly, considering $\vec{u_i} = \langle 1, -1, 1, -1, 1\rangle$ we see both $\langle 1, -1, ?, -1, 1\rangle \odot \langle 1, -1, 1, -1, 1\rangle$ and $\langle 1, ?, 1, -1, 1\rangle \odot \langle 1, -1, 1, -1, 1\rangle$ hold. Thus, $\langle 1, -1, ?, -1, 1\rangle, \langle 1, ?, 1, -1, 1\rangle \in U_i^4$; but as, $d_H(\langle 1, ?, 1, -1, 1\rangle, \langle 1, -1, ?, -1, 1\rangle)$ is not defined both together cannot be considered in $AM_{\rho_i}^4(\langle 1, -1, 1, -1, 1\rangle)$, the set of associative patterns of $\langle 1, -1, 1, -1, 1\rangle$.

[1]Here to be noted that in [3] for this notion a different notation, namely $U_i^n(\rho_i)$ is used.

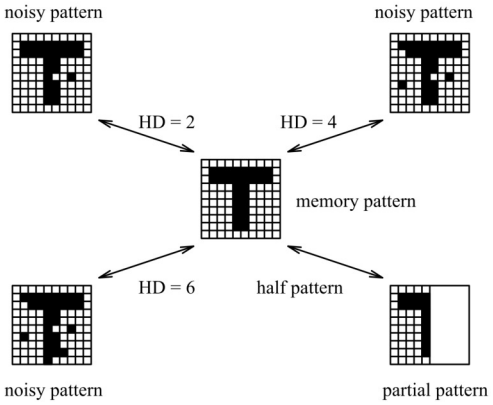


Figure 1. Computing Hamming distance(HD) for noisy patterns [3]

Keeping these limitations in mind our proposal, presented in the following sections, departs from a few aspects.

III. Hamming Distance and Potential Extensions of Partial Patterns

The aim of this section is to develop step-by-step the basic notions, departing from the respective notions proposed in [3], [8], so that the limitations discussed in Section II can be overcome. Our first target is to modify the definition for computing distance between two partial patterns in a way that instead of ignoring it would account for different unknown parts of the input patterns. In this regard, let us first present some notational details below.

Let $\vec{u}$ be any n-dimensional partial binary vector with some unknown components, denoted as $Unp(\vec{u})$; the set of known positions of $\vec{u}$ is denoted by $Bits(\vec{u})$, and $|Bits(\vec{u})|$ determines the number of bits in the vector, alternatively denoted by $bits(\vec{u})$. For example, for $\vec{u} = \langle ?, 1, 0, ?, 1, 0\rangle$ the $bits(\vec{u}) = 4$, $Bits(\vec{u}) = \{2, 3, 5, 6\}$, and $Unp(\vec{u}) = \{1, 4\}$.

A. Hamming distance for partial patterns

Let $\vec{u}$ and $\vec{u'}$ be two n-dimensional partial vectors having (possibly) some unknown components, denoted by $Unp(\vec{u})$ and $Unp(\vec{u'})$ respectively.

Definition III.1. *Given two partial binary vectors $\vec{u}$ and $\vec{u'}$ the Hamming distance between the vectors is defined as $D_H(\vec{u}, \vec{u'}) = d_H(\vec{u^t}, \vec{u^{t'}}) + l$, where $\vec{u^t}$ and $\vec{u^{t'}}$ are obtained by dropping all positions of $Unp(\vec{u}) \cup Unp(\vec{u'})$ from $\vec{u}$ and $\vec{u'}$ respectively, and $l = |Unp(\vec{u}) \cup Unp(\vec{u'})|$.*

It is clear from Definition III.1 that the modified notion of Hamming distance between two partial patterns considers a

pessimistic perspective by admitting the possibility that the missing positions of the vectors might have different content. The results below present in what sense it could be considered as a metric.

Proposition III.2. $D_H(\vec{u},\vec{v}) \leq D_H(\vec{u},\vec{w}) + D_H(\vec{w},\vec{v})$ *for any* $\vec{u},\vec{v},\vec{w}$

Based on Definition III.1, the proof relies on meticulously checking all different possibilities of unknown positions and their interrelations for three vectors $\vec{u},\vec{v},\vec{w}$.

Proposition III.3. D_H *satisfies the following properties of a metric.*

1) $D_H(\vec{u},\vec{v}) \geq 0$ *for any* $\vec{u},\vec{v}$.
2) $D_H(\vec{u},\vec{v}) = k$ *if and only if* $\vec{u^t} = \vec{v^t}$, $|Unp(\vec{u}) \cup Unp(\vec{v})| = k$.
3) $D_H(\vec{u},\vec{v}) = D_H(\vec{v},\vec{u})$ *for any* $\vec{u},\vec{v}$.

The proof of Proposition III.3 is straightforward from Definition III.1. From the property (2) of Proposition III.3, we can notice that $D_H(\vec{u},\vec{u}) = k$ given $|Unp(\vec{u})| = k > 0$. So, in contrary to standard metrics D_H is not reflexive. However, in the context of partial patterns it is quite intuitive that two partial patterns even may look the same on the known parts but there might be some mismatch in content in the undisclosed parts. Moreover, if we stick to the condition that $Unp(\vec{u}) = Unp(\vec{u'})$, then d_H can be considered as a lower estimation of D_H, and thus provides a more general perspective of calculating distance between two partial patterns.

Proposition III.4. *Given* $Unp(\vec{u}) = Unp(\vec{u'})$, $D_H(\vec{u},\vec{u'}) \geq d_H(\vec{u},\vec{u'}) + l$ *where* $|Unp(\vec{u})| = l$.

From Proposition III.4 it can be noticed that as the number of bits of $\vec{u}$ and $\vec{u'}$ increase D_H will be more close to the estimation proposed by d_H.

B. Potential extension for partial patterns

In this section we present a new perspective for the notion associative patterns by considering a modification of the same notion proposed in [3] from two aspects. Firstly, unlike $\odot$ we would introduce a new predicate $\otimes$ which incorporate the possibility of $\vec{u} \otimes \vec{v}$ to hold for a partial vector $\vec{v}$. Secondly, the notion of associative patterns developed based on $\otimes$ would allow two partial patterns $\vec{u}$ and $\vec{u'}$ as associative patterns of a given input pattern $\vec{u_i}$ even they do not share the same unknown parts.

Definition III.5. *Given a partial binary vectors* $\vec{u}$ *and any (partial) binary vector* $\vec{v}$, $\vec{u}$ *is said to be potentially extended to* $\vec{v}$, *denoted by* $\vec{u} \otimes \vec{v}$, *if* $Unp(\vec{u}) \supseteq Unp(\vec{v})$ *and values of all the known components of* $\vec{u}$ *match to that of the respective known components of* $\vec{v}$.

The Theorem below shows that $\otimes$ generalizes the notion of $\odot$.

Theorem III.6. *For a full vector* $\vec{v}$, $\vec{u} \odot \vec{v}$ *is a special case of* $\vec{u} \otimes \vec{v}$.

Proof. Let consider n-dimensional vectors $\vec{u},\vec{v},$. Let $\vec{u} \otimes \vec{v}$ where $Unp(\vec{v}) = \emptyset$ (The condition $Unp(\vec{u}) \supseteq Unp(\vec{v})$ for having $\vec{u} \otimes \vec{v}$ is trivially satisfied.) Now as $\vec{u} \otimes \vec{v}$, $d_H(\vec{u^t},\vec{v^t}) = 0$, where $\vec{u^t},\vec{v^t}$ are the vectors obtained by truncating $Unp(\vec{u})$ from both the vectors. Thus $\vec{v}$ is a full vector and $\vec{u}$ is a partial vector, s.t. $\forall l \in Bits(\vec{u})$, the l-th components of $\vec{u}$ and $\vec{v}$ are the same. $\therefore \vec{u} \odot \vec{v}$ □

Now, analogous to $U^j_{\rho_i}$ presented in Section II in our context for an input vector $\vec{u_i}$ we construct the following set.
$U_i^{bits(\vec{u_i})} = \{\vec{u} \in \mathbf{B}^n | 0 < bits(\vec{u}), \vec{u} \otimes \vec{u_i}\}, 1 \leq i \leq k$.
Considering $|Unp(\vec{u_i})| = p_i$ we can write $U_i^{bits(\vec{u_i})} = U_i^{n-p_i}$.

It is to be worth noticing that in $U_i^{n-p_i}$ we do not allow vectors with full unknown parts or in other words a vector with empty Bits.

IV. Data restriction in choosing precision level for associative memory

In [3], while developing binary mapping and respectively perceptron model for a set of input vectors the authors maintained the constraint of mutually disjoint associative memories for a set of input vectors by tuning the precision level ρ_i for the associative memory constructed for each input vector $\vec{u_i}$. In our work, as the notion of $U^j_{\rho_i}$ is replaced by the notion of $U_i^{n-p_i}$ where $|Unp(\vec{u_i})| = p_i$ to ensure mutually non-overlapping associative patterns for a given input set of vectors we propose to consider non-overlapping unknown components for each pair of vectors $\vec{u_i}$ and $\vec{u_j}$ in the input set of vectors.

Apart from the technical point of view of imposing $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$ for any $i \neq j$ and $1 \leq i,j \leq k$ there is an intuitive perspective of imposing this constraint as well. Often while solving a puzzle related to creating a known image, nowadays prevalently known as jigsaw puzzle[2], we are given a number of pieces each having different known fragments of the image visible. Usually, we aim to fit the unknown and known fragments of two pair of pieces in a way that some part of the whole image can be retrieved. For instance, given two partial input vectors $\vec{u_i} = \langle 1,0,?,0,1\rangle$ and $\vec{u_j} = \langle 1,?,1,?,1\rangle$, in one hand, merging them by replacing one's unknown components with others respective known components may help to visualize a complete pattern $\langle 1,0,1,0,1\rangle$. On the other hand, following the notion of $\otimes$ neither $\vec{u_i} \otimes \vec{u_j}$ nor $\vec{u_j} \otimes \vec{u_i}$ holds. That is, all the input vectors are considered to be independently indispensable to recover the full content of the actual image as neither of them can be potentially extended to the other following the notion of $\otimes$. Moreover, it would be clear in the latter sequels that the impact of this condition is not limited to the above discussed perspectives only.

A. Imposition of Restriction on Data

We start with imposing a few constraint on a data set of n-dimensional input vectors $\vec{u_1},\vec{u_2},\ldots,\vec{u_k}$.

[2]https://en.wikipedia.org/wiki/Jigsaw_puzzle

Definition IV.1 (Dset)**.** *A set of n-dimensional input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ satisfying the following conditions is called a Dset.*

1) $0 \leq |Unp(\vec{u_i})| \leq \lfloor n/2 \rfloor$
2) $Unp(\vec{u_i}) \cup Bits(\vec{u_i}) = \{1, 2, \ldots, n\}$
3) $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$ *for* $1 \leq i, j \leq k, i \neq j$

The condition (1) imposes a restriction, may be called as Maximum Half Uncovered Pattern (MHUP). The condition (2) imposes that unknown and known part of each vector must exhaust the full pattern, may be called Exhaustive Pattern (EP). The condition (3) may be named as Mutually Disjoint Unknown Parts (MDUP).

A series of results below would clarify how the condition (MDUP) may determine the choice of precision level for a given input vector in the context of creating associative memory.

Proposition IV.2. *If $\vec{u} \in U_i^{n-p_i} \cap U_j^{n-p_j}$ then $Unp(\vec{u}) \supseteq Unp(\vec{u_i}) \cup Unp(\vec{u_j})$ and $d_H(\vec{u_i^t}, \vec{u_j^t}) = 0$ where $\vec{u_i^t}$ and $\vec{u_j^t}$ are obtained by dropping all positions of $Unp(\vec{u})$ from $\vec{u_i}$ and $\vec{u_j}$ respectively.*

Proof. $\vec{u} \in U_i^{n-p_i}$ and $\vec{u} \in U_j^{n-p_j}$; that is, $\vec{u} \otimes \vec{u_i}$ and $\vec{u} \otimes \vec{u_j}$. Let $\vec{u_i^t}$ and $\vec{u^t}$ be the vectors obtained by dropping the positions of $Unp(\vec{u})$ from both $\vec{u_i}$ and $\vec{u}$. So $Unp(\vec{u_i}) \subseteq Unp(\vec{u})$ and $d_H(\vec{u_i^t}, \vec{u^t}) = 0$ as well as $Unp(\vec{u_j}) \subseteq Unp(\vec{u})$ and $d_H(\vec{u_j^t}, \vec{u^t}) = 0$. □

The above proposition leads to the following corollary.

Corollary IV.3. *If $\vec{u} \in U_i^{n-p_i} \cap U_j^{n-p_j}$, then $p_i + p_j \leq p$.*

Proof. As $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$, then by Proposition IV.2, $Unp(\vec{u_i}) \cup Unp(\vec{u_j}) \subseteq Unp(\vec{u})$ which implies $p_i + p_j \leq p$. □

Combining Proposition IV.2 and Corollary IV.3 using contraposition we can say that if for a vector $\vec{u}$ with $|Unp(\vec{u})| = p$, $p < p_i + p_j$ or $d_H(\vec{u_i^t}, \vec{u_j^t}) \neq 0$, then $\vec{u} \notin U_i^{n-p_i} \cap U_j^{n-p_j}$, where $\vec{u_i}^t$, $\vec{u_j}^t$ are obtained from $\vec{u_i}$, $\vec{u_j}$ respectively by truncating respective positions of $Unp(\vec{u})$.

So, from Corollary IV.3 it is clear that if two input vectors are not having common unknown parts then that suffices to choose a precision level for a vector $\vec{u}$ and ensure that $\vec{u}$ cannot be regarded as a common associative pattern for both the input vectors. However, this is not a necessary condition ensuring $U_i^{n-p_i} \cap U_j^{n-p_j} = \emptyset$. The sequel below will throw some light in this regard.

Proposition IV.4. *If $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$, then $U_i^{n-p_i} \neq U_j^{n-p_j}$ given $p_i \geq 2$ or $p_j \geq 2$.*

Proof. Let $Unp(\vec{u_i}) = \{r_1, r_2, \ldots, r_{p_i}\}$, $Unp(\vec{u_j}) = \{t_1, t_2, \ldots, t_{p_j}\}$ where $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$ and $p_j \geq 2$. Now, consider $\vec{u}$ with $Unp(\vec{u}) = \{r_1, r_2, \ldots, r_{p_i}\} \cup \{t_l\}$ where $t_l \in Unp(\vec{u_j})$ and $D_H(\vec{u_i}^t, \vec{u}^t) = 0$ where $\vec{u_i}^t$, $\vec{u}^t$ are obtained by truncating all the positions of $Unp(\vec{u})$ from both the vectors. Clearly, $\vec{u} \otimes \vec{u_i}$ and thus $\vec{u} \in U_i^{n-p_i}$. But as $Unp(\vec{u}) \not\supseteq Unp(\vec{u_j})$, $\vec{u} \notin U_j^{n-p_j}$.

In the similar fashion, for $p_i \geq 2$ we can create a vector $\vec{u'}$ such that $\vec{u'} \in U_j^{n-p_j}$ but $\vec{u'} \notin U_i^{n-p_i}$. Thus, we can claim that $U_i^{n-p_i} \neq U_j^{n-p_j}$. □

Proposition IV.4 indicates that by considering non-overlapping unknown parts in the input data and putting the restriction that at most one input vector in the input data may have exactly one unknown component, we will be able to ensure that the set of associative patterns for each input vector will be distinct. However, the condition does not ensure that $U_i^{n-p_i} \cap U_j^{n-p_j} \neq \emptyset$. The example below illustrates the claim.

Example IV.1. *Let us consider $Unp(\vec{u_i})$ and $Unp(\vec{u_j})$ as the same as in Proposition IV.4 with an additional condition that $Unp(\vec{u_i}) \cup Unp(\vec{u_j}) \neq \{1, 2, \ldots, n\}$. Now consider a vector $\vec{u}$ with $Unp(\vec{u}) \supseteq Unp(\vec{u_i}) \cup Unp(\vec{u_j})$ and $D_H(\vec{u}^t, \vec{u_i}^t) = D_H(\vec{u}^t, \vec{u_j}^t) = 0$ where $\vec{u}^t, \vec{u_i}^t, \vec{u_j}^t$ are obtained respectively from the original vectors after removing the components of $Unp(\vec{u})$. This clearly indicates that $\vec{u} \otimes \vec{u_i}$ and $\vec{u} \otimes \vec{u_j}$ and thus $\vec{u} \in U_i^{n-p_i} \cap U_j^{n-p_j}$.*

In [3], the associative patterns for the set of input vectors are imposed to be non-overlapping. The reason behind such consideration is already explained in Section II. The above results indicate that, in our context, in order to avoid overlap in the respective associative patterns for the input vectors we may need a further pre-processing phase for data.

Based on Corollary IV.3, we can notice that given any two input vectors $\vec{u_i}$ and $\vec{u_j}$ if we choose a vector $\vec{u}$ with $Unp(\vec{u}) = p$ such that $p < p_i + p_j$ then $\vec{u} \notin U_i^{n-p_i} \cap U_j^{n-p_j}$. That is, to avoid overlap in the respective associative patterns of any two input vectors of a data set $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ we can impose the following condition.

Condition IV.5. *Given a data set of input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$, for any $\vec{u_i}$, $1 \leq i \leq k$, the precision level ρ_i for the set of associative patterns of $\vec{u_i}$, denoted by $AM(\vec{u_i})$, must satisfy $p_i < \rho_i < p_i + p_j$ for any $j \neq i$.*

Here a natural question arises that whether given an input data we will always have a situation to choose a precision level for an arbitrary input vector following Condition IV.5. Following proposition would clarify the issue further.

Proposition IV.6. *Given the input data $\vec{u_1}, \vec{u_2}, ..., \vec{u_k}$ if for a vector $\vec{u_i}$, $1 \leq i \leq k$, there is no $\rho_i \in \mathbb{Z}^+$ s.t. $p_i < \rho_i < p_i + p_j$ for any $i \neq j$, then there is at least one $j \neq i$ for which $p_j = 1$.*

Proof. From the assumption that there is no $\rho_i \in \mathbb{Z}^+$, s.t. $p_i < \rho_i < p_i + p_j$ for all $i \neq j$ it is clear that $p_i + p_j$ must be equal to $p_i + 1$ for some $j \therefore p_j = 1$ for some j. □

Proposition IV.7. *If $p_j = 1$ for some $j \neq i$, then there is no $\rho_i \in \mathbb{Z}^+$ satisfying $p_i < \rho_i < p_i + p_j$ for $i \neq j$.*

Proof. Let $p_j = 1$ for some $j \neq i$, and there exist a $\rho_i \in \mathbb{Z}^+$ satisfying $p_i < \rho_i < p_i + p_j$ for $i \neq j$. As $p_j = 1$ for some $j \neq i$, we can write $p_i < \rho_i < p_i + 1$ and $0 < \rho_i - p_i < 1$. This is a contradiction as $p_i, \rho_i \in \mathbb{Z}^+$. Hence there is no $\rho_i \in \mathbb{Z}^+$ satisfying $p_i < \rho_i < p_i + p_j$ for $i \neq j$. □

Clearly, if we have a data set of input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ such that $|Unp(\vec{u_i})| = p_i > 1$ for each $i \in \{1, 2, \ldots, k\}$, then following Condition IV.5 finding a precision level $\rho_i \in \mathbb{Z}^+$ for each input vector $\vec{u_i}$ is possible. This result also goes in accordance with our observation from Proposition IV.4. So, allowing only input vectors with more than one unknown components may help to create a unique precision level for the set of associative patterns for each input vectors following Condition IV.5. However, we need to dig further to check how such a condition would impact on the availability of sufficient amount of input data.

B. Impact of data restriction on availability of data

Given the restriction on data imposed in Definition IV.1 and the additional restriction mentioned in Condition IV.5 it is now important to check how much data we may have to train a required perceptron model. In this regard, let us first consider the context of a data set satisfying the conditions (MHUP) $bits(\vec{u_i}) \geq \lceil n/2 \rceil$ and (EP) ($Bits(\vec{u_i})$, $Unp(\vec{u_i})$) forms a partition of $\{1, 2, \ldots n\}$, mentioned in Definition IV.1. Before presenting the result let us define some useful notations.

Notation IV.8. (i) $S_n(bits) = \{X \subseteq \{1, 2, ..., n\} : |X| \geq \lceil n/2 \rceil\}$.
(ii) $S^i(Y) = \{X \subseteq Y : 0 \leq |X| < i\}$

Given the above notations we can notice that $S_n(bits) = P(\{1, 2, ..., n\}) - S^{\lceil n/2 \rceil}(\{1, 2, ..., n\})$.

Proposition IV.9. *$|S_n(bits)| = 2^n - \sum_{i=0}^{m-1} {}^nC_i$, where $n = 2m$ or $n = 2m - 1, m \geq 1$.*[3]

Proof. We know $|P(\{1, 2, ..., n\})| = 2^n$. Now for $S^{\lceil n/2 \rceil}(\{1, 2, ..., n\})$, we have two cases (1) $n = 2m$ and (2) $n = 2m - 1$
(1) If $n = 2m, \lceil n/2 \rceil = m$, then $S^m(\{1, 2, ..., n\}) = \sum_{i=0}^{m-1} {}^{2m}C_i$
(2) If $n = 2m - 1, \lceil n/2 \rceil = m$, then $S^m(\{1, 2, ..., n\}) = \sum_{i=0}^{m-1} {}^{2m-1}C_i$
Therefore, clearly from (1) and (2), $|P(\{1, 2, ..., n\})| - |S^{\lceil n/2 \rceil}(\{1, 2, ..., n\})| = 2^n - \sum_{i=0}^{m-1} {}^nC_i$, where $n = 2m$ or $n = 2m - 1$, $m \geq 1$. □

Proposition IV.10. *$2^n - \sum_{i=0}^{m-1} {}^nC_i = \sum_{i=0}^{\lfloor n/2 \rfloor} {}^nC_i$, where $n = 2m$ or $n = 2m - 1$, $m \geq 1$*

Proof. It is well known $2^n = \sum_{i=0}^{n} {}^nC_i$. Now we need to show $\sum_{i=0}^{n} {}^nC_i = \sum_{i=0}^{m-1} {}^nC_i + \sum_{i=0}^{\lfloor n/2 \rfloor} {}^nC_i$
(1) when $n = 2m, \lceil n/2 \rceil = m$ and $\lfloor n/2 \rfloor = m$
LHS = ${}^{2m}C_0 + {}^{2m}C_1 + {}^{2m}C_2 + ... + {}^{2m}C_{2m}$
RHS[4] = ${}^{2m}C_0 + {}^{2m}C_1 + ... + {}^{2m}C_{m-2} + {}^{2m}C_{m-1} + {}^{2m}C_0 + {}^{2m}C_1 + ... + {}^{2m}C_{m-1} + {}^{2m}C_m$
$= {}^{2m}C_0 + {}^{2m}C_1 + ... + {}^{2m}C_{m-2} + {}^{2m}C_{m-1} + {}^{2m}C_m + {}^{2m}C_{m+1} + ... + {}^{2m}C_{2m-1} + {}^{2m}C_{2m}$
$\therefore$ LHS=RHS

(2) when $n = 2m - 1, \lceil n/2 \rceil = m$ and $\lfloor n/2 \rfloor = m - 1$
LHS = ${}^{2m-1}C_0 + {}^{2m-1}C_1 + {}^{2m-1}C_2 + ... + {}^{2m-1}C_{2m-1}$
RHS = ${}^{2m-1}C_0 + {}^{2m-1}C_1 + ... + {}^{2m-1}C_{m-2} + {}^{2m-1}C_{m-1} + {}^{2m-1}C_0 + {}^{2m-1}C_1 + ... + {}^{2m-1}C_{m-1} + {}^{2m-1}C_m$
$= {}^{2m-1}C_0 + {}^{2m-1}C_1 + ... + {}^{2m-1}C_{m-2} + {}^{2m-1}C_{m-1} + {}^{2m-1}C_m + {}^{2m-1}C_{m+1} + ... + {}^{2m-1}C_{2m-2} + {}^{2m-1}C_{2m-1}$
$\therefore$ LHS=RHS

□

So, combining Propositions IV.9, IV.10 we can say that $|S_n(bits)| = \sum_{i=0}^{\lceil n/2 \rfloor} {}^nC_i$. That is, for 10-dimensional vectors there can be maximum 638 vectors satisfying (MHUP) and (EP) mentioned in Definition IV.1. So, the next question arises that how many vectors out of $|S_n(bits)|$ number of possible vectors can have non-overlapping unknown parts. In this regard, let us consider any input vector $\vec{u_i}$ with $bits(\vec{u_i}) = r$ where $\lceil n/2 \rceil \leq r \leq n$.

Proposition IV.11. *Given an input vector $\vec{u_i}$ with $bits(\vec{u_i}) = r$, where $\lceil n/2 \rceil \leq r \leq n$, the maximum number of vectors in $\boldsymbol{B}^n$ having disjoint unknown parts with $Unp(\vec{u_i})$ is ${}^rC_{n-r}$.*

Proof. Let $\vec{u_i}$ be a vector with $Bits(\vec{u_i}) = \{s_1, s_2, \ldots, s_r\}$ where $\lceil n/2 \rceil \leq r \leq n$ and $|Unp(\vec{u_i})| = n - r$. Following our condition $r \geq n - r$. Then in order to have other r-bits vectors with non-overlapping unknown components we can consider a vector $\vec{u_i}'$ whose $Unp(\vec{u_i}')$ is obtained by exchanging all $n-r$ positions of $Unp(\vec{u_i})$ with positions from $\{s_1, s_2, \ldots, s_r\}$ of $\vec{u_i}$. This allows us to create ${}^rC_{n-r}$ number of different vectors with r-bits having disjoint unknown parts with $\vec{u_i}$. □

Example IV.2. *Following Proposition IV.11, for $n = 10$, we can allow input vectors with number of bits ranging from 5 to 10. Given an input vector $\vec{u_1}$ with $|Unp(\vec{u_1})| = 5$ there can be exactly 1 possible 5-bit vector, say $\vec{u_{11}}$, having non-overlapping unknown parts with $\vec{u_1}$. For an input vector $\vec{u_2}$ with $|Unp(\vec{u_2})| = 4$ there can be exactly 15 possible 6-bit vectors having non-overlapping unknown parts with $\vec{u_2}$. However, these 15 vectors may have common unknown parts within themselves. Thus, at each turn we can select exactly one vector, say $\vec{u_{21}}$, from these 15 possible cases satisfying the condition of non-overlapping unknown components with $\vec{u_2}$. For an input vector $\vec{u_3}$ with $|Unp(\vec{u_3})| = 3$ there can be exactly 35 possible 7-bit vectors having non-overlapping unknown parts with $\vec{u_3}$. So, in one round of selection out of 35 possibilities we can select a pair of vectors, say $\vec{u_{31}}, \vec{u_{32}}$, such that $\vec{u_3}$ and these two vectors satisfy mutually disjoint unknown parts. For an input vector $\vec{u_4}$ with $|Unp(\vec{u_4})| = 2$ there can be exactly 28 possible 8-bit vectors having non-overlapping unknown parts with $\vec{u_4}$. In this case, out of 28 possibilities we can select 4 vectors, say $\vec{u_{41}}, \vec{u_{42}}, \vec{u_{43}}, \vec{u_{44}}$ such that along with $\vec{u_4}$ they satisfy mutually disjoint unknown parts. For an input vector $\vec{u_5}$ with $|Unp(\vec{u_5})| = 1$ there can be exactly 9 possible 9-bit vectors having non-overlapping unknown parts with $\vec{u_5}$ and all of them can be considered together in one round of selection as they satisfy mutually disjoint unknown parts.*

[3] $n = 2m$ or $n = 2m + 1$, is also possible.
[4] LHS and RHS are respectively the abbreviations for Left Hand Side and Right Hand Side of a relational equation.

From the above clarification, two important aspects are surfacing out. One is the chosen number of unknown components for the vectors of a particular dimension. If half of the components are unknown the number of different vectors with non-overlapping unknown parts is very less, and when the number of unknown components is decreasing possible number of different vectors satisfying (MHUP) and (EP) is increasing. However, as it is visible from the example that there seems to be a threshold value for number of unknown components, which is here 3, after which the number of possible vectors with a specified number of unknown components is decreasing. On the other hand, even for a vector $\vec{u_i}$ with $n - r$ ($\leq \lfloor n/2 \rfloor$) number of unknown components not all of the ${}^{r}C_{n-r}$ number of possible vectors having non-overlapping unknown parts with $\vec{u_i}$ can be considered together as they would not satisfy (MDUP). So, in order to train a perceptron model if we wish to use the available data as much as possible satisfying the conditions of a Dset, we may need to split the data over several iterations and learn separately the perceptron model generated in each iteration.

V. Constructing Associative Memory and Splitting Data Over Iterations

Given the above analysis about the number of possible input vectors of dimension n satisfying the conditions of Definition IV.1 and the condition concerning precision level of each input vector (see Proposition IV.6) let us introduce the following method ensuring non-overlapping sets of associative patterns for a given Dset.

Definition V.1 (M-AM)**.** *Let a Dset containing k number of n-dimensional input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ be given. The method, called as M-AM, of creating the associative memory for each input vector is described as follows.*

1) *For each i, if the precision level ρ_i is s.t. $p_i < \rho_i < p_i + p_j$ for any $j \neq i$ holds, then consider $U_{\rho_i} = \{\vec{u} \in U_i^{n-p_i} : |Unp(\vec{u})| \leq \rho_i\}$ as $AM(\vec{u_i})$.*
2) *If for a vector $\vec{u_i}$ there is no ρ_i satisfying Condition IV.5, then follow the following process of modifying U_{ρ_i}.*
   a) *Find $O_i = \{l \mid p_i + p_l = min_j\ p_i + p_j,\ j \neq i\}$*
   b) *Consider $\rho_i = p_i + p_l$*
   c) *$U_{\rho_i}^{mod} = \{\vec{u} \in U_{\rho_i} : Unp(\vec{u}) \not\supseteq Unp(\vec{u_l}),\ l \in O_i\}$ and consider $U_{\rho_i}^{mod}$ as $AM(\vec{u_i})$.*

The method M-AM describes an algorithm for determining a precision level for each input vector keeping an eye on the whole input data and accordingly creating an associative memory for each input vector of a given data set $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$. However, at this stage our next concern is to check what amount of associative patterns can be generated for each input vector following the method M-AM.

Proposition V.2. *Given input vectors $\vec{u_1}, \vec{u_2}, ..., \vec{u_k}$, for an arbitrary $\vec{u_i}$, if the precision level ρ_i is s.t. $p_i < \rho_i < p_i + p_j$ for any $j \neq i$ holds, then $|U_{\rho_i}| = {}^{n-p_i}C_{\rho_i - p_i}$*

Proof. According to M-AM for any $\vec{u} \in U_{\rho_i}$, $Unp(\vec{u}) \supseteq Unp(\vec{u_i})$, $|Unp(\vec{u})| \leq \rho_i$ where $p_i < \rho_i < p_i + p_j$, any $i \neq j$. So, for any $\vec{u} \in U_{\rho_i}$ $Unp(\vec{u})$ contains $Unp(\vec{u_i})$ and any $|Unp(\vec{u}) - Unp(\vec{u_i})|$ = ρ_i - p_i positions out of $|\{1, 2, \ldots, n\} - Unp(\vec{u_i})|$ = n-p_i positions of $Bits(\vec{u_i})$. Thus, $|U_{\rho_i}| = {}^{n-p_i}C_{\rho_i - p_i}$. □

Proposition V.3. *Given input vectors $\vec{u_1}, \vec{u_2}, ..., \vec{u_k}$, if for an arbitrary $\vec{u_i}$, there is no ρ_i s.t. $p_i < \rho_i < p_i + p_j$ for any $j \neq i$, then $|U_{\rho_i}^{mod}| = {}^{n-(p_i+|O_i|)}C_1$ where $O_i = \{L : p_i + p_L = min_j\ p_i + p_j, j \neq i\}$*

Proof. Clearly, $U_{\rho_i}^{mod} \subseteq U_{\rho_i}$ and as there is some $j \neq i$ for which p_j = 1, $\rho_i = p_i + 1$. So, following Proposition V.2 $\rho_i - p_i = 1$. Moreover, while creating $U_{\rho_i}^{mod}$ as we are considering that for any $u \in U_{\rho_i}^{mod}$ if $l \in O_i$ then $l \notin Unp(\vec{u})$, we also need to remove from $\{1, 2, \ldots, n\}$ all unknown positions of the vectors $\vec{u_l}$ such that $l \in O_i$. So, in contrary to Proposition V.2 here we have $n - (p_i + |O_i|)$ remaining positions from where $\rho_i - p_i$ that is 1 position can be chosen to be included in $Unp(\vec{u})$. So, we have ${}^{n-(p_i+|O_i|)}C_1$ many ways to create $Unp(\vec{u})$ by adding one additional position to $Unp(\vec{u_i})$. Hence $|U_{\rho_i}^{mod}| = {}^{n-(p_i+|O_i|)}C_1$. □

Following Proposition V.3, we can see that the more input vectors with unknown components equal to 1, i.e., $|O_i|$ increases, the number of vectors belonging to $U_{\rho_i}^{mod}$ decreases. For this reason, we should avoid having a large number of input data consisting of just one unknown component vector.

A. Stepwise Unknown Full Cover Input Data

Following Proposition V.3, we should avoid a set of input vectors such that each has only one unknown component. The sequel below will make the point more precise.

Proposition V.4. *Following M-AM for any n number of n-dimensional vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_n}$, if $AM(\vec{u_i}) \supsetneq \{\vec{u_i}\}$ then it can not be the case that $p_1 = p_2 = \cdots = p_n = 1$.*

Proof. Contrapositively, if for a data set satisfying conditions of M-AM we assume $p_1 = p_2 = \cdots = p_n = 1$, then for any $\vec{u_i}$ by Proposition IV.7 there is no ρ_i such that $p_i < \rho_i < p_i + p_j$ for any $j \neq i$, $i, j \in \{1, 2, \ldots, n\}$. Thus,$AM(\vec{u_i}) = U_{\rho_i}^{mod}$. But $O_i = \emptyset$. Hence, $AM(\vec{u_i}) = \{\vec{u_i}\}$. □

Definition V.5 (SUFC)**.** *Given a set of n-dimensional input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ with $|Unp(\vec{u_i})| = p_i$ for $1 \leq i \leq k$, and $p_1 = p_2 = \cdots = p_k = 1$ if $\cup_{i=1}^{k} Unp(\vec{u_i}) = \{1, 2, \ldots, n\}$ and $Unp(\vec{u_i}) \cap Unp(\vec{u_j}) = \emptyset$ for any $i \neq j$, then we call such data set as Stepwise Unknown Full Cover (SUFC).*

Thus, along with conditions mentioned in Definition IV.1 let us add another additional condition on the Dset.

Condition V.6. *A Dset of input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ should not satisfy SUFC.*

Based on the above discussion it is clear that apart from the conditions mentioned in Definition IV.1 we need some other conditions to be imposed in order to follow method M-AM in the process of creating associative memory for a Dset. Thus, below we present a modified notion of Dset.

Definition V.7 (Dset$_{AM}$)**.** *Given a set of input vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ we say it as a Dset$_{AM}$, if it is a Dset and satisfies Condition V.6.*

B. Processing Input Data Through Iterations

Based on the results and discussion presented in the above sections, we can notice a few important aspects for effectively using a given data in the process of creating associative patterns and then learning perceptron models based on them. The first is that, given a Dset of n-dimensional input vectors to apply method M-AM for creating associative memories for input vectors we may need to split the Dset over different iterations. Secondly, in order to avoid overlapping sets of associative patterns for all the vectors considered in a particular iteration the input vectors considered in each iteration should form a Dset$_{AM}$. Thirdly, based on Proposition IV.11 and Example IV.2 we recognize a need for selecting a Dset$_{AM}$ for each iteration satisfying a *top down decomposition ordering*, described below, by allowing gradually less unknown parts (or gradually more known parts) in the selection of input vectors.

C. Top-down Decomposition Ordering

We propose a method for ordering the input data to be selected in a $Dset_{AM}$ in such a way that it considers vectors with gradually increasing order of unknown parts while selecting for a $Dset_{AM}$ from a given data.

Definition V.8 (TDDO)**.** *Let us consider a set of input vectors, $In = \{\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}\}$ of dimension n, where $n > 3$.* [5] *A Top-Down Decomposition Ordering of input vectors is composed of the following steps.*

1) *Select a vector $\vec{u_j}$ ($1 \leq j \leq k$) such that $|Unp(\vec{u_j})|$ is minimum. Call the vector $\vec{u_{j1}}$.*
2) *Select the next vector from $In \setminus \{\vec{u_j}\}$ such that the number of unknown components of the selected vector is $|Unp(\vec{u_j})|$+1 and the unknown components of the new vector does not have common overlap with $Unp(\vec{u_j})$. Call the new vector $\vec{u_{j2}}$.*
3) *Continue the process as long as there is no vector from the remaining set of In satisfying condition (2).*

Example V.1. *Given a input data $In = \{\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}\}$ for $n = 10$. Following Definition V.8 consider input vectors from In such that $\vec{u_{j1}}$ with $|Unp(\vec{u_{j1}})| = 1$, $\vec{u_{j2}}$ with $|Unp(\vec{u_{j2}})| = 2$, $\vec{u_{j3}}$ with $|Unp(\vec{u_{j3}})| = 3$ and $\vec{u_{j4}}$ with $|Unp(\vec{u_{j4}})| = 4$, and they satisfy (MDUP). Clearly, for $n = 10$, there are no more than 4 vectors that have no common overlapping unknown components covering $\{1, 2, \ldots, 10\}$. Thus, an iteration consist of this set of vectors $\{\vec{u_1}, \vec{u_2}, \vec{u_3}, \vec{u_4}\}$.*

From Example V.1 we can see that following TDDO if we select input vectors, starting from lowest number of unknown components, in a strict increasing manner depending on the dimension of the vector there seems to be a threshold value for the number of unknown components. For instance, in case of Example V.1 for 10 dimensional vectors this threshold value is 4. The sequel below throws light on this.

Definition V.9 (Threshold and cardinality value for TDDO)**.** *Let a set of input vectors of n-dimension ($n > 3$) be given. To create a Dset$_{AM}$ for one iteration we first select a vector with exactly one unknown component and continue selecting a sequence of vectors following TDDO. When the process reaches to stage 3 of TDDO (see Definition V.8) the maximum number of unknown components of the last selected vector in Dset$_{AM}$ is considered to be the threshold value for TDDO for dimension n, denoted by T_n, and the cardinality of the Dset thus constructed is called the cardinality value for Dset$_{AM}$, denoted as $|Dset_{AM}|$.*

Following Definition V.9 it is clear that for an n-dimensional input vectors once we reach to the stage (3) of TDDO and get hold of a vector reaching threshold T_n we cannot include any vector in the respective Dset$_{AM}$ having more than T_n number of unknown components. So, one important question is how to determine the threshold T_n for any arbitrary dimension n.

Theorem V.10. *Given a set of input vectors of dimension n let T_n and $|Dset_{AM}|$ be respectively the threshold value and cardinality value of the Dset$_{AM}$ obtained following TDDO corresponding to an iteration. Then considering $|Dset_{AM}| = i$ the following relations hold between T_n and i.*

1) *If $n = 2, 3$, $T_n = 1$.*
2) *For any $n > 3$ and $i \geq 2$, $\frac{i(i+1)}{2} \leq n \leq \frac{i(i+1)}{2} + i$ holds and for all such n, $T_n = i$.*

Proof. For $n = 2, 3$ the case is straightforward.

For any n ($n > 3$) the increasing order of unknown components of the vectors selected in $Dset$ are $1, 2, \ldots T_n$. So, $|Dset_n| = T_n = i$. Following TDDO, $n - (1 + 2 + \cdots + i) \leq i$.
$\therefore n \leq \frac{i(i+1)}{2} + i$ and clearly for $n > 3$, the value i must satisfy $i \geq 2$.
Moreover, total number of positions covered by the unknown parts of the vectors of $Dset_n$ cannot exceed n. $\therefore \frac{i(i+1)}{2} \leq n$
$\therefore \frac{i(i+1)}{2} \leq n \leq \frac{i(i+1)}{2} + i$, for all $n > 3$ and $i \geq 2$.
Moreover, as for any $n \in [\frac{i(i+1)}{2}, \frac{i(i+1)}{2} + i]$ after covering gradually $1, 2, \ldots, i$ number of unknown positions, i.e., total $\frac{i(i+1)}{2}$ positions out of n dimension, there can be at most i positions remaining to be considered as unknown part of a vector. Thus, the threshold value T_n for all such n would be the same as i. □

Example V.2. *Following Theorem V.10, let us consider that $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ be a set of input vectors of dimension $n > 3$. For $i = 2$ we see that 4, $5 \in [\frac{i(i+1)}{2}, \frac{i(i+1)}{2} + i]$ and $T_n = 2$; that is following TDDO there cannot be any vector of more than 2 unknown parts in a $Dset_{AM}$. For $i = 3$, $n \in [6, 9]$ and for any n lying in this interval $T_n = 3$ as the gradual increasing order of the number of unknown components of any vector entitled to enter in a $Dset_{AM}$ of an iteration are 1, 2, 3. Similarly, for $i = 4$, $[\frac{i(i+1)}{2}, \frac{i(i+1)}{2} + i] = [10, 14]$ and*

[5] Following the condition that for a n-dimensional vector maximum number unknown components can be $\lfloor n/2 \rfloor$, for a vector of dimension one would not have any unknown component, and for a vector of dimension 2 or dimension 3 can have exactly one unknown component. So, TDDO does not apply to the vectors with dimension $n \leq 3$.

for any $n \in [10, 14]$ the threshold value for maximum unknown part following TDDO is $T_n = 4$.

D. Splitting data over iterations

As it is clear from previous sections, for n-dimensional vectors satisfying first two conditions mentioned in Definition IV.1 we may have $\sum_{i=0}^{\lfloor n/2 \rfloor} {}^nC_i$ of vectors in the input data In. However, following method M-AM in one iteration we can use only one small fragment of of data from In using TDDO. So, below we propose a complete procedure for using the data of In over some iterations.

Definition V.11 (SDOI)**.** *For a input data In containing n-dimensional vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$ by Splitting Data Over Iterations (SDOI) we mean the following procedure. In the first iteration, say I_0, we choose the vectors for $DSet^0_{AM}$ as follows.*

(i) *Select the first appeared vector $\vec{u_j} \in In$ with minimum $|Unp(\vec{u_j})|$ ($\leq T_n$) and rename it $\vec{u_{j1}}$*

(ii) *Following TDDO select vectors from $In \setminus \{\vec{u_j}\}$ with gradually increasing order of unknown components (considering first such appeared in input data) and continue till it reaches the threshold value T_n*

(iii) *Step (ii) ends when at most T_n vectors, say $\vec{u_{j1}}, \vec{u_{j2}} \ldots, \vec{u_{jT_n}} \in Dset^0_{AM}$. For the remaining $n - \frac{T_n(T_{n+1})}{2}$ positions select vectors from $In \setminus Dset^0_{AM}$ such that the gradual selection of the vectors satisfy monotonically decreasing order of unknown parts (satisfying Definition IV.1) as long as $\cup_{\vec{u} \in Dset^0_{AM}} Unp(\vec{u})$ = $\{1, 2, \ldots, n\}$.*

(iv) *Once Step (iii) ends, a new iteration of creating next $Dset_{AM}$ starts from step (i) over the remaining data $In \setminus Dset^0_{AM}$.*

(v) *Step (iv) continues as long as it is possible to create a new $Dset_{AM}$ following steps (i)-(iii).*

Example V.3. *For $n = 14$, in the iteration I_0 we may have $Dset_{AM} = \{\vec{u_1}, \vec{u_2}, \vec{u_3}, \vec{u_4}, \vec{u_5}, \vec{u_6}\}$ where $|Unp(\vec{u_i})| = i$ for $1 \leq i \leq 4$, and after reaching $T_{14} = 4$, $|Unp(\vec{u_5})| = 3$ and $|Unp(\vec{u_6})| = 1$. There can be several possibilities of constructing such a $Dset_{AM}$. For instance, one such is $Unp(\vec{u_1}) = \{1\}$, $Unp(\vec{u_2}) = \{2, 3\}$, $Unp(\vec{u_3}) = \{4, 5, 6\}$, $Unp(\vec{u_4}) = \{7, 8, 9, 10\}$, $Unp(\vec{u_5}) = \{11, 12, 13\}$, and $Unp(\vec{u_6}) = \{14\}$.*

Now let us present another example in order to understand how the associative memories for a given set of input vectors of a specific dimension n can be constructed following M-AM method and the data selection method described in Definition V.11.

Example V.4. *Given a input data In containing 8-dimensional vectors $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_k}$. Following Definitions V.1, V.8, V.9 V.11 consider the input vectors from In such that $\vec{u_1}$ with $Unp(\vec{u_1}) = \{1\}$, $\vec{u_2}$ with $Unp(\vec{u_2}) = \{2, 3\}$, $\vec{u_3}$ with $Unp(\vec{u_3}) = \{4, 5, 6\}$, [6] $\vec{u_4}$ with $Unp(\vec{u_4}) = \{7, 8\}$, that is, $I_0 = \{\vec{u_1}, \vec{u_2}, \vec{u_3}, \vec{u_4}\}$. We can check that the precision level for $\vec{u_1}$ is obtained following the first condition of Definition V.1. Since $\rho_1 = 2$ satisfying Condition IV.5, we can consider $\vec{u} \in U_{\rho_1}$, s.t. $Unp(\vec{u}) = \{1, i\}$ where $i \neq 1$, $i \leq 8$. That is $U_{\rho_1} = \{\langle ?, u_{12}, u_{13}, \ldots, u_{18}\rangle$: exactly one of $u_{1j} = ?$ for $j \neq 1$ and $u_{1k} = \vec{u_{ik}}$, $k \neq j, 1\}$. To illustrate more precisely, let us assume $\vec{u_1} = \langle ?, 0, 1, 0, 1, 0, 1, 0\rangle$. Then $U_{\rho_1} = \{\langle ?, ?, 1, 0, 1, 0, 1, 0\rangle, \langle ?, 0, ?, 0, 1, 0, 1, 0\rangle, \langle ?, 0, 1, ?, 1, 0, 1, 0\rangle, \ldots, \langle ?, 0, 1, 0, 1, 0, 1, ?\rangle\}$. For the other vectors $\vec{u_i}$, $i = 2, 3, 4$, the precision levels cannot be obtained following the first condition of Definition V.1 (see Proposition IV.7). Therefore, we follow the process of modifying U_{ρ_i}. According to Proposition IV.7, $O_i = \{1\}$, $\rho_i = p_i + p_1$ for any $i \neq 1$, $1 < i \leq 4$. [7] Therefore we should consider $\vec{u} \in U^{mod}_{\rho_i}$, s.t. $Unp(\vec{u}) = Unp(\vec{u_i}) \cup \{j\}$ where $1 < i \leq 4$, $j \neq 1$, $j \leq 8$, $Unp(\vec{u_i}) \cap \{j\} = \emptyset$. That is $U^{mod}_{\rho_2} = \{\langle u_{21}, ?, ?, u_{24}, \ldots, u_{28}\rangle$: exactly one of $u_{2j} = ?$ for $j \neq 1, 2, 3$ and $u_{2k} = \vec{u_{ik}}, k \neq j, 2, 3\}$. To illustrate this, let us say a given vector $\vec{u_2} = \langle 1, ?, ?, 0, 1, 0, 1, 0\rangle$. Then $U^{mod}_{\rho_2} = \{\langle 1, ?, ?, ?, 1, 0, 1, 0\rangle, \langle 1, ?, ?, 0, ?, 0, 1, 0\rangle, \langle 1, ?, ?, 0, 1, ?, 1, 0\rangle, \langle 1, ?, ?, 0, 1, 0, ?, 0\rangle \langle 1, ?, ?, 0, 1, 0, 1, ?\rangle\}$. Similarly, $U^{mod}_{\rho_3} = \{\langle u_{31}, u_{32}, u_{33}, ?, ?, ?, u_{37}, u_{38}\rangle$: exactly one of $u_{3j} = ?$ for $j \neq 1, 4, 5, 6$ and $u_{3k} = \vec{u_{ik}}, k \neq j, 4, 5, 6\}$ and $U^{mod}_{\rho_4} = \{\langle u_{41}, \ldots, u_{46}, ?, ?\rangle$: exactly one of $u_{4j} = ?$ for $j \neq 1, 7, 8$ and $u_{4k} = \vec{u_{ik}}, k \neq j, 7, 8\}$.*

VI. Information Retrieval from Associative Memory Using Perceptron

In this section, we attempt to sketch the idea of projecting a 2-layer perceptron from a $Dset_{AM}$ generated by an iteration following SDOI. In contrary to [3], in our case a $Dset_{AM}$ contains input data with unknown components. Following [9], [10], if the perceptron is trained on samples with incomplete feature vectors, i.e., partial input data, significant propagation errors may occur during the inference process. In addition, if important features are frequently missing, the learned decision boundaries may become biased or suboptimal, which can degrade the model's prediction performance. To get rid of this problem, it is essential to implement a process to replace missing data before training. This allows the perceptron to learn from partially known data, thereby outputting incomplete input data as complete data. So, below we propose a way to replace the partial input vectors with a potential full vector.

A. Replacing Partial Vectors By Full Vectors

The $Dset_{AM}$ generated by an iteration of SDOI contains vectors satisfying (MDUP). That is, the unknown parts of any pair of vectors do not have common overlap with each others. This feature can be utilized to reconstruct full vectors before training the perceptron model. We replace each vector from a $Dset_{AM}$ as follows.

1) Given $\vec{u_i}, \vec{u_j} \in Dset_{AM}$ for any $i \neq j$, $1 \leq i, j \leq T_n$ the unknown component of the vector $\vec{u_i}$ is replaced with the

[6] Following Theorem V.10, the threshold value $T_n = 3$ for $n = 8$.

[7] We only consider the given vectors $\vec{u_1}, \vec{u_2}, \vec{u_3}, \vec{u_4}$.

known components of the vector $\vec{u_j}$. For vector $\vec{u_i}$ only the positions from $Unp(\vec{u_i}) \cap Bits(\vec{u_j})$ are replaced with that of $\vec{u_j}$. The new vector created in this way is called $\vec{u_{ij}}$. Following similar process $\vec{u_{ji}}$ is also obtained.

2) Check Hamming distance between $\vec{u_{ij}}$ and $\vec{u_{ji}}$, for any $i \neq j$
3) Among the vectors $\vec{u_{ij}}$ $j \neq i$ obtained by exchanging parts with $\vec{u_i}$ the one with minimum Hamming distance, would determine the full vector that can be selected to replace $\vec{u_i}$ considering as the most plausible ideal representation of $\vec{u_i}$.

Following the above method, we can expect that the reconstructed input vector will match more closely the ideal input data set without noise and unknown parts. Using the Hamming distance as a similarity measure, we can systematically select the most reasonable candidate from the pool of potential complete vectors, obtaining an input data set that better matches the intended representation of the original set of partial input vectors.

Example VI.1. *For $n = 10$, in an iteration I_0 we may have $Dset^0_{AM} = \{\vec{u_1}, \vec{u_2}, \vec{u_3}, \vec{u_4}\}$, where $|Unp(\vec{u_i})| = i$, for $1 \leq i \leq 4$, $T_n = 4$.*
Let the vectors be $\vec{u_1} = \langle 0,1,0,0,?,1,1,1,0,0\rangle$, $\vec{u_2} = \langle 1,1,0,0,0,1,0,?,0,?\rangle$, $\vec{u_3} = \langle 1,0,0,0,0,?,?,1,?,0\rangle$, $\vec{u_4} = \langle ?,?,?,?,0,1,1,1,0,0\rangle$, where ? denotes missing value in a vector.
Replacing $Unp(\vec{u_1})$ with $Bits(\vec{u_2})$, $Bits(\vec{u_3})$ and $Bits(\vec{u_4})$, we respectively have $\vec{u_{12}} = \langle 0,1,0,0,0,1,1,1,0,0\rangle$, $\vec{u_{13}} = \langle 0,1,0,0,0,1,1,1,0,0\rangle, \vec{u_{14}} = \langle 0,1,0,0,0,1,1,1,0,0\rangle$.
Similarly, replacing the unknown components $\vec{u_2}$ with the known components of other vectors we have, $\vec{u_{21}} = \langle 1,1,0,0,0,1,0,1,0,1\rangle$, $\vec{u_{23}} = \langle 1,1,0,0,0,1,0,1,0,1\rangle$, $\vec{u_{24}} = \langle 1,1,0,0,0,1,0,1,0,0\rangle$. For $\vec{u_3}$ we have $\vec{u_{31}} = \langle 1,0,0,0,0,1,1,1,0,0\rangle$, $\vec{u_{32}} = \langle 1,0,0,0,0,1,0,1,0,0\rangle$, $\vec{u_{34}} = \langle 1,0,0,0,0,1,1,1,0,0\rangle$, and for $\vec{u_4}$ we have $\vec{u_{41}} = \langle 0,1,0,0,0,1,1,1,0,0\rangle$, $\vec{u_{42}} = \langle 1,1,0,0,0,1,1,1,0,0\rangle$, $\vec{u_{43}} = \langle 1,0,0,0,0,1,1,1,0,0\rangle$.
Next, we check the Hamming distance between each pair of exchanged vectors. $D_H(\vec{u_{12}}, \vec{u_{21}}) = 3$, $D_H(\vec{u_{13}}, \vec{u_{31}}) = 2$, $D_H(\vec{u_{14}}, \vec{u_{41}}) = 0$, $D_H(\vec{u_{23}}, \vec{u_{32}}) = 2$, $D_H(\vec{u_{24}}, \vec{u_{42}}) = 1$, $D_H(\vec{u_{34}}, \vec{u_{43}}) = 0$. [8]
From the above calculation we see, the minimum Hamming distance of $D_H(\vec{u_{1j}}, \vec{u_{j1}})$ where $j \in \{2,3,4\}$, is $D_H(\vec{u_{14}}, \vec{u_{41}})$. So partial vectors $\vec{u_1}$ and $\vec{u_4}$ are exchanged the parts within themselves the newly obtained full vectors $\vec{u_{14}}$ appears to be most closely fitting. Thus, $\vec{u_1}$ can be substituted with full vector $\vec{u_{14}}$. Similarly, partial vector $\vec{u_2}$ can be substituted with full vector $\vec{u_{24}}$, partial vector $\vec{u_3}$ can be substituted with full vector $\vec{u_{34}}$ and partial vector $\vec{u_4}$ can be substituted with full vector $\vec{u_{41}}$ or $\vec{u_{43}}$.

[8]Since Hamming metric, i.e., $D_H(\vec{u_{12}}, \vec{u_{21}}) = D_H(\vec{u_{21}}, \vec{u_{12}})$, there are only six results left in this case excluding duplicate results.

B. Perceptron model

As described in Section II, creating perceptron model following [3] solely depends on two facts; one is choosing precision levels for the input patterns in a way that their associative memories do not overlap and the second is to convert partial associative patterns of a given input (full) pattern to full patterns. The first point is important in establishing the claim that the set of input-outputs pairs in the extended function f_A are linearly separable and it is required to develop the perceptron model for f_A. In our case, we opt to depart from several aspects considered in [3], and these are mentioned in Introduction as well as in Section II. However, from the development and results presented in the sections above, starting from Section III, it is clear that both the contexts responsible for smooth projection of the function f_A in a perceptron model are guaranteed. Thus, in our case the same strategy as in [3] of creating perceptron model can be followed.

VII. Conclusion

The paper presents some limitations of the approach to creating associative memory in the context of retrieving information from partial patterns using perceptron model presented in [3] and proposes a few more general perspectives of different notions used in [3], [8]. First of all, unlike [3] we propose incomplete information in the input data by allowing partial vectors in the input data. Secondly, the paper proposes a new perspective of measuring distance between two partial vectors using the notion of Hamming distance. According to [3] two partial vectors with different unknown fragments cannot be comparable. Following our approach this limitation is overcome. Thirdly, our proposal for creating associative patterns for an input (partial) vector $\vec{u_i}$ depends on a notion of '$\vec{u}$ is a potential extension of $\vec{u_i}$', denoted by $\vec{u} \otimes \vec{u_i}$. The idea is developed generalizing the notion of $\odot$ presented in [3]. Moreover, in the process of converting incomplete or partial patterns by some plausible full patterns in [3] a random process of replacing unknown parts by 0's is followed. Whereas in our context, we consider an approach of finding plausible full pattern of an input partial pattern by exchanging content of the given pattern with other input vectors of the training data. The idea is similar to solving a puzzle related to constructing a full image of a desired object from several available pieces where different fragments of the full image are visible.

From theoretical perspective, though we manage to resolve the limitations mentioned before, following our proposal a given data set of n-dimensional input vectors cannot be used together in one iteration to train a perceptron model. However, there seems to be a possibility of creating perceptron model from aggregating data from different iterations. In [12], the authors proposed different models for generalizing Hamming associative memory. In particular, they introduced a notion of *local Hamming memory* which is based on splitting the input vectors (input data) into smaller clusters and learning the dependence of all the input vectors in that cluster with their respective outputs. Modifying the idea of local

Hamming memory the authors [12] also defined a notion of *decoupled Hamming associative memory* which considers non-overlapping windows (clusters) of input vectors. Though there is a similarity of their considered idea of non-overlapping windows of input vectors [12] with our idea of splitting the input vectors over iterations, there is a difference in basic consideration. In our case, the input vectors can be partial and in [12] the input vectors are full binary vectors. Following [3] non-overlapping associative memories of the input vectors allows a direct projection of the extended function f_A into a multilayered perceptron model. In our context, though we have departed from [3] on many aspects, we stick to the technical advantage of non-overlapping associative memories of the input vectors. However, following the notion of Distributed Associative Memory (DAM), in [13], the authors envisaged a possibility of representing multiple association of an input data with different sets of patterns. This idea can be worth investigating further in our context by allowing associative patterns of two input vectors to have common intersection. But this needs additional research and experiments in order to optimize the performance of the proposed method and check the changes needed to develop the set-up based on DAM.

REFERENCES

[1] P. M. Churchland, *Plato's camera: How the physical brain captures a landscape of abstract universals*. MIT press, 2012. DOI: 10.1080/10848770.2016.1139356. [Online]. Available: https://doi.org/10.7551/mitpress/9116.001.0001.

[2] T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997, ISBN: 0070428077. [Online]. Available: https://dl.acm.org/doi/10.5555/541177.

[3] C.-H. Chen and V. Honavar, "A neural architecture for content as well as address-based storage and recall: Theory and applications," *Connection Science*, vol. 7, no. 3, pp. 281–300, 1995. DOI: 10.1080/09540099509696194.

[4] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982. DOI: 10.1073/pnas.79.8.2554.

[5] B. Prasad, K. Prasad, S. Yeruva, P. Murty, and S. Rama, "A study on associative neural memories," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 1, no. 6, pp. 124–133, 2010. DOI: 10.14569/IJACSA.2010.010619.

[6] G. R. Barker and E. C. Warburton, "Multi-level analyses of associative recognition memory: The whole is greater than the sum of its parts," *Current opinion in behavioral sciences*, vol. 32, pp. 80–87, 2020. DOI: 10.1016/j.cobeha.2020.02.004.

[7] J. Steinberg and H. Sompolinsky, "Associative memory of structured knowledge," *Scientific Reports*, vol. 12, no. 1, p. 21 808, 2022. DOI: 10.1038/s41598-022-25708-y.

[8] C. H. Chen and V. Honavar, "Neural network architecture for high-speed database query processing," *Microcomputer Applications*, vol. 15, no. 1, pp. 7–13, 1996, ISSN: 0820-0750. [Online]. Available: https://dr.lib.iastate.edu/handle/20.500.12876/20228.

[9] C. F. Caiafa, Z. Wang, J. Solé-Casals, and Q. Zhao, "Learning from incomplete data by simultaneous training of neural networks and sparse coding," *arXiv preprint arXiv:2011.14047*, 2020. DOI: 10.48550/arXiv.2011.14047.

[10] W.-D. Chang, J. Shin, M.-Y. Song, and T.-S. Chon, "Enhanced back propagation algorithm for estimating ecological data with missing values," *WSEAS Transactions on Computers*, vol. 5, no. 9, pp. 2043–2048, Sep. 2006.

[11] C. Chen and V. Honavar, "Neural network automata," in *Proc. of World Congress on Neural Networks*, vol. 4, 1994, pp. 470–477.

[12] P. Watta and M. H. Hassoun, "Generalizations of the hamming associative memory," *Neural processing letters*, vol. 13, no. 2, pp. 183–194, 2001. DOI: 10.1023/A:1011384407294.

[13] T. Park, I. Choi, and M. Lee, "Distributed associative memory network with memory refreshing loss," *Neural Networks*, vol. 144, pp. 33–48, 2021. DOI: 10.1016/j.neunet.2021.07.030.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 145–154
DOI: 10.15439/2025F3136
ISSN 2300-5963 ACSIS, Vol. 43

Hybrid mutation MTD solution with dedicated SDN agents

Mateusz Fronczyk

Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
Email: mateusz.fronczyk.stud@pw.edu.pl

Mariusz Rawski
0000-0002-7489-0785
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
Email: mariusz.rawski@pw.edu.pl

***Abstract*—Reconnaissance is a crucial stage of cyberattacks, enabling attackers to gather information about system vulnerabilities. In computer networks, data regarding addressing and transmission paths is especially sensitive. This paper introduces a concept for a hybrid mutation system based on Software Defined Networking (SDN), combining address and data path mutation to improve security. The system employs central network management and pseudorandom, temporary addressing, which is periodically reconfigured. In contrast to existing Moving Target Defense (MTD) address mutation methods, end devices connect through dedicated agents, which can be implemented using a SmartNIC. The agent modifies packet headers at a network edge to obfuscate address information, minimizing the processing burden on SDN switches inside a network. The objective is to hinder attackers from discovering network details that could be exploited. A prototype was implemented using typical SDN components and containerized end devices. Tests confirmed the system's correctness and effectiveness in protecting the network structure and communication paths. This approach enhances the confidentiality of network parameters and limits the information available to potential attackers, making reconnaissance significantly more difficult, while minimizing SDN network control overhead.**

I. Introduction

NETWORK security is critical in today's digital landscape, as it protects sensitive data from unauthorized access and cyber threats while ensuring the integrity and availability of network resources. The network attack cycle spans several steps. The Cyber Kill Chain illustrates this cycle, detailing the stages an attacker typically goes through to successfully breach a network, from reconnaissance and weaponization to delivery, exploitation, installation, command and control, and finally, actions on objectives [1].

The reconnaissance is the initial phase where attackers gather information about their target, identifying potential vulnerabilities and entry points that can be exploited to gain unauthorized access. Stopping an attacker at the reconnaissance phase of a cyber attack provides significant security advantages by preventing further exploitation and reducing the overall risk to a network. However, spotting reconnaissance activity can be challenging due to several factors that make it stealthy and difficult to differentiate from legitimate network traffic. Traditional computer networks are inherently static, making them predictable targets for attackers who exploit static IP addresses, unchanged software stacks, and consistent routing paths to plan and execute cyberattacks. This predictability enables adversaries to conduct prolonged reconnaissance, identify vulnerabilities, and launch precise, well-timed exploits with minimal resistance. The static nature of computer networks gives the attacker a significant advantage over the defender, as attackers can use stealthy techniques to gather information over a long time to identify weaknesses without raising alarms [2].

A proposed solution to this issue was introduced in [3] as part of the USA Federal Networking and Information Technology Research and Development program, which seeks innovative approaches to address critical cybersecurity challenges. This approach, known as Moving Target Defense (MTD), represents a fundamental shift in securing computer networks. The core principle of MTD is that an attack, if successful, can only be effective once, as the system is continuously reconfigured to prevent the reuse of the same attack vector. By dynamically altering the attack surface, MTD increases the complexity and cost for attackers, reduces the window of vulnerability exposure, and enhances overall system resilience. Moving Target Defense mitigates this risk by introducing continuous, randomized changes to network attributes such as IP addresses, system configurations, and routing paths. By dynamically altering the attack surface, MTD disrupts adversarial strategies, forcing attackers to constantly reassess their tactics, thereby increasing the cost and complexity of cyber intrusions [4]. Unlike traditional static defenses relying on perimeter security mechanisms, MTD adopts a proactive security posture that introduces unpredictability while ensuring system reliability for legitimate users.

Numerous MTD strategies have been proposed for securing computer networks [5]. One of the key principles of MTD in networks is IP Address Mutation, which involves frequently changing the IP addresses of critical assets, preventing attackers from accurately mapping the network [6][7]. This technique disrupts traditional reconnaissance methods and limits the effectiveness of targeted attacks. Another effective approach is Dynamic Network Topology, which emphasizes

The research was carried out on devices cofunded by the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme.

dynamic changes in network architecture to confound attackers [8]. Route Mutation and Traffic Obfuscation techniques further enhance network security by dynamically altering packet routes, making it difficult for attackers to eavesdrop or execute man-in-the-middle attacks. This method ensures that sensitive data does not travel predictable paths, increasing the difficulty of successful network intrusions [9]. Service and Port Randomization also prevents attackers from leveraging known vulnerabilities by frequently changing open ports and service access points [10].

The benefits of MTD for networks are substantial. It disrupts reconnaissance efforts, reduces the attack window, and forces adversaries to adapt continuously, increasing their resource expenditure and reducing their ability to execute long-term attacks. Moreover, MTD enhances cyber resilience by introducing self-adaptive security measures that dynamically respond to emerging threats. However, challenges such as performance overhead, compatibility with legacy systems, and the need for intelligent automation must be addressed to optimize MTD implementation.

Software-Defined Networking (SDN) technology plays a key role in implementing most Moving Target Defence strategies, enabling dynamic, automated, and adaptive networking mechanisms. This technology provides programmability, flexibility, and scalability, making it possible to rapidly adjust network configurations, which constitutes the core of most network-based MTD methods. However, these frequent and unpredictable changes introduce significant stress on SDN switches and controllers, impacting performance, stability, and resource utilization.

As noted in [11], while centralization in SDN improves control and flexibility, it also introduces new attack vectors. Offloading certain security functions to edge components can mitigate these risks and improve scalability. With the advent of Smart Network Interface Cards (SmartNICs), there is potential to alleviate some of these pressures by offloading certain processing tasks from the SDN controllers and switches, thereby enhancing overall system efficiency and enabling more robust MTD solutions that can adapt in real-time without compromising network integrity. SmartNICs represent a significant evolution in network interface technology, offering enhanced functionality that extends far beyond traditional NICs. SmartNICs integrate programmable processors, memory, and dedicated hardware accelerators at their core, enabling them to offload network processing tasks [12].

This paper presents a hybrid addressing and route mutation mechanism that utilizes dedicated host agents with a lightweight SDN switch system. These host agents are designed to be implemented using SmartNICs, offloading specific tasks from core SDN switches. This approach not only improves the scalability of network management but also enhances performance by allowing core switches to focus on high-level routing decisions while host agents handle lower-level address mutation tasks efficiently.

II. Related Works

Several solutions based on the concept of address mutation and dynamic routing in network environments built on SDN architecture can be found in scientific literature.

One example of applying address mutation in SDN networks is the OpenFlow Random Host Mutation (OF-RHM) mechanism [13]. Each host is assigned a virtual IP address used during network communication in this solution. Importantly, these addresses are periodically changed, significantly complicating attacks based on traffic monitoring (e.g., sniffing) or network topology analysis. The address mutation process is handled by the switches, which also manage data traffic. Although this approach improves transmission security, it increases the load on the switches, which can negatively impact the overall performance and scalability of the network.

Another approach was presented in [14], where a routing algorithm based on fake addressing was proposed. In this model, switches establish paths between hosts using randomly assigned addresses, while the SDN controller manages the mapping of temporary addresses to actual ones. Despite its security advantages, this solution involves several significant challenges. Firstly, the mutation process executed on the switches generates additional load for these devices. Secondly, the path selection may lead to temporary network congestion, especially in more complex topologies, due to excessive packet flooding during route updates.

In [15], the authors proposed using local NAT tables on each switch. In this model, a dedicated MTD unit responsible for generating temporary addressing is connected to each switch. As a result, the SDN controller is relieved of some responsibility – its role is limited to initializing data transmission paths. Although this reduces the controller's load, the switches still need to handle the address translation process, which remains challenging for their performance.

On the other hand, the solution presented in [16] introduces an additional layer of switches called agents directly connected to end devices. These agents take over the task of address mutation, thereby offloading the main switches responsible for packet routing in the network. This approach limits the impact of mutation on the core infrastructure; however, it introduces another challenge – temporary addressing is created for each individual connection, which may lead to significant load on the SDN controller, especially in high-traffic environments.

III. MTD Concept

This paper proposes a solution based on the Moving Target Defense technique, utilizing a hybrid mutation approach. The proposed method introduces dynamic mutations at both the data transmission path level and the addressing level, specifically targeting the second and third layers of the OSI model, namely the MAC and IP address layers.

Our solution leverages the flexibility and centralized control offered by Software-Defined Networking, where the SDN controller plays a pivotal role. The controller is tasked with orchestrating both path and address mutations and dynamically installing flow rules on SDN switches throughout the network

via the OpenFlow protocol. To support the address mutation process, we introduce lightweight, specialized agents in the form of thin SDN switches that are designed to manipulate packet headers in real time.

These agents are deployed at the network edge, typically on end devices or SmartNICs. They are responsible for executing packet-level transformations in accordance with mutation policies received from the SDN controller. Each agent is capable of performing real-time mutation of MAC and IP addresses. The agents operate transparently to the applications running on the host devices, ensuring seamless communication while continuously altering address mappings to thwart reconnaissance and tracking attempts by potential adversaries.

By offloading the mutation tasks from the core SDN switches to these edge agents, the architecture reduces the processing overhead on network infrastructure components, improves scalability, and enables fine-grained control over address mutation. The result is a dynamic, resilient, and secure communication environment that significantly raises the bar for attackers attempting to exploit static network configurations.

The overall architecture of the proposed system is illustrated in Fig. 1, highlighting the interplay between the SDN controller, the network switches, and the deployed agents.

To establish a connection between devices, the controller performs a pseudo-random selection of one of the available paths labeled 1, 2, and 3 in the diagram. Data exchange takes place over the selected route for the duration defined by the mutation interval.

During operation, packets transmitted in the network are subject to modification to obscure their actual source and destination addresses. Temporary false addressing is applied, referred to as $tMAC$ (temporary MAC address) and tIP (temporary IP address). After the mutation interval expires, the controller generates new temporary addresses and selects new transmission paths for each end device connected to the network.

The real addresses, corresponding to the physical configuration of device interfaces, are denoted as $pMAC$ (permanent MAC address) and pIP (permanent IP address). From the perspective of communicating hosts, it is impossible to discover the actual address of the machine involved in data exchange, ensuring complete address obfuscation within the system. As a result, all packets transmitted within the network contain only temporary, falsified source and destination addresses in their headers.

This mechanism can operate successfully provided all communicating devices, including SDN switches, remain under the organization's control. The system also assumes that the organization has access to a large address pool within a given subnet, allowing for effective masking of real IP addresses. To avoid raising suspicion from a potential attacker, temporary IP addresses are selected from the available address space within the corresponding subnet.

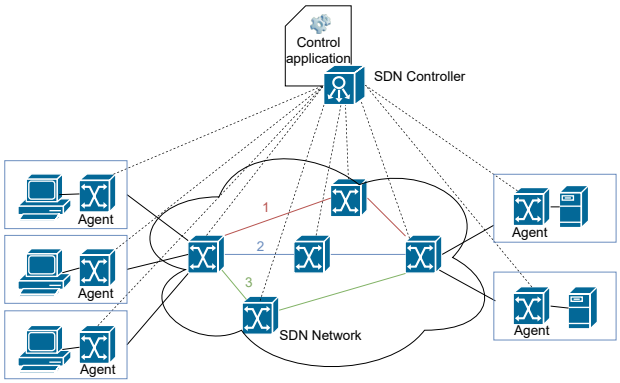


Fig. 1. The architecture of the proposed network

A. Connection Establishment and Handling

Fig. 2 presents a sequence diagram illustrating the phases of a sample connection establishment between *Host 1* and *Server 1*. This process includes:

1) **ARP Request for the *DNS Server***
   Communication with the *DNS Server* is the first step in establishing the connection. Assuming that the ARP tables on the end devices are initially empty, *Host 1* sends an ARP request to the broadcast address (ff:ff:ff:ff:ff:ff) asking for the MAC address of the *DNS Server*. The *H1 Agent* intercepts this request and replies to *Host 1* with a packet containing a temporary MAC address for the *DNS Server*, marked as *tMACDNS* in the diagram.
2) **DNS Query**
   Host 1, now knowing the temporary MAC address of the *DNS Server* (*tMACDNS*), sends a query requesting the IP address of a specific domain. In the presented diagram, the DNS query concerns the domain *server1.pl*, corresponding to the device *Server 1*. The query is received by the *H1 Agent* and then sent to the *SDN Controller* using a `PACKET_IN` message.
   Upon receiving the query, the *SDN Controller* performs a pseudo-random path selection to determine through which switch the communication will proceed. It also verifies whether the requested domain is handled by the *DNS Server*. If the domain does not exist, the connection attempt fails. If the domain is valid, the controller uses temporary addresses created for the current mutation state.
   One of these temporary addresses is the fake IP address of *Server 1* (*tIPS1*). The *SDN Controller* uses this address to modify the DNS server's configuration file, which maps actual server addresses to domain names.
   In the next step, the controller uses the temporary addressing to generate rules for appropriate agents and switches in the network. These rules are installed using `FLOW_MOD` messages. Rules are installed on the *H1*, *S1*, and *DNS* agents. The rules for *H1* and *S1* handle address obfuscation and ARP query handling

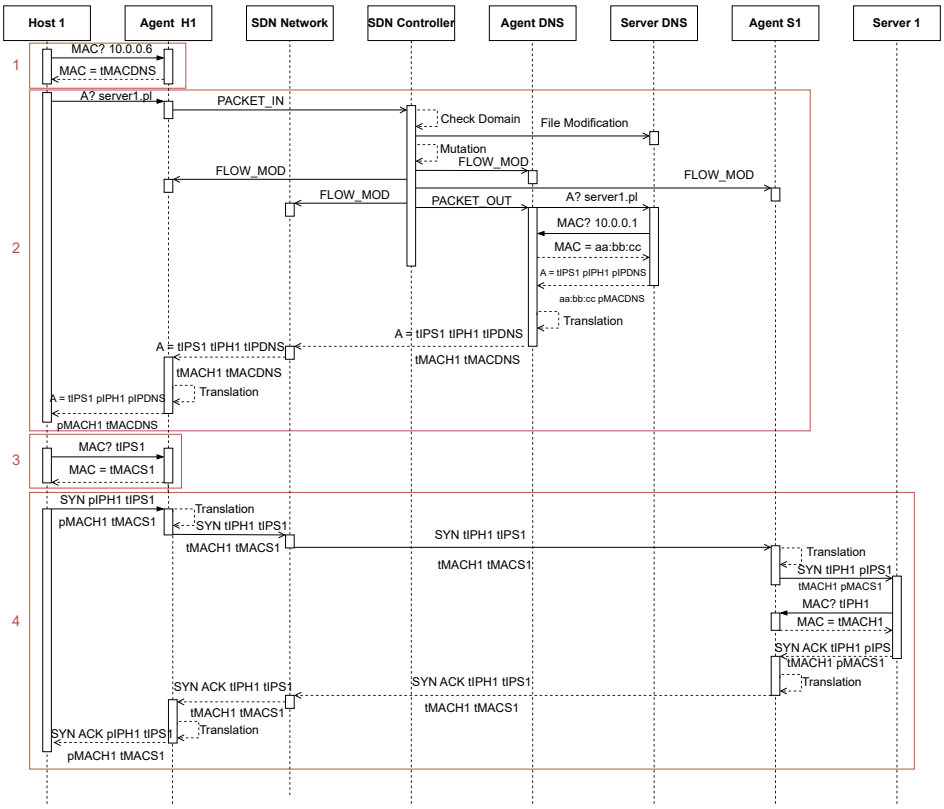


Fig. 2. Connection Establishment and Handling Diagram

for temporary IP addresses. The rule installed on the *DNS Agent* ensures that the DNS response is delivered to *Host 1* through the network. To achieve this, the *SDN Controller* additionally installs rules on *Host 1*'s agent and on the switches in the *SDN Network* to allow proper delivery of the response from the *DNS Server*. The rules installed on the *SDN Network* switches serve exclusively to forward packets between end devices using obfuscated addressing.

The DNS query is directly forwarded to the *DNS Agent* and then to the *DNS Server*. To send a response, the *DNS Server* must obtain the MAC address of *Host 1*, from which the original query was received. It sends an ARP request, which is intercepted by the *DNS Agent*, who replies with a fake MAC address `00:00:00:aa:bb:cc`. Due to the complexity of the proposed system's logic, the *DNS Agent* is configured with a flow rule that returns this MAC address for any IP in response to an ARP request. After receiving the ARP response, the *DNS Server* sends back a DNS response containing the fake IP address corresponding to the requested domain (*tIPS1*). The *DNS Agent*, upon receiving this response, sends it to the *SDN Controller*, which then forwards the packet to the *H1 Agent* via the *SDN Network*. While sending the response into the network, the *DNS Agent* performs address translation, inserting temporary addressing into the packet header: *tMACDNS* and *tIPDNS* for the source, and *tMACH1* and *tIPH1* for the destination. When the *H1 Agent* receives the response, it performs reverse translation by replacing the destination addresses with the real addressing of *Host 1* (*pMACH1* and *pIPH1*).

3) **ARP Request for *Server 1***
After receiving the DNS response containing *tIPS1*, *Host 1* sends an ARP request to obtain the MAC address assigned to this IP address. The rule installed by the *SDN Controller* on the *H1 Agent* handles the request, returning the temporary MAC address of *Server 1*, marked in the diagram as *tMACS1*.

4) **Communication Between Host and Server**
The next part of the diagram presents a sample TCP connection established between *Host 1* and *Server 1*. In the first phase, *Host 1* sends a SYN message as the initial step of the TCP three-way handshake. The packet header includes the real source address (*pIPH1* and *pMACH1*) and the temporary destination address (*tIPS1* and *tMACS1*). Upon receiving it, the *H1 Agent* performs address translation using *tMACH1* and *tIPH1*. The packet sent into the network has fully obfuscated addressing until it reaches the *S1 Agent*, who then

performs reverse translation by replacing the temporary destination address (*tMACS1* and *tIPS1*) with the real address (*pMACS1* and *pIPS1*). After *Server 1* successfully receives the segment, it replies with a SYN-ACK. Meanwhile, the *S1 Agent* handles an ARP query for the *Host 1* address mapping, returning the temporary address *tMACH1*. Since communication is bidirectional, *S1 Agent* once again performs address translation for the response, and on the receiving side, the *H1 Agent* performs reverse translation.

B. Mutation State Change During Connection Duration

One of the assumptions of the system operation is the continuation of the connection when a mutation state change occurs. Fig. 3 illustrates the sequence diagram that fulfills this assumption. The diagram shows a fragment of communication between *Server 1* and *Host 1*, which is downloading a file from the server. It is assumed that the file is large enough for the download process to be time-consuming. It is also assumed that the mutation interval *Tm* is short enough to change a mutation state during the file download process.

After the connection is successfully established, the file download from *Server 1* begins. Communication between *Host 1* and *Server 1* takes place using segments containing the $PUSHACK$ and ACK flags. The segment sent by *Server 1* with the $PUSHACK$ flag includes the data segment of the file requested by *Host 1*. The ACK segment, on the other hand, is the response from *Host 1* indicating successful reception of the segment from *Server 1*. This communication continues until *Server 1* sends all segments. At this point, *Server 1* sends a FIN flag segment, informing *Host 1* that all segments have been transmitted and initiating the connection termination process.

During the connection establishment, the *SDN Controller* places flow rules in the switch flow tables, enabling communication using temporary addressing marked in the diagram as addresses *t0*. These rules have a defined time *Tm*, after which they are permanently removed from the switch. When the rule responsible for address obfuscation on the *Agent H1* is removed, it sends a FLOW_REMOVED message to the *SDN Controller*, informing it that the rule has been deleted. The *SDN Controller* then re-selects the path for communication. Next, using the newly generated addressing in state *t1*, it installs flow rules on the switches responsible, in the case of *Agent H1* and *Agent S1*, for translating addresses *t1*, and for switches in the *SDN Network*, for communication using the *t1* addresses along the new path. After the rules are installed, the client-server communication continues with the new data path state and obfuscated addressing.

IV. Implementation

The experimental environment for the designed system was implemented using the Containernet network emulator, running on a virtual machine with Ubuntu 20.04. Open vSwitch (OVS) was used to implement the agents and SDN switches. These components form the foundation of a software-defined network architecture, which supports advanced dynamic behavior and fine-grained traffic control.

The SDN network is managed by the Ryu controller, which runs a dedicated control application written in Python. This application serves as the central intelligence of the system, orchestrating all major network operations such as address mutation, communication path management and DNS response manipulation. It interacts continuously with switches and agents, using the OpenFlow protocol to install and update flow rules that define how packets are forwarded throughout the network.

One of the core features of the control application is its dynamic addressing mechanism. At fixed intervals, the system triggers a mutation process that generates random temporary IP and MAC addresses for all endpoints in the network, including hosts, servers, and the DNS server. These addresses replace the real ones in communication flows, significantly reducing the feasibility of reconnaissance and tracking by potential attackers. The mutation runs asynchronously in the background, ensuring uninterrupted operation of other system components.

In addition to address mutation, the control application configures agent-hosted rules to intercept and respond to ARP queries. When hosts send ARP requests for the DNS server, the agents reply with dynamically assigned fake MAC addresses.

Another essential component of the architecture is a DNS server built using the Python $Dnslib$ library. It processes traditional DNS queries but is tightly integrated with the control logic. Upon receiving a DNS request, the control application extracts the queried domain, verifies it against a known mapping, and assigns a temporary IP address corresponding to the correct service endpoint. The DNS server configuration is then updated in real time to reflect the new address, ensuring all clients receive the current valid mapping.

Once a DNS query is resolved, the control application establishes the necessary flow rules to guide the DNS response and subsequent communication across the SDN switches and agents. For every transmission, source and destination addresses are translated to temporary counterparts, and then restored on the receiving end, enabling seamless and transparent bidirectional communication.

Communication paths themselves are also subject to mutation. The system periodically selects different routes for data transfer from a predefined set of options. New flow rules are installed along the newly chosen path, while outdated ones are removed.

All flow rules are configured with hard timeouts to ensure that temporary addressing remains effective. The control application is notified when a rule expires and installs a new rule set with updated addresses and paths. This dynamic rule replacement ensures continued communication fidelity while preserving the ephemeral nature of all addressing and routing information.

The overall logic of the control application described above is presented in Algorithm 1.

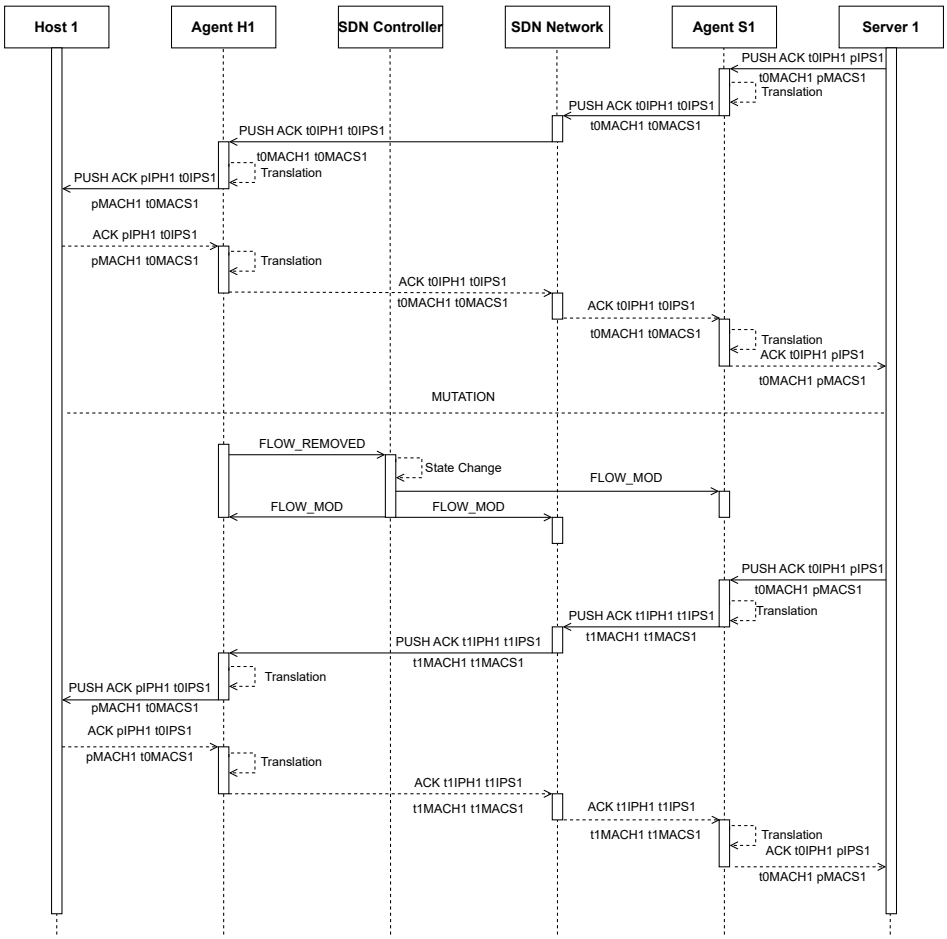


Fig. 3. Mutation State Change During the Connection

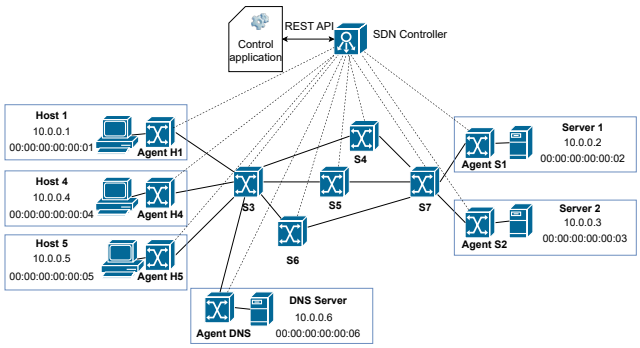


Fig. 4. SDN Network Topology in the Research Environment

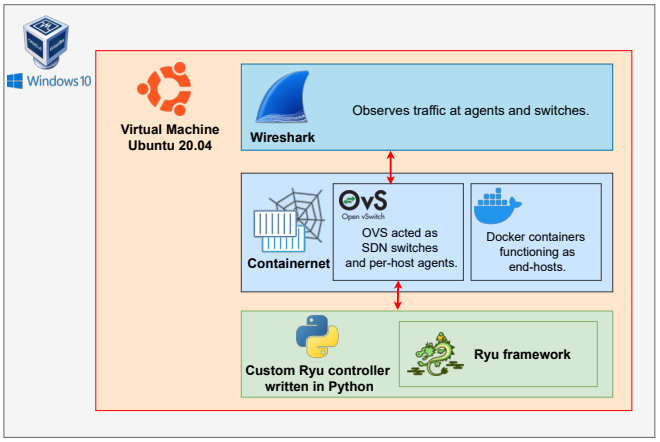


Fig. 5. Logical architecture of the implemented environment

The structure of the implemented network, including address allocation for individual components, is illustrated in Fig. 4, and the logical system architecture is shown in Fig. 5.

V. Validation and Results

A series of test scenarios were developed to verify the implemented system's operation.

Algorithm 1 Control Application Logic for SDN Mutation System

Require: Active SDN controller, initialized switches, known domain-to-server mappings
Ensure: Continuous operation with dynamic addressing and path mutation

Initialize mappings and path set
Start background **MutationLoop**
OnSwitchFeaturesReceived
if host agent **then**
    Install ARP rule, redirect DNS to controller
else if DNS agent **then**
    Install ARP rule
end if
OnPacketIn
if DNS query for known domain **then**
    Select endpoint, assign temporary IP/MACs
    Store mappings, update DNS config
    Install translation and ARP rules
    Select path, install forwarding rules
    Forward modified DNS packet
end if
OnFlowRemoved
if ARP rule expired **then**
    Reinstall ARP rule
else if IP flow expired **then**
    Identify flow, retrieve temp addresses
    Select path, reinstall updated rules
end if
MutationLoop (background)
while true **do**
    Sleep for interval, generate temp IP/MACs
    Update mappings, reinstall rules
end while

A. Address Mutation

The first test scenario illustrates network traffic on various devices belonging to the selected path during communication between *H4* and *Server 2* using the $ping$ command.

Fig. 8 shows traffic on the interface of the agent connected to *H4*. Initially, the agent responds to an ARP query regarding the DNS server, returning the temporary MAC address (*t_MAC_DNS*). Then, the host sends a DNS query for the IP address of the server domain $server2.pl$. After forwarding the query to the controller, it is sent to the DNS server, which returns a response containing the temporary IP address of the target server (*t_IP_S2*).

To begin the transmission of ICMP packets, the host must still obtain the MAC address assigned to the server. In response to the ARP query, the agent returns the temporary MAC address of the server (*t_MAC_S2*). Following this, correct communication occurs, using the temporary destination addressing.

The source addressing is obscured when the packet passes through the agent. The agent then replaces the real addressing with the temporary one corresponding to the given host (*t_MAC_H4* and *t_IP_H4*). Fig. 9 shows the observation of traffic on a selected switch in the *SDN Network* located on the communication path. The ICMP packets contain both temporary source and destination addressing.

The observation of traffic on the *Agent S2* device port, connecting the agent to the server, is shown in Figure 10. Upon receiving the appropriate packet from the network, the agent performs a reverse translation, replacing the temporary server addressing with the real one, labeled *p_MAC_S2* and *p_IP_S2*, allowing the response to be sent to the host.

B. Change of Address Mutation State

As part of the address mutation mechanism, after the time *Tm* has passed, communication in the system takes place using the temporary addressing of the new state. In the following Fig. 6 and Fig. 7, the communication between *Host 1* and *Server 1* during the download of an HTTP page is presented, using temporary addresses in two different states of the system.

During the first two connections (lines 1-27 and 3-6 of listings in Fig. 6 and Fig. 7 respectively), *Host 1* communicates with *Server 1* using the temporary address 10.0.0.165, and the server 10.0.0.53. After the time *Tm*, defined in the test as 10 seconds, the subsequent page download (lines 29-41 and 7-8) takes place using a new pair of temporary addresses: 10.0.0.237 for *Host 1* and 10.0.0.102 for *Server 1*.

```
1  root@h1:/# wget http://server1.pl:8080
2  --2025-01-07 17:28:49--  http://server1.pl:8080/
3  Resolving server1.pl (server1.pl)... 10.0.0.53
4  Connecting to server1.pl (server1.pl)
5  |10.0.0.53|:8080... connected.
6  HTTP request sent, awaiting response... 200 OK
7  Length: 1028 (1.0K) [text/html]
8  Saving to: 'index.html'
9
10 100%[============>] 1,028       --.-K/s   in 0s
11
12 2025-01-07 17:28:49 (244 MB/s) -
13 'index.html' saved [1028/1028]
14
15 root@h1:/# wget http://server1.pl:8080
16 --2025-01-07 12:28:52--  http://server1.pl:8080/
17 Resolving server1.pl (server1.pl)... 10.0.0.53
18 Connecting to server1.pl (server1.pl)
19 |10.0.0.53|:8080... connected.
20 HTTP request sent, awaiting response... 200 OK
21 Length: 1028 (1.0K) [text/html]
22 Saving to: 'index.html.1'
23
24 100%[============>] 1,028      --.-K/s   in 0s
25
26 2025-01-07 17:28:52 (213 MB/s) -
27 'index.html.1' saved [1028/1028]
28
29 root@h1:/# wget http://server1.pl:8080
30 --2025-01-07 17:29:09--  http://server1.pl:8080/
31 Resolving server1.pl (server1.pl)... 10.0.0.102
32 Connecting to server1.pl (server1.pl)
33 |10.0.0.102|:8080... connected.
34 HTTP request sent, awaiting response... 200 OK
35 Length: 1028 (1.0K) [text/html]
36 Saving to: 'index.html.2'
37
38 100%[============>] 1,028       --.-K/s   in 0s
39
40 2025-01-07 17:29:09 (213 MB/s) -
41 'index.html.2' saved [1028/1028]
```

Fig. 6. Downloading a page from *Server 1* on device *Host 1*

```
root@server1:/# python3 -m http.server 8080
Serving HTTP on 0.0.0.0 port 8080 ...
10.0.0.165 - - [07/Jan/2025 17:28:49]
"GET / HTTP/1.1" 200 -
10.0.0.165 - - [07/Jan/2025 17:28:52]
"GET / HTTP/1.1" 200 -
10.0.0.237 - - [07/Jan/2025 17:29:09]
"GET / HTTP/1.1" 200 -
```

Fig. 7. Received requests of *Server 1*

C. Mutation of Path and Address State During the Connection

The implemented system supports mutation during the connection, ensuring continuous communication. In the test on device *Host 5*, a 1GB file was downloaded from *Server 1*. The time *Tm* was set to 5 seconds, allowing for the observation of frequent mutation state changes in the system.

Upon establishing a connection with *Server 1*, *Host 1* received the following messages, shown in Fig. 11:

- temporary MAC address of the *DNS Server* 23:DA:A9:89:27:51 (*t_MAC_DNS*),
- temporary IP address of *Server 1* 10.0.0.185 (*t_IP_S1_t0*),
- temporary MAC address of *Server 1* B1:56:ED:C4:2D:06 (*t_MAC_S1_t0*).

Subsequent communication involves downloading the file "plik.bin" from *Server 1*. Fig. 12 shows the traffic observed on switch *S5* (chosen for path mutation) with completely obfuscated addressing. *Host 5* uses the temporary address D1:58:C0:F7:AE:3C (*t_MAC_H5_t0*) and 10.0.0.247 (*t_IP_H5_t0*). Fig. 13 depicts the situation of a mutation state change. After the time *Tm* has passed, the flows on the corresponding switches are updated, enabling continued communication with a new address and path. In the new mutation state for the path, the switch changes to *S4*, through which the segments are transmitted. Additionally, new temporary addresses appear in the packet headers:

- MAC address of *Host 5* 93:97:15:01:69:DC (*t_MAC_H5_t1*),
- IP address of *Host 5* 10.0.0.236 (*t_IP_H5_t1*),
- MAC address of *Server 1* D3:03:0A:06:8F:8A (*t_MAC_S1_t1*),
- IP address of *Server 1* 10.0.0.21 (*t_IP_S1_t1*).

Upon the mutation state change, retransmission of the segments occurs. This happens so quickly that the connection continues with the new addressing shortly afterward, enabling the entire file to be downloaded.

VI. Discussion

The test scenarios described above validate the correct functioning of the proposed Moving Target Defense system based on hybrid mutations at both the MAC and IP layers, as well as dynamic path adjustments during communication.

The results from the Address Mutation test confirm that the agents correctly perform address obfuscation and translation at the network edges. Temporary addresses are assigned and managed without impacting the continuity of communication from the hosts' perspective. This behavior is evident from the consistent and uninterrupted exchange of ICMP packets and successful DNS resolution processes involving mutated addressing.

The Change of Address Mutation State scenario demonstrates the system's ability to transition between different temporary address states during normal operations seamlessly. After the predefined timeout interval (*Tm*), new sets of temporary IP addresses are allocated and applied transparently, while ongoing services (such as HTTP downloads) continue unaffected. This dynamic reassignment capability significantly increases the complexity for any adversary attempting to map the network or track devices, reinforcing the defensive posture.

Finally, the Mutation of Path and Address State During the Connection scenario showcases the most demanding feature of the system—live mutation of both the data path and addressing during an active session. During a large file transfer, the system's behavior proves that the agents and controller can coordinate effectively to update switch flows and reassign addressing without disrupting the data stream. Segment retransmissions and address switching occur within such short intervals that the host application layer remains unaware of the underlying changes.

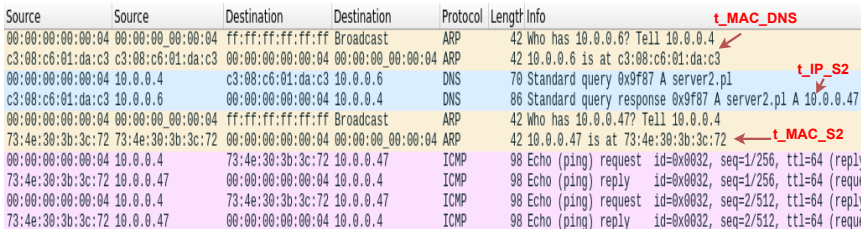
The tests and observed results prove that the proposed solution functions as intended. It achieves its core objective of creating a dynamic and unpredictable network communication environment, effectively increasing potential attackers' difficulty in exploiting static addressing schemes or consistent routing paths.

Although effective in the test environment, the system uses temporary IPv4 addresses from a limited subnet. In larger networks, this may lead to address exhaustion. Supporting IPv6 would expand the address pool and enhance scalability and obfuscation.

VII. Conclusion

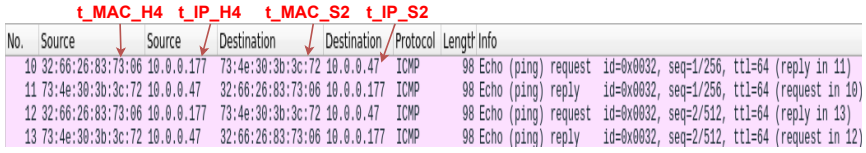
This paper presents an MTD system employing a hybrid mutation strategy involving address and path alterations to enhance network security. By cyclically mutating these elements during connection initiation and throughout the session duration, the proposed approach significantly increases the complexity of network reconnaissance for potential attackers. The design leverages dedicated packet processing agents, which can be efficiently implemented on SmartNICs at end hosts, enabling high-performance execution with minimal impact on system resources.

The prototype implementation, tested in an SDN-based environment utilizing OVS switches, demonstrated that the mutation mechanism operates seamlessly without disrupting normal network operations. Experimental results validated the system's effectiveness in improving the security posture of corporate networks by dynamically and unpredictably altering attack surfaces. Future work will optimize mutation intervals, integrate adaptive intelligence for mutation decisions, and evaluate performance in larger and more diverse network topologies.



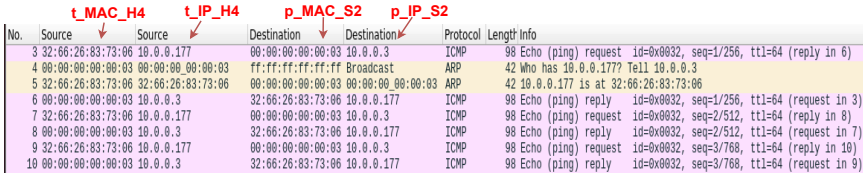
Source	Source	Destination	Destination	Protocol	Length	Info
00:00:00:00:00:04	00:00:00_00:00:04	ff:ff:ff:ff:ff:ff	Broadcast	ARP	42	Who has 10.0.0.6? Tell 10.0.0.4
c3:08:c6:01:da:c3	c3:08:c6:01:da:c3	00:00:00:00:00:04	00:00:00_00:00:04	ARP	42	10.0.0.6 is at c3:08:c6:01:da:c3
00:00:00:00:00:04	10.0.0.4	c3:08:c6:01:da:c3	10.0.0.6	DNS	70	Standard query 0x9f87 A server2.pl
c3:08:c6:01:da:c3	10.0.0.6	00:00:00:00:00:04	10.0.0.4	DNS	86	Standard query response 0x9f87 A server2.pl A 10.0.0.47
00:00:00:00:00:04	00:00:00_00:00:04	ff:ff:ff:ff:ff:ff	Broadcast	ARP	42	Who has 10.0.0.47? Tell 10.0.0.4
73:4e:30:3b:3c:72	73:4e:30:3b:3c:72	00:00:00:00:00:04	00:00:00_00:00:04	ARP	42	10.0.0.47 is at 73:4e:30:3b:3c:72
00:00:00:00:00:04	10.0.0.4	73:4e:30:3b:3c:72	10.0.0.47	ICMP	98	Echo (ping) request id=0x0032, seq=1/256, ttl=64 (repl
73:4e:30:3b:3c:72	10.0.0.47	00:00:00:00:00:04	10.0.0.4	ICMP	98	Echo (ping) reply id=0x0032, seq=1/256, ttl=64 (requ
00:00:00:00:00:04	10.0.0.4	73:4e:30:3b:3c:72	10.0.0.47	ICMP	98	Echo (ping) request id=0x0032, seq=2/512, ttl=64 (repl
73:4e:30:3b:3c:72	10.0.0.47	00:00:00:00:00:04	10.0.0.4	ICMP	98	Echo (ping) reply id=0x0032, seq=2/512, ttl=64 (requ

Fig. 8. Traffic observed on the port of the *Agent H4* device connecting the agent to the host



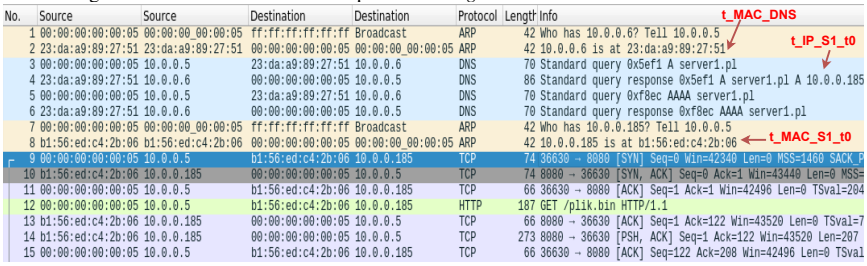
No.	Source	Source	Destination	Destination	Protocol	Length	Info
10	32:66:26:83:73:06	10.0.0.177	73:4e:30:3b:3c:72	10.0.0.47	ICMP	98	Echo (ping) request id=0x0032, seq=1/256, ttl=64 (reply in 11)
11	73:4e:30:3b:3c:72	10.0.0.47	32:66:26:83:73:06	10.0.0.177	ICMP	98	Echo (ping) reply id=0x0032, seq=1/256, ttl=64 (request in 10)
12	32:66:26:83:73:06	10.0.0.177	73:4e:30:3b:3c:72	10.0.0.47	ICMP	98	Echo (ping) request id=0x0032, seq=2/512, ttl=64 (reply in 13)
13	73:4e:30:3b:3c:72	10.0.0.47	32:66:26:83:73:06	10.0.0.177	ICMP	98	Echo (ping) reply id=0x0032, seq=2/512, ttl=64 (request in 12)

Fig. 9. Traffic observed on the switch in the *SDN Network*



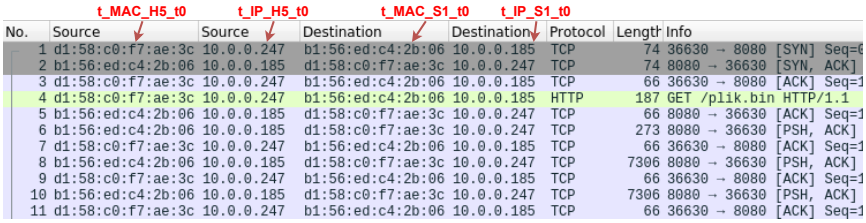
No.	Source	Source	Destination	Destination	Protocol	Length	Info
3	32:66:26:83:73:06	10.0.0.177	00:00:00:00:00:03	10.0.0.3	ICMP	98	Echo (ping) request id=0x0032, seq=1/256, ttl=64 (reply in 6)
4	00:00:00:00:00:03	00:00:00_00:00:03	ff:ff:ff:ff:ff:ff	Broadcast	ARP	42	Who has 10.0.0.177? Tell 10.0.0.3
5	32:66:26:83:73:06	32:66:26:83:73:06	00:00:00:00:00:03	00:00:00_00:00:03	ARP	42	10.0.0.177 is at 32:66:26:83:73:06
6	00:00:00:00:00:03	10.0.0.3	32:66:26:83:73:06	10.0.0.177	ICMP	98	Echo (ping) reply id=0x0032, seq=1/256, ttl=64 (request in 3)
7	32:66:26:83:73:06	10.0.0.177	00:00:00:00:00:03	10.0.0.3	ICMP	98	Echo (ping) request id=0x0032, seq=2/512, ttl=64 (reply in 8)
8	00:00:00:00:00:03	10.0.0.3	32:66:26:83:73:06	10.0.0.177	ICMP	98	Echo (ping) reply id=0x0032, seq=2/512, ttl=64 (request in 7)
9	32:66:26:83:73:06	10.0.0.177	00:00:00:00:00:03	10.0.0.3	ICMP	98	Echo (ping) request id=0x0032, seq=3/768, ttl=64 (reply in 10)
10	00:00:00:00:00:03	10.0.0.3	32:66:26:83:73:06	10.0.0.177	ICMP	98	Echo (ping) reply id=0x0032, seq=3/768, ttl=64 (request in 9)

Fig. 10. Traffic observed on the port of the *Agent S2* device connected to the server



No.	Source	Source	Destination	Destination	Protocol	Length	Info
1	00:00:00:00:00:05	00:00:00_00:00:05	ff:ff:ff:ff:ff:ff	Broadcast	ARP	42	Who has 10.0.0.6? Tell 10.0.0.5
2	23:da:a9:89:27:51	23:da:a9:89:27:51	00:00:00:00:00:05	00:00:00_00:00:05	ARP	42	10.0.0.6 is at 23:da:a9:89:27:51
3	00:00:00:00:00:05	10.0.0.5	23:da:a9:89:27:51	10.0.0.6	DNS	70	Standard query 0x5ef1 A server1.pl
4	23:da:a9:89:27:51	10.0.0.6	00:00:00:00:00:05	10.0.0.5	DNS	86	Standard query response 0x5ef1 A server1.pl A 10.0.0.185
5	00:00:00:00:00:05	10.0.0.5	23:da:a9:89:27:51	10.0.0.6	DNS	70	Standard query 0xf8ec AAAA server1.pl
6	23:da:a9:89:27:51	10.0.0.6	00:00:00:00:00:05	10.0.0.5	DNS	70	Standard query response 0xf8ec AAAA server1.pl
7	00:00:00:00:00:05	00:00:00_00:00:05	ff:ff:ff:ff:ff:ff	Broadcast	ARP	42	Who has 10.0.0.185? Tell 10.0.0.5
8	b1:56:ed:c4:2b:06	b1:56:ed:c4:2b:06	00:00:00:00:00:05	00:00:00_00:00:05	ARP	42	10.0.0.185 is at b1:56:ed:c4:2b:06
9	00:00:00:00:00:05	10.0.0.5	b1:56:ed:c4:2b:06	10.0.0.185	TCP	74	36630 → 8080 [SYN] Seq=0 Win=42340 Len=0 MSS=1460 SACK_P
10	b1:56:ed:c4:2b:06	10.0.0.185	00:00:00:00:00:05	10.0.0.5	TCP	74	8080 → 36630 [SYN, ACK] Seq=0 Ack=1 Win=43440 Len=0 MSS=
11	00:00:00:00:00:05	10.0.0.5	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=1 Ack=1 Win=42496 Len=0 TSval=204
12	00:00:00:00:00:05	10.0.0.5	b1:56:ed:c4:2b:06	10.0.0.185	HTTP	187	GET /plik.bin HTTP/1.1
13	b1:56:ed:c4:2b:06	10.0.0.185	00:00:00:00:00:05	10.0.0.5	TCP	66	8080 → 36630 [ACK] Seq=1 Ack=122 Win=43520 Len=0 TSval=7
14	b1:56:ed:c4:2b:06	10.0.0.185	00:00:00:00:00:05	10.0.0.5	TCP	273	8080 → 36630 [PSH, ACK] Seq=1 Ack=122 Win=43520 Len=207
15	00:00:00:00:00:05	10.0.0.5	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=122 Ack=208 Win=42496 Len=0 TSval

Fig. 11. Traffic observed on the *Agent H5* device



No.	Source	Source	Destination	Destination	Protocol	Length	Info
1	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	TCP	74	36630 → 8080 [SYN] Seq=0
2	b1:56:ed:c4:2b:06	10.0.0.185	d1:58:c0:f7:ae:3c	10.0.0.247	TCP	74	8080 → 36630 [SYN, ACK]
3	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=1
4	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	HTTP	187	GET /plik.bin HTTP/1.1
5	b1:56:ed:c4:2b:06	10.0.0.185	d1:58:c0:f7:ae:3c	10.0.0.247	TCP	66	8080 → 36630 [ACK] Seq=1
6	b1:56:ed:c4:2b:06	10.0.0.185	d1:58:c0:f7:ae:3c	10.0.0.247	TCP	273	8080 → 36630 [PSH, ACK]
7	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=1
8	b1:56:ed:c4:2b:06	10.0.0.185	d1:58:c0:f7:ae:3c	10.0.0.247	TCP	7306	8080 → 36630 [PSH, ACK]
9	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=1
10	b1:56:ed:c4:2b:06	10.0.0.185	d1:58:c0:f7:ae:3c	10.0.0.247	TCP	7306	8080 → 36630 [PSH, ACK]
11	d1:58:c0:f7:ae:3c	10.0.0.247	b1:56:ed:c4:2b:06	10.0.0.185	TCP	66	36630 → 8080 [ACK] Seq=1

Fig. 12. The traffic observed on switch *S5* in state *t0*

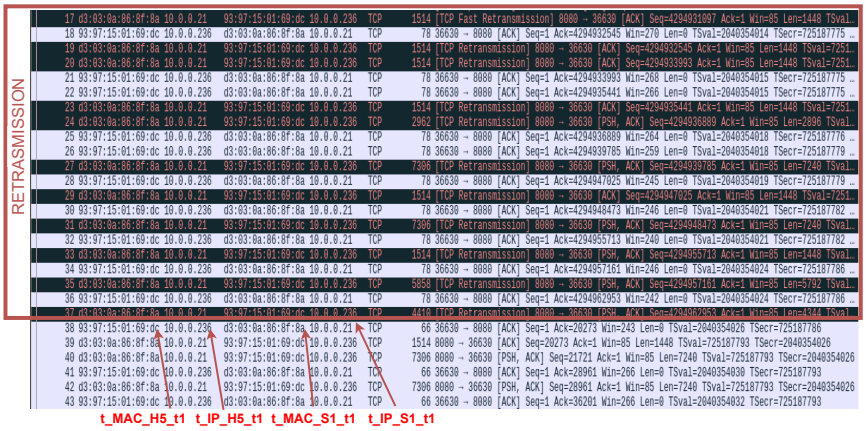


Fig. 13. Traffic observed on switch *S4* in state *t1*

REFERENCES

[1] M. S. Khan, S. Siddiqui, and K. Ferens, "A cognitive and concurrent cyber kill chain model," *Computer and Network Security Essentials*, pp. 585–602, 8 2017. doi: 10.1007/978-3-319-58424-9_34/FIGURES/4

[2] M. Albanese, "From cyber situational awareness to adaptive cyber defense: Leveling the cyber playing field," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11170 LNCS, pp. 1–23, 2018. doi: 10.1007/978-3-030-04834-1_1/FIGURES/12

[3] Fred, R. Lee, A. Acquisti, W. Horne, C. Palmer, A. Ghosh, D. Pendarakis, W. Sanders, E. Fleischman, H. Teufel III, and others Chong, "National cyber leap year summit 2009 co-chairs' report, networking and information technology research and development," 9 2009.

[4] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, and X. S. Wang, *Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats*, 1st ed. Springer Publishing Company, Incorporated, 2011. ISBN 1461409764

[5] L. Jalowski, M. Zmuda, and M. Rawski, "A survey on moving target defense for networks: A practical view," *Electronics 2022, Vol. 11, Page 2886*, vol. 11, p. 2886, 9 2022. doi: 10.3390/ELECTRONICS11182886

[6] E. Al-Shaer, Q. Duan, and J. H. Jafarian, "Random host mutation for moving target defense," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, vol. 106 LNICS, pp. 310–327, 2013. doi: 10.1007/978-3-642-36883-7_19

[7] C. Gudla and A. H. Sung, "Moving target defense discrete host address mutation and analysis in sdn," *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, pp. 55–61, 12 2020. doi: 10.1109/CSCI51800.2020.00017

[8] M. Rawski, "Network topology mutation as moving target defense for corporate networks," *International Journal of Electronics and Telecommunications*, vol. 65, pp. 571–577, 2019. doi: 10.24425/IJET.2019.129814

[9] L. Zhang, Q. Wei, K. Gu, and H. Yuwen, "Path hopping based sdn network defense technology," *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016*, pp. 2058–2063, 10 2016. doi: 10.1109/FSKD.2016.7603498

[10] A. Chowdhary, D. Huang, A. Alshamrani, and H. Liang, "Mtd analysis and evaluation framework in software defined network (mason)," *SDN-NFVSec 2018 - Proceedings of the 2018 ACM International Workshop on Security in Software Defined Networks and Network Function Virtualization, Co-located with CODASPY 2018*, vol. 2018-January, pp. 43–48, 3 2018. doi: 10.1145/3180465.3180473

[11] K. Cabaj, J. Wytrębowicz, S. Kukliński, P. Radziszewski, and K. T. Dinh, "SDN Architecture Impact on Network Security," in *Position Papers of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 3. IEEE/Polish Information Processing Society, 2014. doi: 10.15439/2014F473 pp. 143–148. [Online]. Available: http://dx.doi.org/10.15439/2014F473

[12] E. F. Kfoury, S. Choueiri, A. Mazloum, A. Alsabeh, J. Gomez, and J. Crichigno, "A comprehensive survey on smartnics: Architectures, development models, applications, and research directions," *IEEE Access*, vol. 12, pp. 107 297–107 336, 2024. doi: 10.1109/ACCESS.2024.3437203

[13] J. H. Jafarian, E. Al-Shaer, and Q. Duan, "Openflow random host mutation: transparent moving target defense using software defined networking," in *Proceedings of the First Workshop on Hot Topics in Software Defined Networks*, ser. HotSDN '12. New York, NY, USA: Association for Computing Machinery, 2012. doi: 10.1145/2342441.2342467. ISBN 9781450314770 p. 127–132. [Online]. Available: https://doi.org/10.1145/2342441.2342467

[14] S. Wang, L. Zhang, and C. Tang, "A new dynamic address solution for moving target defense," in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, 2016. doi: 10.1109/ITNEC.2016.7560545 pp. 1149–1152.

[15] A. Almohaimeed and A. Asaduzzaman, "A novel moving target defense technique to secure communication links in software-defined networks," in *2019 Fifth Conference on Mobile and Secure Services (MobiSecServ)*, 2019. doi: 10.1109/MOBISECSERV.2019.8686530 pp. 1–4.

[16] D. C. MacFarland and C. A. Shue, "The sdn shuffle: Creating a moving-target defense using host-based software-defined networking," in *Proceedings of the Second ACM Workshop on Moving Target Defense*, ser. MTD '15. New York, NY, USA: Association for Computing Machinery, 2015. doi: 10.1145/2808475.2808485. ISBN 9781450338233 p. 37–41. [Online]. Available: https://doi.org/10.1145/2808475.2808485

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 155–163
DOI: 10.15439/2025F1225
ISSN 2300-5963 ACSIS, Vol. 43

Towards a Framework for Systematic API Migrations

Nikolas Jíša
0009-0005-1551-2740
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic
Email: jisaniko@fit.cvut.cz

Robert Pergl
0000-0003-2980-4400
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic
Email: robert.pergl@fit.cvut.cz

***Abstract*—This work presents a framework for systematic Application Programming Interface (API) migrations that provides guidance for both manual and automated migration processes. The approach starts by matching methods from the old API version with those of the new version, followed by comparative syntactic and semantic analyses to produce migration steps. As a proof of concept, we implement the most challenging parts using structured, parameterized Artificial Intelligence (AI) queries, which also serve as a basis for evaluation. Our results demonstrate the feasibility and usefulness of this approach. Future work will focus on completing the full migration process while reducing reliance on AI due to its current limitations in reasoning. We also plan to evaluate the framework on real-world APIs to assess its effectiveness and general applicability.**

I. Introduction

IN RECENT years, technological progress has accelerated rapidly. One clear example of this is the widespread integration of the Internet into everyday life, supporting activities such as online shopping, banking, social networking, and information retrieval through Artificial Intelligence (AI) queries [1]. Kurzweil [2] expands Moore's Law by applying the idea of exponential growth not only to hardware but also to software and other technological areas, with the Internet serving as one such example. The Internet consists of many interconnected systems that rely on each other. Some of these systems are client-server applications, which are often built using Application Programming Interfaces (APIs).

The problem we focus on is API evolvability, with a particular emphasis on the challenge of API migration. Lamothe et al. [3] identified several research challenges related to API evolvability. Among these, we are especially interested in the following: **(a)** Combining textual merging with syntactic and semantic approaches, **(b)** Providing a commercially viable API migration solution, **(c)** Incorporating domain-specific information into tools, **(d)** Developing more tools to assist with Web API, **(e)** Automatically identifying factors driving API changes, **(f)** Addressing API semantics and dependencies, and **(g)** Supporting the context sensitivity of API migration tools.

Although the article [3] was published in late 2021, our extended review of the API evolvability literature, presented in section III, shows that many of these challenges continue to be highly relevant.

In this paper, our research objective is: "**To create a framework for systematic API migrations that provides guidelines for performing API migrations, whether fully manual or semi-automated**." Our approach targets scenarios where an application uses a specific version of an API and needs to migrate its API calls to a newer version. A key requirement is to know the method signatures of both the old API version and the new API version, at least at the syntactic level. However, method signatures alone allow for only basic migration guidelines. If the source code of the new API version is available, static analysis can be used to generate more detailed guidance. Furthermore, if semantic documentation artifacts are available, changes at the semantic level—even if the source code remains unchanged—can be incorporated to further improve the quality of the migration guidelines.

Apart from mapping the landscape in the area of API migration research and reporting the progress of our own API migration work, the contribution of this paper lies in presenting a high-level approach to API migration. This approach consists of several components, each of which can be implemented in different ways. Nevertheless, we do not provide implementations here.

The remainder of this paper is structured as follows. Section II presents the research methodology. Related work is discussed in section III. The core of our contribution is detailed in section IV, where we explain our approach to API migration. Section V provides the evaluation of our approach, and section VI provides further discussion. Finally, section VII concludes the paper and outlines directions for future work.

II. Research Methodology

We follow the Design Science Research Methodology (DSRM), first introduced by Hevner et al. [4] and later improved in [5]. DSRM focuses on creating and evaluating new artifacts — such as models, methods, processes, or systems — that aim to solve real-world problems. The key components are: **Environment**, the real-world context including people, organizations, technologies, and problems. It sets the requirements and limits for the research. **Knowledge Base** holds existing theories, methods, frameworks, and tools that support the research. **Design Science Research** is an

iterative process of `Design Cycles` that build and test artifacts to solve problems. It connects to the **Environment** via the `Relevance Cycle` and to the **Knowledge Base** through the `Rigor Cycle`.

This paper builds on our previous work aimed at supporting API usage and migration. We initially explored various API documentation formats, such as OpenAPI Specification (OAS) [6] and API Blueprint (APIB) [7], with the goal of designing a unified format and mapping mechanisms between common specifications. So far, we have implemented a mapping from OAS to APIB [8], with plans to develop a central model and mappings to other formats. In the approach proposed in this paper, one possible input for step **M2.b** is documentation represented in this central model. We also developed an API ontological model presented in [9], which illustrates how changes made by an API provider can impact API consumers. The model includes several key observations, such as the distinction between syntactic and semantic changes, which forms the foundation of the approach presented in this paper.

In this paper, first, we do another `Relevance Cycle` by looking at updates in API documentation and investigating recent breaking changes in open-source APIs. We provide some remarks about this part in section III and section V. Next, we do the `Rigor Cycle` by reviewing new research on API evolvability using snowballing [10], as described in section III. Finally, we do the `Design Cycle` by creating a framework for systematic API migration in section IV and provide evaluation in section V.

III. Related Work

According to Lamothe et al. in [3], the key areas of focus in API evolvability at the time of their work, published at the end of 2021, were addressing breaking API changes, enhancing API usability, and reducing API misuses. In the same paper, Lamothe et al. also identify several open challenges in the field of API evolvability. In this section, we review challenges relevant for this paper and describe some additional research works discovered by snowballing [10]. The challenges are categorized with identifiers from the original article, where "EC" stands for "Existing Challenges" (which have existing publications attempting to find solutions) and "UC" for "Unresolved Challenges" (which lack existing solutions).

EC-1*) Combining textual merging with syntactic and semantic approaches:* This challenge is based on article [11] where the context is to merge two or more versions of the same software together. The article is from year 2002, nevertheless, even today, Version Control System (VCS) tools such as Git [12], are based on textual merging without employing syntactic nor semantic approaches. In a way, we can consider API migration as merge of new version of consumed API with the consumed application itself. Therefore, this can be considered as a relevant challenge for this paper as we also employ syntactic and semantic approaches.

EC-2*) Providing a commercially viable API migration solution:* At the time of writing of [3], there had already been existing API migration solutions such as tool SemDiff [13] or ApiDiff [14]. SemDiff can recommend replacements of API calls for methods which were removed in given API based on static analysis of Abstract Syntax Trees (ASTs) corresponding to code files in VCS. ApiDiff tool is based on static analysis and similarity heuristics of Java libraries hosted on Git [12] repositories in order to detect breaking changes and non-breaking changes between two versions. ApiDiff and SemDiff have not been proven commercially viable at the time of writing of [3]. One of the problems was that, as for most API Evolvability works, there was a lack in area of evaluation, because there had not been established standardized approaches and datasets to evaluate API evolvability research works which would be in wide use [3]. Both the SemDiff and ApiDiff approaches employ static analysis taking into account syntactic information but without utilizing semantic information.

In [15] Deshpande et al. tackle the problem of API migration using multi-objective evolutionary algorithms. Their approach is not restricted to cases where a single source method is always transformed into a single target method (one-to-one mappings). Instead, it is also applicable to scenarios where one or multiple source methods are mapped to multiple target methods (one-to-many and many-to-many mappings).

In [16] Ramos et al. present the MELT system, which extracts API Consumer-side transformations by analyzing pull requests on the API side. This process combines static analysis with Natural Language Processing (NLP) of pull request descriptions and comments.

Some approaches leverage program synthesis to generate transformation procedures by using examples of API call mappings between different versions. Notable examples include the APIFIX tool introduced in [17] and the ReFazer tool introduced in [18].

In [19] Beuer-Kellner et al. propose an API migration approach that utilizes a service to manage the conversion of data structures between different API versions.

In [20], Huang et al. propose an API mapping approach called MATL, which employs transfer learning to automate API mapping without requiring knowledge of the underlying source code of the API involved.

Another approach involves having developers on the API side create transformation scripts to assist API Consumers in updating API calls between versions [3]. Similarly to our goal, the mentioned techniques deal with API migration. There already exist multiple works with employ both static and semantic information in order to automatize API Migration. The idea of our approach is to employ combination of static analysis with analysis of additional semantic information.

EC-3*) Incorporating domain-specific information into tools:* This challenge is based on [11] in which Mens points out that more research is needed for the detection and resolution of structural merge conflicts that arise in the presence of restructuring transformation, and that we need more domain-specific information because it cannot be inferred from code. Our approach also relies on additional information beyond

source code.

***EC-10**) More tools to help with Web APIs:* According to Wittern in [21], Web API providers also control runtime of APIs and can do changes anytime with severe consequences to their API consumers as opposed to Library APIs. Additionally, Web APIs often lack machine-undrestandable specifications, and data are often passed over strings. Our research idea should be applicable to both Web APIs and Library APIs, therefore it should contribute to Web APIs tooling making EC-10 a relevant challenge for our research.

***EC-16**) Dealing with API semantics and dependencies:* In [22] Amman et al. describes following calls to action:

1) *We need precise definition of API usages, considering usage properties such as usage location and call multiplicities*
2) *We need a representation of such usages that captures all code details necessary to distinguish correct usages from misuses and more precise analyzes to identify usages in code*
3) *We need detectors that retrieve sufficiently many usage examples using project-external sources, such as large project sets or code-search engines*
4) *We need detectors that go beyond native assumption that a deviation from the most-frequent usage corresponds to a misuse, but consider program semantics, such as type hierarchies and implicit dependencies between objects. We hypothesize that probabilistic models might be a way to tackle this problem*
5) *We need strategies to properly match patterns and usages in the presence of violations*
6) *We need strategies to properly handle alternative patterns for the same API*
7) *Finally, we need good ranking strategies, to reduce the cost of reviewing findings.*

In the same work Amman et al. present MuBench and MuBenchPipe as foundational tools to support repeatable and replicable studies that enable systematic evaluation and analysis of alternative approaches and strategies in order to move forward with the mentioned calls to action.

Dealing with API semantics is definitely in scope of our research.

***UC-2**) Supporting the context sensitivity of API migration tools:* It is not yet clear how to best approach the context sensitivity in API migration tools. One of the proposed directions is the incorporation of domain-specific information into tools. Our research idea is related to this challenge, because we also consider incorporation of domain-specific information in form of semantic documentation in structured API documentation formats (such as OAS).

IV. Our Approach

Our proposed approach for API migration should serve as a guideline for performing API migration for developers managing consumption of an evolving API in applications - regardless whether it is performed manually or with some automated steps. The most relevant inputs are as follows:

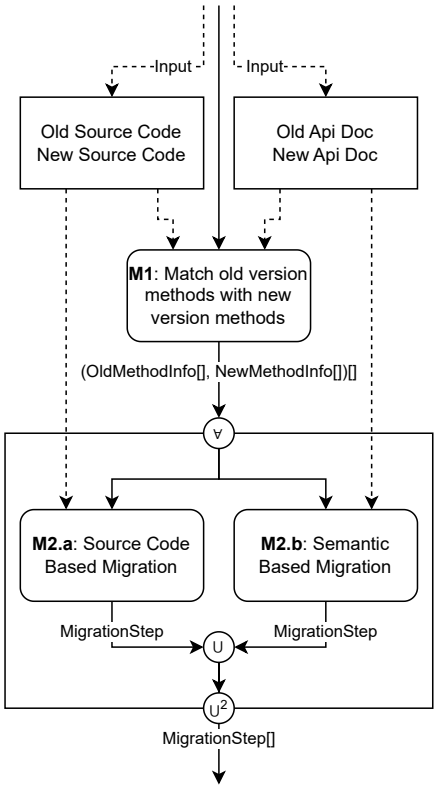


Fig. 1: Overview of our API migration approach

1) Structured API documentation for both the old and the new versions of the consumed API such as OAS documentation. **2)** Links to source code repositories (e.g., in Git), corresponding to both the old and new API versions, when available.

Upon execution, the API migration tool follows a two-step process depicted in fig. 1: **1)** Matching API methods between the old and new API versions (**M1**). **2)** Performing comparative syntax analysis (when source code is available) alongside semantic analysis for each match identified in the previous step, in order to generate a set of *MigrationStep*s (**M2.a** and **M2.b**).

MigrationStep represents a transformation applicable to old-version API method calls, specifying how to adapt them for the new version. The resulting *MigrationStep*s provide developers with a structured foundation to carry out the migration in manual or semi-automated manner.

For clarity, we present a demonstrative example comprising two versions of the same API, shown in example 4.2 and example 4.3. Both versions utilize the `Product` class defined in example 4.1, where the `[Key]` attribute — pro-

Example 4.1 (ASP.NET Core (C#)): Definition of `Product`:

```
class Product : Item
{
  public string Name { get; set; }
  public decimal Price { get; set; }
}
class Item
{
  [Key]public int Id { get; set; }
}
```

vided by Entity Framework Core [23] — designates the `Id` property as the primary key for the database. Additionally, the `dbContext` referenced in the examples corresponds to an instance of `DbContext` [24], which is used to access and interact with the database.

An example of the output from our proposed approach is shown in example 4.4, which migrates old version API calls from example 4.2 to new version API calls from example 4.3. This output is represented as JSON, where each object in the top-level array corresponds to a possible *MigrationStep*. The strings labeled "`RegexMatch`" denote regular expressions, each paired with an associated "`Replacement`" string. Within these replacement strings, "`$1`" refers to the first captured group in the match.

The primary purpose of these examples is to illustrate the concept of inputs and outputs rather than to present fully-fledged input/output artifacts.

M1: Match old version methods with new version methods

The purpose of this step is to identify correspondences between API methods in the old version and the new version of given API. Specifically, it determines which methods in the old version align with which methods in the new version, based on structural and semantic similarity. The output of this step is a collection of tuples `(OldMethodInfo[], NewMethodInfo[])[]`, where each `MethodInfo` instance encapsulates identifying and basic metadata about a given method, such as its name, return type and parameter types.

There are multiple possible strategies for implementing this matching step. We propose implementation in two parts as follows: **M1.1) Exact Signature Match**: Methods that share identical signatures — that is, matching method names, return types, and parameters — in both the old and new versions are matched directly. For example, during the migration from example 4.2 to example 4.3, only the method `DeleteProduct` is matched this way. **M1.2) Remaining Methods Match**: For methods which were not matched by the previous step, several heuristic strategies may be applied: **a)** Manual matching by the developers **b)** Matching based solely on method names **c)** Matching based on parameters and return types **d)** Matching via AST comparison, when source code is available... During the migration from example 4.2 to example 4.3, the methods `CreateProduct` and `UpdateProduct` in the old version are expected to be matched with the method `CreateOrUpdateProduct` in the new version. Similarly, the method `GetAllProducts()` from the old version should correspond to `GetProducts(decimal minPrice)` in the new version. Conversely, the method `GetProduct` from the old version does not have a matching counterpart in the new version.

Methods in the old API version that have no corresponding counterparts in the new version are classified as *deleted methods*, while those introduced only in the new version are considered *added methods*.

It is important to note that, depending on the implementation, the matching algorithm may produce false negatives—failing to correctly associate semantically equivalent methods across versions. In such cases, logically related method pairs might be misclassified as a deletion and an addition, rather than as an evolution of a single method, which could ultimately render our approach non-functional for involved methods. This underscores the necessity of maximizing the effectiveness of step **M1.2**.

Furthermore, the mapping between methods in the old and new versions is not limited to one-to-one or zero-to-one relationships. Many-to-one (M:1), one-to-many (1:N), and even many-to-many (M:N) mappings are possible, as illustrated in migration of `CreateProduct` and `UpdateProduct` in example 4.2 to `CreateOrUpdateProduct` in example 4.3. A baseline implementation that accounts for this could analyze method names by identifying segments such as "And" or "Or", and then checking whether corresponding methods exist for the parts before and after these segments. It should also verify that the parameter and return types match appropriately. A more advanced implementation could incorporate static analysis when the source code is available.

M2.a: Source Code Based Migration

The primary objective of this step is to derive the necessary migration transformations to ensure that the updated version of the API-consuming application remains syntactically correct. Specifically, it aims to determine how each method call to the old API version should be transformed into a corresponding call to the new API version, preserving equivalent application behavior. A key prerequisite for this process is access to the source code of the consumed API.

To illustrate the inputs and outputs, consider migrating method calls from `GetAllProducts()` in example 4.2 to method calls `GetProducts(decimal minPrice)` in example 4.3. This step should establish that calls to `GetAllProducts()` correspond to calls to `GetProducts(decimal.MinValue)` in the new version and generate the appropriate *MigrationStep*, as demonstrated in example 4.4.

There are several possible approaches to implementing this step. Below, we outline three representative strategies:

1) **Manual Transformation**: In this baseline approach, the developer manually inspects the old and new versions of each API method and defines the required transformations manually.

Example 4.2 (ASP.NET Core (C#)): Old version of API:

```
Product? GetProduct(int productId) =>
 ↪  dbContext.Products.FirstOrDefault(p
 ↪  => p.Id == productId);
IEnumerable<Product> GetAllProducts() =>
 ↪  dbContext.Products;
Product? CreateProduct(Product product)
 ↪  {
  dbContext.Products.Add(product);
  dbContext.SaveChanges();
  return product;
}
Product? UpdateProduct(Product product)
 ↪  {
  dbContext.Products.Update(product);
  dbContext.SaveChanges();
  return product;
}
void DeleteProduct(int productId) {
  var product =
   ↪  dbContext.Products.FirstOrDefault(p
   ↪  => p.Id == productId);
  if (product is null)
    return;
  var result =
   ↪  dbContext.Products.Remove(product);
  dbContext.SaveChanges();
}
```

Example 4.3 (ASP.NET Core (C#)): New version of API:

```
IEnumerable<Product> GetProducts(decimal
 ↪  minPrice) =>
 ↪  dbContext.Products.Where(p =>
 ↪  p.Price >= minPrice);

Product? CreateOrUpdateProduct(Product
 ↪  product) {
  if (product.Id == default)
    dbContext.Products.Add(product);
  else
    dbContext.Products.Update(product);
     ↪  ;
  dbContext.SaveChanges();
  return product;
}


void DeleteProduct(int productId) {
  var product =
   ↪  dbContext.Products.FirstOrDefault(p
   ↪  => p.Id == productId);
  if (product is null)
    return;
  var result =
   ↪  dbContext.Products.Remove(product);
  dbContext.SaveChanges();
}
```

Example 4.4 (ASP.NET Core (C#)): Output of our approach:

```
[{
  "OldMethodSignatures": ["Product? CreateProduct(Product product)", "Product?
 ↪  UpdateProduct(Product product)"],
  "NewMethodSignatures": ["Product? CreateOrUpdateProduct(Product product)"],
  "ReplacementRules": [{"RegexMatch": "CreateProduct\\((\w+)\)", "Replacement":
 ↪  "CreateOrUpdateProduct($1)"},
    {"RegexMatch": "UpdateProduct\\((\w+)\)", "Replacement":
    ↪  "CreateOrUpdateProduct($1)"}]
},
{
  "OldMethodSignatures": ["IEnumerable<Product> GetAllProducts()"],
  "NewMethodSignatures": ["IEnumerable<Product> GetProducts(decimal minPrice)"],
  "ReplacementRules": [{"RegexMatch": "GetAllProducts\\(\)", "Replacement":
 ↪  "GetProducts(decimal.MinValue)"}]
}
{
  "OldMethodSignatures": ["Product? GetProduct(int productId)"],
  "ReplacementRules": [{"RegexMatch": "GetProduct\\((\w+)\)", "Replacement":
 ↪  "GetProducts(decimal.MinValue).FirstOrDefault(p => p.Id == $1)"}]
}]
```

2) **Static analysis**: This strategy compares the ASTs of the old and new method implementations. While potentially precise, it poses significant technical challenges due to the complexity of code analysis and the wide variety of programming language constructs.
3) **AI-Based Methods**: Advances in AI, particularly large language models, offer promising opportunities for semi-automated transformation. Although current AI systems have limitations— especially in tasks requiring deep reasoning [25] — our experience with modern tools such as ChatGPT [26] indicates they can effectively address transformation problems, provided the task is well-specified.

***M2.b**: Semantic Based Migration*

The purpose of this step is to derive the migration steps necessary to ensure that the migrated version of the API-consuming application remains semantically correct. This includes preserving the intended meaning of method inputs and outputs.

To illustrate the inputs and outputs of this step, consider a scenario where the `Price` property of a `Product` is represented in US dollars in the older version, but in euros in the newer version. If this change is not properly accounted for, it can cause misinterpretation of price values, leading to errors. Importantly, such semantic differences are usually not detectable from the source code alone, even though they may sometimes be suggested by comments or naming conventions. To address this limitation, we propose extracting semantic metadata from structured documentation artifacts, such as OAS. In OAS, `Product` would be defined as a component schema, and its `Price` property could be annotated using an OAS extension (e.g., `x-semantic-meaning`), with values set to `USD` in the old version and `EUR` in the new version. Example of the `USD`-annotated OAS schema component for `Product` is shown inexample 4.5. The output of our approach would be a set of *MigrationSteps* for each method utilizing the `price` of `Product`. For instance, migrating the `GetAllProducts` method from using `USD` to `EUR` pricing would yield a *MigrationStep* as illustrated in example 4.6.

Example 4.5 (OAS JSON): Product in component schema:

```
{
  "type": "object",
  "properties": {
    "id": {"type": "string"},
    "name": {"type": "string"},
    "price": {"type": "number",
     ↪  "x-semantic-meaning": "USD"}
  }
}
```

Example 4.6 (JSON): MigrationStep for `Price` change from `USD` to `EUR`:

```
{
  "OldMethodSignatures":
   ↪  ["IEnumerable<Product>
   ↪  GetAllProducts()"],
  "NewMethodSignatures":
   ↪  ["IEnumerable<Product>
   ↪  GetAllProducts()"],
  "OldSemanticMeaning": {"Path":
   ↪  "Product.Price", "Value": "USD"},
  "NewSemanticMeaning": {"Path":
   ↪  "Product.Price", "Value": "EUR"}
}
```

A baseline implementation of this step could involve simply checking whether the semantic annotations have changed and, if so, requesting manual intervention. More advanced approaches might attempt to perform semantic migration automatically. However, such techniques are beyond the scope of this paper and represent a promising direction for future research.

V. Evaluation

A. API changes for evaluation

Initially, we aimed to identify an open-source API with a well-documented history of breaking changes to evaluate our approach. However, this task proved more difficult than anticipated. Although we identified some promising candidates, clearly documented and representative breaking changes among real-world open-source APIs were generally scarce. One example of a promising API for evaluation purposes is Elasticsearch [27], which documents several breaking changes - such as the one described and implemented in [28] where our approach would likely be effective. Nevertheless, the complexity of many APIs hindered analysis, especially when changes were undocumented or lacked sufficient context. Therefore, we decided to create our own API changes based on examples 4.2 and 4.3, which are listed in table I.

As our approach is not yet fully implemented, we conducted an evaluation focusing on the implementability and applicability of each individual step. In some cases, this evaluation was supported by leveraging AI tools, specifically ChatGPT [26]. The results of this assessment are summarized in table II.

```
I am consuming old version of API in my
 ↪  application and I have to migrate it to
 ↪  use the new version. The API is written
 ↪  in `{{API_FRAMEWORK_OR_LANGUAGE}}`. I
 ↪  need to know which steps to take in my
 ↪  API-consuming application created in
 ↪  `{{API_CONSUMER_FRAMEWORK_OR_LANGUAGE}}`,
 ↪  specifically, what should I replace the
 ↪  calls to `{{OLD_METHOD_SIGNATURES}}` from
 ↪  old version to migrate it to calls to new
 ↪  `{{NEW_METHOD_SIGNATURES}}` with behavior
 ↪  in terms of input-output mapping being
 ↪  intact? Mark the source code I should use
 ↪  in place of calls to
 ↪  `{{OLD_METHOD_SIGNATURES}}` - i.e. exact
 ↪  string I should use as replacement for
 ↪  `{{OLD_METHOD_CALLS}}` - in the answer
 ↪  between XML tags `<REPLACEMENT>` and
 ↪  `</REPLACEMENT>`.

Definition of relevant structures:
`{{SHARED_DATA_TYPE_DEFINITIONS}}`

Old version source code:
`{{OLD_CODE}}`

New version source code:
`{{NEW_CODE}}`
```

Fig. 2: AI query implementation of **M2.a** step

B. Test implementation of *M2.a* over AI

Our implementation of **M2.a** for AI-based processing is defined by a structured AI query that accepts the following self-explanatory parameters:

1) `API_FRAMEWORK_OR_LANGUAGE`
2) `API_CONSUMER_FRAMEWORK_OR_LANGUAGE`
3) `OLD_METHOD_SIGNATURES`
4) `NEW_METHOD_SIGNATURES`
5) `OLD_METHOD_CALLS`
6) `SHARED_DATA_TYPE_DEFINITIONS`
7) `OLD_CODE`
8) `NEW_CODE`

The structure of this query is illustrated in fig. 2, where parameters are enclosed using double curly braces.

We used the examples of API method changes from table I to create AI queries, which we submitted to ChatGPT using the ChatGPT-4 Turbo engine [26]. The `REPLACEMENT` parts of the generated responses are listed in table III. The responses were generally satisfactory. However, certain changes may require additional manual adjustments. In general, developers may not be aware of which modifications necessitate further intervention, and thus must manually review all changes. This limitation aligns with the intended scope of our approach, which is designed to assist—rather than fully automate—manual API migration at this stage. Nevertheless, minimizing manual effort remains a key objective, representing an important direction for future research. Also, for cases involving the weakening of preconditions or postconditions, the suggested

TABLE I: API changes for evaluation

Identifier	Change Kind	Version A	Version B
C1	Create / Remove Method **C1**		`IEnumerable<Product> GetProductsWithPriceLowerThan(decimal price) => dbContext.Products.Where(p => p.Price < price);`
C2	Add / Remove Parameter	`IEnumerable<Product> GetProducts() => dbContext.Products`	`IEnumerable<Product> GetProducts(decimal minPrice) => dbContext.Products.Where(p => p.Price >= minPrice)`
C3	Change Parameter Type to More / Less Abstract	`Product? UpdateProduct(Product product) { dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`	`Product? UpdateProduct(Item item) { dbContext.Items.Update(item); dbContext.SaveChanges(); return item; }`
C4	Data Coupling to / from Stamp Coupling	`Product? UpdateProduct(Product product) { dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`	`Product? UpdateProduct(int id, string name, decimal price) { var dbProduct = dbContext.Products.First(p => p.Id == id); dbProduct.Name = name; dbProduct.Price = price; dbContext.SaveChanges(); return dbProduct; }`
C5	Stronger / Weaker Pre-Condition with Exception	`Product? UpdateProduct(Product product) { dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`	`Product? UpdateProduct(Product product) { if (product.Price < 0) throw new ArgumentException(); dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`
C6	Stronger / Weaker Pre-Condition with Null	`Product? UpdateProduct(Product product) { dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`	`Product? UpdateProduct(Product product) { if (product.Price < 0) return null; dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`
C7	Stronger / Weaker Post-Condition with Exception	`Product? GetProduct(int productId) => dbContext.Products.First(p => p.Id == productId);`	`Product? GetProduct(int productId) { var result = dbContext.Products.First(p => p.Id == productId); if (result.Price < 0) throw new InvalidOperationException(); return result; }`
C8	Stronger / Weaker Post-Condition with Null	`Product? GetProduct(int productId) => dbContext.Products.First(p => p.Id == productId);`	`Product? GetProduct(int productId) { var result = dbContext.Products.First(p => p.Id == productId); if (result.Price < 0) return null; return result; }`
C9	Method composition / decomposition	`Product? CreateProduct(Product product) { dbContext.Products.Add(product); dbContext.SaveChanges(); return product; } Product? UpdateProduct(Product product) { dbContext.Products.Update(product); dbContext.SaveChanges(); return product; }`	`Product? CreateOrUpdateProduct(Product product) { if (product.Id == default) dbContext.Products.Add(product); else dbContext.Products.Update(product);; dbContext.SaveChanges(); return product; }`
C10	Change Primitive Type from / to Smaller	`Product? GetProduct(int productId) => dbContext.Products.FirstOrDefault(p => p.Id == productId)`	`Product? GetProduct(long productId) => dbContext.Products.FirstOrDefault(p => p.Id == productId)`
C11	Add / Remove Field to / from Parameter	`class Filter { decimal MinPrice { get; set; } decimal MaxPrice { get; set; } } IEnumerable<Product> GetProducts() => dbContext.Products; IEnumerable<Product> GetProducts(Filter filter) { var allProducts = GetProducts(); var result = allProducts.Where(p => p.Price >= filter.MinPrice && p.Price <= filter.MaxPrice); return result; }`	`class Filter { string SearchPhrase { get; set; } decimal MinPrice { get; set; } decimal MaxPrice { get; set; } } IEnumerable<Product> GetProducts() => dbContext.Products; IEnumerable<Product> GetProducts(Filter filter) { var allProducts = GetProducts(); var result = allProducts.Where(p => p.Price >= filter.MinPrice && p.Price <= filter.MaxPrice && p.Name.Contains(filter.SearchPhrase)); return result; }`

TABLE II: Evaluation summary

Step	Comment
M1.1	Matching methods of identical signatures is straightforward and thus excluded from the evaluation.
M1.2	We have confirmed that ChatGPT can successfully perform this step using the example presented in section IV.
M2.a	This step is regarded as the most challenging, and consequently, we implemented it using a structured ChatGPT query described in section V-B, which produced satisfactory results.
M2.b	We have confirmed that ChatGPT can successfully execute this step using the example outlined in section IV.

changes could be undesirable. One possible improvement would be to explicitly request `RegexStrings` with corresponding `Replacement` values, so that they could be directly used in the resulting *MigrationSteps*, as demonstrated in example 4.4.

VI. Discussion

We fulfill the research objective outlined in section I by developing a framework for systematic API migration (manual or semi-automated), which is described in section IV and later evaluated in section V. However, several aspects of our proposed approach require further discussion.

Firstly, it would be valuable to compare our API migration methodology with existing approaches. However, as noted by Lamothe et al. [3], comparing API migration techniques remains an open and unresolved challenge. Additionally, we currently lack a complete implementation of our methodology. Instead, we have implemented and evaluated only individual components using AI queries. It would be interesting to see what a full implementation would look like and how it would perform on real-world APIs.

Secondly, during the evaluation, we repeatedly observed that changes preserving backward compatibility might not always be desirable. This is particularly true for modifications that weaken preconditions or postconditions. These observations prompt an important question: how should we determine

TABLE III: Summary of AI responses

Change Identifier	Direction	Response part between "<REPLACEMENT>" and "</REPLACEMENT>"	Manual steps needed?	Comment
C1	A to B	N/A	No	No changes necessary
C1	B to A	N/A	Yes	Injecting code of the removed method might work
C2	A to B	`GetProducts(0)`	No	`decimal.MinValue` would be better than `0`
C2	B to A	`GetProducts().Where(p => p.Price >= minPrice)`	Yes	There needs to be `minPrice` variable
C3	A to B	`(Product?)UpdateProduct(product)`	No	
C3	B to A	`UpdateProduct((Product)item)`	No	
C4	A to B	`UpdateProduct(product.Id, product.Name, product.Price)`	No	
C4	B to A	`new Product { Id = id, Name = name, Price = price }`	No	
C5	A to B	Same as before	Yes	Exception handling might be desired

C5	B to A	`if (product.Price < 0)`
		`  throw new ArgumentException();`
		`UpdateProduct(product)`
			Yes	Keeping the same behavior as before might be undesired here

C6	A to B	Same as before	Yes	Null handling might be desired
C6	B to A	`product.Price < 0 ? null : UpdateProduct(product)`	Yes	Keeping the same behavior as before might be undesired here

C7	A to B	`try {`
		`  GetProduct(productId);`
		`} catch (InvalidOperationException) {`
		`  return null;`
		`}`
			Yes	Should have proper exception handling

C7	B to A	`var product = GetProduct(productId);`
		`if (product.Price < 0)`
		`  throw new InvalidOperationException();`
			Yes	Keeping the same behavior as before might be undesired here

C8	A to B	`var product = GetProduct(productId);`
		`if (product == null)`
		`  throw new InvalidOperationException("Product price is invalid");`
			Yes	Throwing an exception might be undesired

C8	B to A	`var product = GetProduct(productId);`
		`if (product.Price < 0)`
		`  return null;`
			Yes	Keeping the same behavior as before might be undesired here

C9	A to B	`CreateOrUpdateProduct(product);`	No	
C9	B to A	`(product.Id == default ? CreateProduct(product) : UpdateProduct(product));`	No	
C10	A to B	`GetProduct((long)productId)`	No	
C10	B to A	`GetProduct((int)productId);`	Yes	There might be some data loss like this
C11	A to B	`GetProducts(new Filter { MinPrice = filter.MinPrice, MaxPrice =` `↪  filter.MaxPrice }).Where(p => p.Name.Contains(filter.SearchPhrase))`	No	
C11	B to A	`GetProducts(new Filter { MinPrice = filter.MinPrice, MaxPrice =` `↪  filter.MaxPrice, SearchPhrase = "" })`	No	

which changes are desirable and which are not? Moreover, is it possible to evaluate the desirability of such changes in an automated or semi-automated manner? These questions remain open for future research.

Thirdly, several steps in our approach are defined at a high level of abstraction, and it remains uncertain whether all of them can be feasibly implemented. For instance, step **M2.a** is intended to leverage static analysis, potentially supported by AI tools. However, the precise mechanism and its practical efficiency are still unclear. Furthermore, methods involving AI introduce additional challenges related to effectiveness and reproducibility, due to their inherently black-box nature.

Next, although our approach draws on many ideas from the research reviewed in section III, we have done little work on integrating it with existing solutions. This remains an interesting direction for future research.

Finally, since we successfully demonstrated the applicability of AI in step **M2.a**, a natural question arises—could an AI-based approach be employed for the entire API migration process, leveraging source code, structured API documentation, and other related artifacts? While we have not yet explored this possibility in detail, we hypothesize that such an approach could be feasible to some extent. Nonetheless, considering the limitations of current AI systems, particularly their challenges with consistent and transparent reasoning, we propose a more cautious direction. In future work, we aim to reduce reliance on AI and instead focus on more deterministic and interpretable alternatives.

VII. Conclusion

In this work, we presented a systematic framework for API migrations. The goal of this framework is to guide developers during migration, whether they choose to do it manually or with the help of automated tools. The migration process starts by matching methods from the old version of the API with those in the new version. It then continues with both comparative syntactic and semantic analyses to produce clear and structured migration steps.

As part of our evaluation, we implemented the comparative syntactic analysis as a structured AI query. We observed that this step was the most technically challenging part of the process. Nevertheless, our working prototype shows that it is not only feasible but also valuable in practice. It lays a solid foundation for future improvements of the framework.

In future work, we plan to complete the full implementation of the migration process. One of our aims is to reduce reliance on AI, since current models still struggle with consistent and reliable reasoning across different migration scenarios.

Another important direction for future work is to apply our framework to real-world APIs, rather than relying only on synthetic examples. This will allow us to evaluate how well the framework performs in practical environments and to what extent it can be generalized to different kinds of APIs.

Statement on the use of AI

AI technologies (ChatGPT [26]) were used to improve the language of the paper.

Acknowledgements

This research was supported by the grant of Czech Technical University in Prague No. SGS23/206/OHK3/3T/18.

References

[1] L. Rainie and B. Wellman, "The Internet in Daily Life: The Turn to Networked Individualism," in *Society and the Internet*. Oxford University Press, Jul. 2019, pp. 27–42. ISBN 978-0-19-884349-8 978-0-19-187932-6. [Online]. Available: https://academic.oup.com/book/35088/chapter/299127482

[2] R. Kurzweil, "The Law of Accelerating Returns," in *Alan Turing: Life and Legacy of a Great Thinker*, C. Teuscher, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 381–416. ISBN 978-3-642-05744-1 978-3-662-05642-4. [Online]. Available: http://link.springer.com/10.1007/978-3-662-05642-4_16

[3] M. Lamothe, Y.-G. Guéhéneuc, and W. Shang, "A Systematic Review of API Evolution Literature," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–36, Nov. 2022. doi: 10.1145/3470133. [Online]. Available: https://dl.acm.org/doi/10.1145/3470133

[4] Hevner, March, Park, and Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004. doi: 10.2307/25148625. [Online]. Available: https://www.jstor.org/stable/10.2307/25148625

[5] A. Hevner, "A Three Cycle View of Design Science Research," *Scandinavian Journal of Information Systems*, vol. 19, Jan. 2007.

[6] "OpenAPI Specification v3.1.1," accessed: 2025-06-26. [Online]. Available: https://spec.openapis.org/oas/latest.html

[7] "API Blueprint Specification | API Blueprint," accessed: 2025-06-26. [Online]. Available: https://apiblueprint.org/documentation/specification.html

[8] R. Pergl and N. Jísa, "Semantic Analysis of API Blueprint and OpenAPI Specification," in *Czech Technical University Prague*, A. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. Poniszewska-Maranda, Eds., vol. 989, 2024. doi: 10.1007/978-3-031-60227-6_15. ISBN 2367-3370 pp. 172–181.

[9] N. Jíša and R. Pergl, "Towards Evolvable APIs through Ontological Analysis," in *Annals of Computer Science and Information Systems*, vol. 41. PTI, Nov. 2024. doi: 10.15439/2024f3164. ISSN 2300-5963 pp. 61–68.

[10] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, ser. Ease '14. London, England, United Kingdom and New York, NY, USA: Association for Computing Machinery, 2014. doi: 10.1145/2601248.2601268. ISBN 978-1-4503-2476-2

[11] T. Mens, "A state-of-the-art survey on software merging," *IEEE Transactions on Software Engineering*, vol. 28, no. 5, pp. 449–462, May 2002. doi: 10.1109/TSE.2002.1000449. [Online]. Available: http://ieeexplore.ieee.org/document/1000449/

[12] "Git," accessed: 2025-06-28. [Online]. Available: https://git-scm.com/

[13] B. Dagenais and M. P. Robillard, "SemDiff: Analysis and recommendation support for API evolution," in *2009 IEEE 31st International Conference on Software Engineering*. Vancouver, BC, Canada: IEEE, 2009. doi: 10.1109/ICSE.2009.5070565. ISBN 978-1-4244-3453-4 pp. 599–602. [Online]. Available: http://ieeexplore.ieee.org/document/5070565/

[14] A. Brito, L. Xavier, A. Hora, and M. T. Valente, "APIDiff: Detecting API breaking changes," in *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. Campobasso: IEEE, Mar. 2018. doi: 10.1109/SANER.2018.8330249. ISBN 978-1-5386-4969-5 pp. 507–511. [Online]. Available: http://ieeexplore.ieee.org/document/8330249/

[15] N. Deshpande, M. W. Mkaouer, A. Ouni, and N. Sharma, "Third-party software library migration at the method-level using multi-objective evolutionary search," *Swarm and Evolutionary Computation*, vol. 84, p. 101444, Feb. 2024. doi: 10.1016/j.swevo.2023.101444. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S221065022300216X

[16] D. Ramos, H. Mitchell, I. Lynce, V. Manquinho, R. Martins, and C. L. Goues, "MELT: Mining Effective Lightweight Transformations from Pull Requests," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. Luxembourg, Luxembourg: IEEE, Sep. 2023. doi: 10.1109/ASE56229.2023.00117. ISBN 979-8-3503-2996-4 pp. 1516–1528. [Online]. Available: https://ieeexplore.ieee.org/document/10298355/

[17] X. Gao, A. Radhakrishna, G. Soares, R. Shariffdeen, S. Gulwani, and A. Roychoudhury, "APIfix: output-oriented program synthesis for combating breaking changes in libraries," *Proceedings of the ACM on Programming Languages*, vol. 5, no. OOPSLA, pp. 1–27, 2021, publisher: ACM New York, NY, USA.

[18] R. Rolim, G. Soares, L. D'Antoni, O. Polozov, S. Gulwani, R. Gheyi, R. Suzuki, and B. Hartmann, "Learning Syntactic Program Transformations from Examples," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*. Buenos Aires: IEEE, May 2017. doi: 10.1109/ICSE.2017.44. ISBN 978-1-5386-3868-2 pp. 404–415. [Online]. Available: http://ieeexplore.ieee.org/document/7985680/

[19] L. Beurer-Kellner, J. Von Pilgrim, C. Tsigkanos, and T. Kehrer, "A Transformational Approach to Managing Data Model Evolution of Web Services," *IEEE Transactions on Services Computing*, pp. 1–1, 2022. doi: 10.1109/TSC.2022.3144613. [Online]. Available: https://ieeexplore.ieee.org/document/9689952/

[20] Z. Huang, J. Chen, J. Jiang, Y. Liang, H. You, and F. Li, "Mapping APIs in Dynamic-typed Programs by Leveraging Transfer Learning," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 4, pp. 1–29, May 2024. doi: 10.1145/3641848. [Online]. Available: https://dl.acm.org/doi/10.1145/3641848

[21] E. Wittern, "Web APIs - challenges, design points, and research opportunities: invited talk at the 2nd international workshop on API usage and evolution (WAPI '18)," in *Proceedings of the 2nd International Workshop on API Usage and Evolution*. Gothenburg Sweden: ACM, Jun. 2018. doi: 10.1145/3194793.3194801. ISBN 978-1-4503-5754-8 pp. 18–18. [Online]. Available: https://dl.acm.org/doi/10.1145/3194793.3194801

[22] S. Amann, H. Nguyen, S. Nadi, T. Nguyen, and M. Mezini, "A Systematic Evaluation of Static API-Misuse Detectors," *IEEE TRANSACTIONS ON SOFTWARE ENGINEERING*, vol. 45, no. 12, pp. 1170–1188, Dec. 2019. doi: 10.1109/TSE.2018.2827384

[23] "Overview of Entity Framework Core - EF Core," accessed: 2025-07-11. [Online]. Available: https://learn.microsoft.com/en-us/ef/core/

[24] "DbContext Class (Microsoft.EntityFrameworkCore)," accessed: 2025-07-11. [Online]. Available: https://learn.microsoft.com/en-us/dotnet/api/microsoft.entityframeworkcore.dbcontext?view=efcore-9.0

[25] G. Gendron, Q. Bao, M. Witbrock, and G. Dobbie, "Large Language Models Are Not Strong Abstract Reasoners," in *IJCAI Int. Joint Conf. Artif. Intell.*, Larson K., Ed. International Joint Conferences on Artificial Intelligence, 2024. ISBN 10450823 (ISSN); 978-195679204-1 (ISBN) pp. 6270–6278, journal Abbreviation: IJCAI Int. Joint Conf. Artif. Intell. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204293915&partnerID=40&md5=b5ebdcbbf62f13f3c21b695461a5ab2d

[26] "ChatGPT," accessed: 2025-07-11. [Online]. Available: https://chatgpt.com

[27] "elastic/elasticsearch," Jul. 2025, accessed: 2025-07-11. [Online]. Available: https://github.com/elastic/elasticsearch

[28] elastic, "Store outcome values in servicemetrics 'transaction.success_count' by carsonip · Pull Request #9791 · elastic/apm-server," accessed: 2025-07-11. [Online]. Available: https://github.com/elastic/apm-server/pull/9791

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 165–174
DOI: 10.15439/2025F5420
ISSN 2300-5963 ACSIS, Vol. 43

Toward Conversational Decision Support Systems: Integrating LLMs in the Operations Research Methodology

Mariusz Kaleta
0000-0002-2225-8956
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: mariusz.kaleta@pw.edu.pl

***Abstract*—This paper introduces the concept of Conversational Decision Support Systems (C-DSS)—a novel, agent-based framework that leverages Large Language Models (LLMs) to enhance the Operations Research (OR) methodology. We focus on the modeling and coding stages of decision support systems, where language-based interaction is crucial. The paper evaluates the effectiveness of LLMs in generating mathematical models and AMPL code for a curated set of 20 LP/MILP artifacts. Four architectural setups are analyzed: a monolithic LLM agent (M/C), its enhancement with a code verifier (M/C+V), agent-based decomposition with RAG-enhanced coding (M+C^R+V), and full specialization with RAG-enhanced modeling and coding (M^R+C^R+V). Experimental results on two benchmark problems reveal that the targeted retrieval-augmented generation technique (RAG) significantly improves performance for complex modeling patterns such as piecewise functions, indicator constraints, and nested logic. We also propose a broader vision of C-DSS as a multi-agent ecosystem—including agents for visualization, explanation, verification, and orchestration—suggesting a path toward more explainable, adaptable, and intelligent decision support systems.**

I. Introduction

A. Backgrounds

OPERATIONS Research (OR) is grounded in analytical methods applied to improve decision-making. OR projects are often characterized by high complexity and a substantial risk of failure. Similar to software projects, they require the development and deployment of software; however, they also introduce a number of additional challenges. These include the need for a deeper understanding of the problem domain, a thorough comprehension of the decision-makers' goals and intentions, the construction of mathematical models, the selection or development of computational algorithms, and uncertainty regarding the quality of the resulting solutions. To address these challenges, OR methodology typically relies on a structured and systematic approach [1]. The individual steps of this methodology are illustrated in Figure 1.

The methodology begins with the real-world system that is the object of the decision-making process. The first step is to identify the decision problem. Next, the scope of the implementation is defined—this does not necessarily have to fully cover the previously identified area. The subsequent steps



Fig. 1. Operations research methodology enhanced with communication and potential languages used at various stages of the methodology

involve constructing a mathematical model and developing a solution method. The resulting solution is then tested using real data. If these tests are successful, the optimization results can be incorporated into actual decision-making processes, leading to full implementation and real-world impact—thus completing the methodology cycle. In practice, however, this process is rarely linear or strictly cyclical; backtracking or skipping steps may occur at various stages.

Effective communication is critical to the success of an OR project, and both its form and language vary across different phases of the methodology. Figure 1 enriches the classical OR methodology with examples of the communication modes used at each stage. In the early phases, interactions among domain experts dominate—for example, requirement elicitation and project scoping are typically conducted through expert meetings. During the modeling and method development

phases, mathematical and algorithmic languages are primarily used. Implementation requires programming languages, while solution validation often involves further expert consultations, sometimes supported by statistical language. Finally, the results must be communicated to end users via a human-machine interface, such as a dedicated GUI.

Large Language Models (LLMs) have emerged as transformative tools for effective communication across a wide range of domains. By leveraging vast amounts of linguistic data and advanced machine learning techniques, LLMs are capable of understanding context, generating coherent text, and engaging in meaningful dialogue. These capabilities enable them to assist with tasks such as summarizing information, translating languages, drafting content, and facilitating real-time interaction between users. Their ability to perform logical reasoning and apply mathematics to solve logical problems is rapidly advancing. In the domain of programming, they have become significant tools for automating source code generation and testing.

These capabilities make LLMs promising candidates for supporting the OR methodology at various stages. Their potential can be utilized in system discovery and requirement elicitation. They are naturally suited to support modeling and coding. Moreover, areas such as test automation, results explanation and visualization, or GUI generation are also promising fields where LLMs may play a significant role.

Recognizing these possibilities leads to the concept of Conversational Decision Support Systems (C-DSS)—a system in which an LLM serves as a proxy across multiple stages of Decision Support System (DSS) development and operation. The C-DSS concept touches on various aspects of building and operating decision support systems and opens up broad opportunities for future research.

In this paper, we introduce the C-DSS concept. However, due to the subject's complexity and scope, we focus specifically on the modeling and programming stages, and investigate how the effectiveness of language models can be enhanced at these steps. This work is a contribution to a broader vision—one that may significantly advance the field of decision support systems.

B. State of the art

Efforts to automate the process of solving optimization problems began gaining traction in 2022. One of the earliest contributions came from Ramamonjison et al. [2], who proposed a system that automatically formulates mathematical models from natural language descriptions. To evaluate their approach, they introduced the Linear Programming Word Problems (LPWP) dataset.

Building on this work, the same authors organized the NL4Opt Competition [3], which aimed to identify the most effective techniques for automated mathematical model formulation. The competition was based on the LPWP dataset—hereafter referred to as NL4OPT. In their conclusions, the organizers noted that solutions utilizing Large Language Models (LLMs) outperformed all other approaches.

Chen et al. introduced a tool named OptiChat that employs LLMs in conjunction with multiple agents, solvers (e.g., Gurobi, Mosek), and the algebraic modeling language Pyomo to support mathematical modeling [4]. Its core functionality includes detecting issues in model formulation and offering corrective suggestions to the user. At around the same time, Yang et al. proposed OPRO [5], a solution that bypasses traditional solvers by leveraging LLMs directly for optimization tasks. OPRO iteratively refines its solutions based on previous outputs and newly generated meta-feedback. However, this approach has so far only been tested on relatively simple problems like the traveling salesman problem and linear regression, leaving its applicability to more complex scenarios an open question.

Li et al. extended the research to Mixed-Integer Linear Programming (MILP) problems [6]. Their work expands the NL4OPT dataset and focuses on training LLMs to produce correct mathematical formulations, including not just quantitative constraints (e.g., resources, demand), but also logical constraints, which were largely neglected in earlier research. The issue of the simplicity of problems in data sets has been also addressed by Ziyang et al. [7]. The authors proposed the ComplexOR dataset, which features more challenging optimization tasks. Alongside the dataset, the authors introduced a multi-agent system implementing the idea of Chain-of-Experts framework. Although, they report performance improvement, there is no analysis on which problem aspects are hard and most problematic for LLMs.

Beyond the simplicity of tasks, the lack of real-world relevance was another concern. This gap was addressed by the creators of CoPilot [8], who focused on practical business scenarios. Their system guides users through understanding the problem, formulating the model, and solving the optimization task—once again, leveraging LLMs as the core engine.

One of the first comprehensive solutions attempting to address the entire modeling pipeline is OptiMUS [9]. This system not only formulates the optimization model but also generates executable solver code to find and return optimal solutions. Alongside the system, the authors introduced a new dataset—NLP4LP—containing both LP and MILP tasks to support their experiments. However, the limitation of OptiMUS was the need for users to prepare problem descriptions in a specific format. This issue has been addressed in a follow-up study [10], where the authors enhanced OptiMUS to automatically convert natural language inputs into the required structured format.

A significant challenge across all of the above efforts is the variability in both solver performance and LLM behavior. This issue is the focus of LM4OPT [11], a project that seeks to standardize evaluation procedures for such systems. Additionally, the researchers investigate the benefits of fine-tuning LLMs, including compact models like Llama-7B, to improve efficiency.

C. Identified gaps and our contribution

As the literature review indicates, this research area is relatively new, and current studies focus primarily on the use of LLMs for modeling and coding optimization problems, often overlooking the broader perspective of the OR methodology. Even within this narrow scope, there is currently no established knowledge regarding optimal architectures or methods for integrating LLMs into the OR process. With the exception of the limited work by Li et al. [6], there is a lack of analysis and identification which modeling artifacts pose the greatest challenges for language models. Such insights could help address specific issues and improve the effectiveness of automated modeling and coding efforts. Additionally, existing work mostly centers around the use of general-purpose programming languages, which rely on libraries for solving optimization problems. In particular, there is a notable absence of research evaluating the effectiveness of modeling using dedicated declarative languages, such as AMPL (A Mathematical Programming Language) [12]. Most of the research focuses on imperative languages, mainly Python, while optimization problems are natively defined in a declarative way.

Our contribution can be summarized as follows:

- We provide a broader context for mathematical modeling and coding by introducing the concept of Conversational Decision Support Systems (C-DSS) and taking a more comprehensive view of the OR methodology.
- We define a catalog of common artifacts in LP and MILP problems, along with associated sets of textual optimization problem descriptions.
- For the identified artifacts, we analyze the performance of LLMs within various agent-based architectures.
- Unlike previous studies, we employ LLMs to encode problems using the AMPL mathematical modeling language.
- We briefly demonstrate that LLM capabilities are also relevant at other stages of the OR methodology, opening opportunities for further research within the OR community.

II. Methods

A. Modeling artifacts

Based on practices in mathematical modeling, we have identified 20 typical modeling artifacts, summarized in Table I. The first five pertain to linear programming (LP) models and include: a simple LP model (artifact 1), minimization of the maximum value (artifact 2), a weighted sum of absolute values with weights a and b depending on the sign of the argument (absolute value in the objective, artifact 3), the absolute value in a constraint (artifact 4), and a piecewise linear approximation of a convex function $f_i(x)$ (artifact 5). By the basic LP model (artifact 1), we mean a situation where all variables directly represent the problem described in natural language, i.e., there is no need for intermediate variables.

The second group of artifacts concerns mixed-integer linear programming (MILP) problems and includes standard logical constraints typical for binary variables. These constraints typically involve the selection of at most/at least/exactly N elements (artifacts 6–8) and a range of conditional relationships (artifacts 9–14), in which binary variables may or must take certain values if other binary variables satisfy a given Boolean condition.

TABLE I
Modeling artifacts

LP	1	simple LP	no implicit variables
	2	min max	$\min x$
	3	absolute value in objective	$\min(ax^+ + bx^-)$
	4	absolute value in constraint	$\min(ax^+), x^- \leq X^-$
	5	easy linearization	$\min x, x \geq f_i(x)$
MILP (logic constr.)	6	at most N of x, y, z,...	$x + y + z + \cdots \leq N$
	7	at least N of x, y, z,...	$x + y + z + \cdots \geq N$
	8	exactly N of x, y, z,...	$x + y + z + \cdots = N$
	9	if x then y	$x \geq y$
	10	if x then not y	$x + y \leq 1$
	11	if x then y and z	$y \geq x, z \geq x$
	12	if y or z then x	$x \geq y, x \geq z$
	13	if y and z then x	$x \geq \frac{y+z}{2}$
	14	if M or more from x, y, z then v	$v \geq \frac{x+y+\ldots M+1}{N-M+1}$
MILP	15	min min	$\min \min f_i(x)$
	16	disjunction	$x \in \langle a; b\rangle \cup \langle b; d\rangle$
	17	min. activity level	$x = 0$ or $x \in \langle a, b\rangle$
	18	indicator constraint	$x \to C, x \in \{0, 1\}$
	19	concave linearization	$\min f(x), f(x)$ concave
	20	non-monotonic linearization	$\min f(x), f(x)$ non-mon.

The third group of artifacts involves more complex MILP models, which usually include non-convexities or discontinuities. This group comprises minimization of the minimum value (artifact 15), disjoint variable domains (artifact 16), minimum activity level (artifact 17), indicator constraint (activation of a linear constraint C based on a binary variable x, artifact 18), linearization of a concave function (artifact 19), and linearization of a general non-monotonic function (artifact 20).

B. Datasets

To evaluate which artifacts a large language model handles better or worse, we prepared two base problems: problem #64 from the NL4Opt dataset (the pharmaceutical paste problem, with minor linguistic adjustments) and our own original problem (the sewage discharging problem), which has never been published online. Each of the base problems represents artifact 1 — a simple LP problem. Subsequently, the content of each base problem was modified to sequentially introduce artifacts 2 through 20.

Although the dataset may seem small, it is important to note that we currently lack both the capability to automatically generate problem statements in a way that ensures the inclusion of specific artifacts, and methods for automatically verifying the correctness of models proposed by LLMs for given problem descriptions. Modifying problem statements to explicitly induce a given artifact is non-trivial and was carried out manually. Similarly, each solution in the form of a mathematical model obtained during the study was manually verified by an expert in operations research. Given the 20

artifacts, four agent configurations discussed later in the paper, and five runs for each problem and configuration, this results in 800 models whose correctness had to be manually verified.

The content of both problems is presented in Appendix A.

C. Environment settings

The conducted experiments cover four distinct language-agent environments, differing in the number of agents involved and their assigned roles. Each experimental setup processes a set of 40 test tasks, composed of two base problems, each instantiated in 20 variants corresponding to the predefined artifacts. The goal of each environment is to generate a mathematical model in textual form as well as a corresponding AMPL model.

The mathematical model (in text form) includes the definition and description of sets, parameters, and variables, along with the objective function and constraints, formulated in the style typical for LP/MILP problems. For the AMPL model, the system is allowed to produce either a single combined file containing both the model and data, or separate files—‘.mod‘ for the model and ‘.dat‘ for the problem data.

Since the experimental output consists of two distinct results—a textual mathematical model and AMPL source code—errors may occur independently in each. Therefore, we report separate success rates for modeling and coding.

Modeling success is defined as the correct formulation of the optimization problem in textual form, manually verified by an operations research (OR) expert. Any identified flaw in the model leads to its classification as incorrect. The ratio of correct mathematical models is reported in the result tables under the column labeled "Model."

Coding success is defined as the correct encoding of the previously generated mathematical model into AMPL syntax. Correctness was verified both manually by an OR expert and by executing the model in the AMPL environment. A model is considered correctly implemented if it runs without errors and matches the expected logic, regardless of whether the original formulation was valid. Hence, a syntactically correct AMPL implementation of an incorrect model still counts as a success in the coding phase. The ratio of correct AMPL models is reported in the column labeled "Code."

Overall success requires both a valid mathematical model and its correct AMPL implementation, i.e., success in both the modeling and coding stages. The ratio of fully correct and functional solutions is reported in the column labeled "Success."

All experiments were conducted using the GPT-4o model. For each combination of base problem, artifact, and agent configuration, the generation and evaluation process was repeated five times.

III. Results

A. Monolith architecture (M/C)

In the monolith architecture, we use an LLM as both modeler and coder (M/C) in a single call. The architecture is illustrated in Figure 2. The LLM acts as a single agent responsible for both modeling and coding. As input, it receives a word problem inserted into the following zero-shot prompt with Chain-of-Thought (CoT) reasoning enabled by default in GPT-4o:

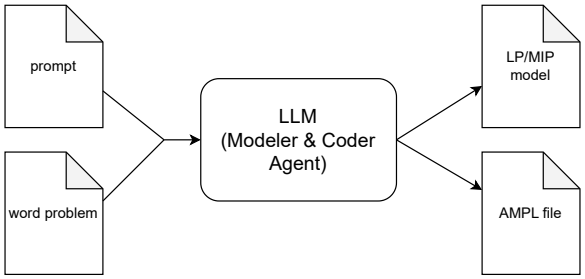


Fig. 2. Monolith architecture (M/C)

> *Formulate a linear programming model for the following problem.*
> ...
> *Provide the model as an AMPL file.*

Results are presented in Table II. Simple LP problems—artifacts 1 to 3—show a high success rate, reaching nearly 100% (only one coding failure occurred). These problems are considered easy for language models in both modeling and coding. The presence of an absolute value in a constraint (artifact 4) reduces the modeling success rate to 60%, though the coding success rate remains relatively high (80%). Artifact 5, which requires a linearization of a piecewise convex function, posed a significant challenge to the LLM and resulted in a very low modeling success rate (10%).

In the second group of artifacts, we observe high modeling success rates for most logical constraints, alongside surprisingly low coding success rates. Interestingly, artifact 13 was particularly challenging; the LLM often interpreted it as artifact 12. We believe this may be due to a linguistic ambiguity, as many people tend to confuse "and" with "or" in everyday language. Artifact 14 also resulted in a relatively low modeling success rate.

In the third group of artifacts, only the minimum activity level (artifact 17) yielded a high success rate. Other artifacts significantly reduced the LLM's ability to model correctly, with three artifacts never resulting in a valid model. However, this group showed a comparatively higher coding success rate than the second group.

Since the obtained results were not satisfactory, we considered several possible directions that may lead to improvements:

1) Fine tuning
2) Special prompting technique, including:
   - Tree-of-thoughts, Graph-of-thoughts
   - Generated knowledge – first get more knowledge about the problem, then model

TABLE II
RESULTS FOR M/C ARCHITECTURE

		Variant	Model	Code	Success
LP	1	simple LP	100%	100%	100%
	2	min max	100%	100%	100%
	3	abs. val. in obj.	100%	90%	90%
	4	abs. val. in constr.	60%	80%	40%
	5	piece-wise fun.	10%	70%	10%
MILP (logic)	6	at most N	100%	50%	50%
	7	at least N	100%	50%	50%
	8	exactly N	80%	40%	40%
	9	$x \Longrightarrow y$	90%	40%	40%
	10	$x \Longrightarrow \neg y$	100%	60%	60%
	11	$x \Longrightarrow y \wedge z$	100%	40%	40%
	12	$y \vee z \Longrightarrow x$	100%	50%	50%
	13	$y \wedge z \Longrightarrow x$	20%	40%	10%
	14	if $\geq M$ from $y \Longrightarrow y, z, \ldots$	50%	40%	30%
MILP	15	min min	0%	100%	0%
	16	disjun.	50%	90%	50%
	17	min. activ. level	100%	90%	90%
	18	indic. constr.	20%	80%	20%
	19	concave linear.	0%	50%	0%
	20	non-monoton. linear.	0%	50%	0%

- Cognitive verifier
- Few-shot – providing examples
- Persona/audiance pattern
- Reflection
- Chain-of-Experts
- ...

3) Agent-based decomposition (mixture-of-experts)
4) Retrieval Augmented Generation (RAG)

For further research, we exclude fine-tuning, as it requires additional examples that are difficult to generate in this domain and may be computationally expensive. We also do not include any special prompting techniques, as it remains unclear which approaches could reliably improve modeling and coding quality. Our observations suggest that for different artifacts, either modeling or coding may be the more challenging task. Following this insight, we believe that decomposition into specialized agents—potentially equipped with additional context and expert knowledge and implemented as RAG—is the most promising solution. The next three environment settings follow this concept.

B. Enhancing M/C with Code Verifier agent

In this setting, we build on the observation that the coding success rate is relatively low, particularly for the second group of artifacts, and that most failures are due to simple syntax issues. We introduced a new component—the verifier agent—responsible for validating the AMPL source file by executing it in the AMPL environment and reporting any errors encountered. The resulting architecture is illustrated in Figure 3.

If an error occurs, the verifier sends the following prompt, including the error message, back to the Modeler/Coder (M/C) agent:

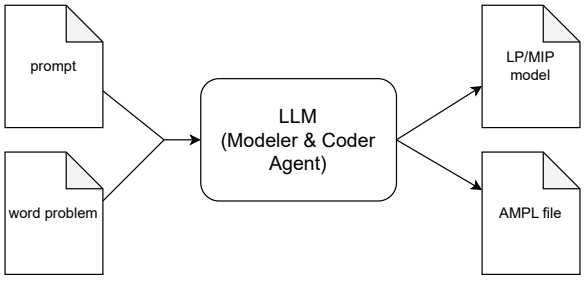


Fig. 3. M/C architecture enhanced with verifier agent (M/C+V)

TABLE III
RESULTS FOR M/C+V ARCHITECTURE

		Variant	Model	Code	Success
LP	1	simple LP	100%	100%	100%
	2	min max	100%	100%	100%
	3	abs. val. in obj.	100%	**100%**	**100%**
	4	abs. val. in constr.	60%	**90%**	**50%**
	5	piece-wise fun.	10%	**80%**	10%
MILP (logic)	6	at most N	100%	**60%**	**60%**
	7	at least N	100%	**50%**	**50%**
	8	exactly N	80%	**70%**	50%
	9	$x \Longrightarrow y$	90%	**50%**	**50%**
	10	$x \Longrightarrow \neg y$	100%	60%	60%
	11	$x \Longrightarrow y \wedge z$	100%	**50%**	**50%**
	12	$y \vee z \Longrightarrow x$	100%	**60%**	**60%**
	13	$y \wedge z \Longrightarrow x$	20%	**50%**	10%
	14	if $\geq M$ from $y \Longrightarrow y, z, \ldots$	50%	**60%**	**40%**
MILP	15	min min	0%	100%	0%
	16	disjun.	50%	**100%**	50%
	17	min. activ. level	100%	**100%**	100%
	18	indic. constr.	20%	**90%**	20%
	19	concave linear.	0%	**90%**	0%
	20	non-monoton. linear.	0%	50%	0%

AMPL reported the following issues:

```
...
test.mod, line 2 (offset 54):
 A is not defined
context: set CONTAINERS := A >>> B <<< C D;
No variables declared.
...
```

Fix all errors.

The M/C agent may respond by correcting the identified issues and regenerating the AMPL source code, which is then returned to the verifier. We allow only a single round of back-propagation. Although further iterations are technically possible, we observed that they do not improve success rates and may lead to repetitive error loops.

Results for the M/C+V architecture are presented in Table III. Changes compared to the M/C architecture (Table II) are shown in bold. We observe improvements in the coding success rate for nearly all artifacts where it was previously below 100%. The typical improvement is around 10 percentage points, but it reaches as much as 40 percentage points for artifact 19.

TABLE IV
RESULTS FOR M+C^R+V ARCHITECTURE

		Variant	Coder	Coder + RAG
LP	1	simple LP	100%	100%
	2	min max	100%	100%
	3	abs. val. in obj.	100%	100%
	4	abs. val. in constr.	80%	80%
	5	piece-wise fun.	60%	60%
MILP (logic)	6	at most N	60%	60%
	7	at least N	20%	**60%**
	8	exactly N	80%	80%
	9	$x \Longrightarrow y$	40%	**80%**
	10	$x \Longrightarrow \neg y$	40%	**60%**
	11	$x \Longrightarrow y \wedge z$	0%	**20%**
	12	$y \vee z \Longrightarrow x$	20%	**60%**
	13	$y \wedge z \Longrightarrow x$	0%	**40%**
	14	if $\geq M$ from $y \Longrightarrow y, z, \ldots$	20%	**60%**
MILP	15	min min	100%	100%
	16	disjun.	100%	100%
	17	min. activ. level	100%	100%
	18	indic. constr.	100%	100%
	19	concave linear.	80%	80%
	20	non-monoton. linear.	80%	**100%**

TABLE V
SHARE OF CODING ERRORS IN M/C AND M+C^R+V ARCHITECTURES

	CODER	CODER + RAG
set definition syntax	54.6%	38.1%
many objectives	4.6%	0.0%
two-side constr.	9.0%	2.4%
multiply constr. in one "s.t."	4.6%	4.8%
double definitions	4.6%	0.0%
syntax	22.7%	54.8%

TABLE VI
IMPROVEMENTS FOR M^R+C^R+V ARCHITECTURE VERSUS M/C ARCHITECTURE

		Variant	Model M/C	Model RAG
MILP LP	5	piece-wise fun.	10%	**100%**
MILP	13	$y \wedge z \Longrightarrow x$	20%	**40%**
Hard MILPs	15	min min	0%	**50%**
	18	indic. constr.	20%	**80%**
	19	concave linear.	0%	**70%**
	20	non-monoton. linear.	0%	**80%**

C. Separation of Modeler and Coder (M+C^R+V)

To further improve coding quality, we decomposed the monolithic M/C architecture into separate modeling and coding agents. This allowed us to augment the coder with Retrieval-Augmented Generation (RAG). Initially, we used chapters 1–9 from the original AMPL book [12], with each chapter provided as a separate PDF file. However, we observed no significant improvement.

Subsequently, we created a concise custom RAG resource focused on recurring coding issues. The RAG content is included in Appendix B. This short but targeted resource significantly improved outcomes—particularly for the second problem, which had previously suffered from poor coding quality. Table IV presents coding success rates for the second problem under the monolithic architecture (column "Coder") and the M+C^R+V architecture (column "Coder + RAG"). Changes are marked in bold.

As shown, coding was particularly weak in the second group of artifacts. The addition of a compact RAG significantly improved the coding success rate—in some cases, tripling it (e.g., artifacts 7 and 14).

Table V presents the distribution of error types reported by the AMPL environment. Several types of errors occur frequently. In the M/C architecture, more than half of the errors relate to syntax issues in set definitions. Nine percent concern the use of two-sided constraints, while similar proportions pertain to multiple constraints under a single 'subject to' statement, double definitions, and multiple objective functions. The remaining errors are categorized as general syntax issues.

When RAG is introduced, the distribution shifts. The frequency of specific known errors drops—sometimes to zero. As a result, the share of general syntax errors increases. However, the total number of errors decreases, indicating that the targeted RAG resource is effective. It successfully reduces both the frequency and proportion of the common, previously observed issues.

D. Enhancing Modeler agent with RAG: M^R+C^R+V

Encouraged by the positive impact of a relatively small RAG set for the Coder, we identified common modeling issues and developed a dedicated RAG for the Modeler agent as well. The resulting RAG is a two-page document that provides guidelines for addressing the most frequent challenges faced by the Modeler agent. Appendix C presents its contents, divided into three sections: piecewise linear function modeling, the minimum of minima problem, and indicator constraints.

Table VI shows the improvements achieved over the monolithic M/C model for a selection of the most challenging artifacts covered by the new RAG. The observed gains are significant, and the effectiveness of the targeted RAG should be considered high.

Table VII presents the complete results for the M^R+C^R+V architecture, which includes decomposition into specialized agents, each enhanced with its own RAG. Although the results are significantly improved compared to the original monolithic M/C architecture, some limitations remain. Notably, coding artifacts from the second group still show signs of weakness. Additionally, modeling performance remains suboptimal for artifacts 13–16. Among the simpler problems, artifact 4—featuring absolute values in constraints—underperforms relative to others in this group.

IV. THE CONCEPT OF C-DSS

Our experience with decomposing the architecture into specialized agents leads us to propose a general framework for Conversational Decision Support Systems (C-DSS), as illustrated in Figure 4. The agents marked in bold represent the components studied in this work: the Modeler, Coder,

TABLE VII
RESULTS FOR M^R+C^R+V ARCHITECTURE

		Variant	Model	Code	Success
LP	1	simple LP	100%	100%	100%
	2	min max	100%	100%	100%
	3	abs. val. in obj.	100%	100%	100%
	4	abs. val. in constr.	60%	90%	60%
	5	piece-wise fun.	100%	80%	80%
MILP (logic)	6	at most N	100%	60%	60%
	7	at least N	100%	50%	50%
	8	exactly N	80%	70%	50%
	9	$x \Longrightarrow y$	90%	50%	50%
	10	$x \Longrightarrow \neg y$	100%	60%	60%
	11	$x \Longrightarrow y \wedge z$	100%	50%	50%
	12	$y \vee z \Longrightarrow x$	100%	60%	60%
	13	$y \wedge z \Longrightarrow x$	40%	50%	40%
	14	if $\geq M$ from $y \Longrightarrow y, z, \ldots$	50%	60%	40%
MILP	15	min min	50%	100%	50%
	16	disjun.	50%	100%	50%
	17	min. activ. level	100%	100%	100%
	18	indic. constr.	80%	90%	80%
	19	concave linear.	70%	90%	60%
	20	non-monoton. linear.	80%	50%	80%

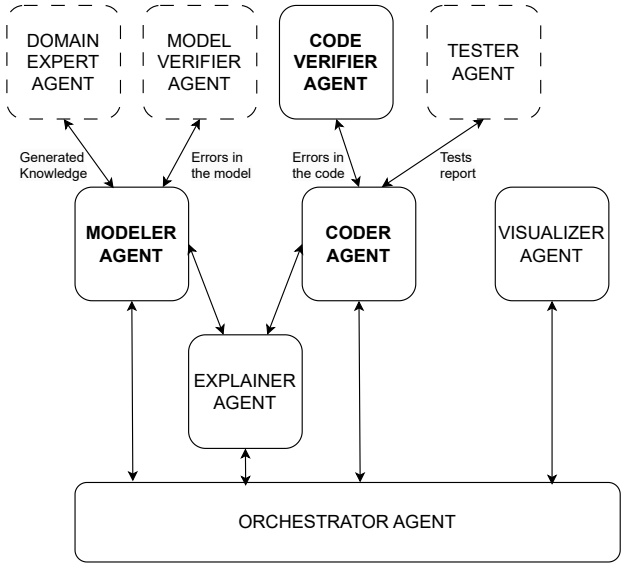


Fig. 4. The concept of Conversational Decision Support Systems

and Code Verifier. However, other elements of a classical Decision Support System (DSS) also involve communication and can be implemented within a language-agent-based architecture.

One example is the DSS user interface, whose role can be taken over by a *Visualizer agent*, dynamically responding to user requests for result visualization. Figure 5 shows a sample response from the Visualizer agent for base problem one, given the following prompt:

> *Visualize the results for the Management Board*

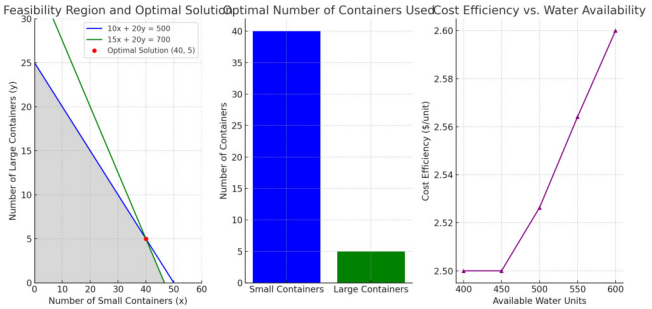


Fig. 5. Visualization of results proposed by the Visualizer agent (overlapping text left as it was generated by the agent)

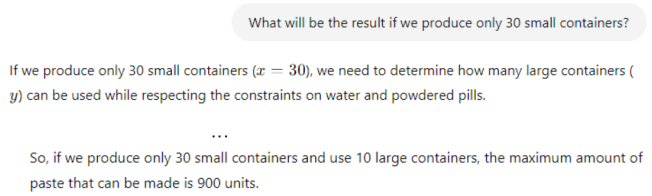


Fig. 6. Exemplary session 1 with Explainer agent

The agent proposed a visualization of the feasible solution set, which may not be suitable for a Management Board. It also presented the solution using a simple bar chart. Interestingly, the agent suggested performing an efficiency analysis with respect to water availability, thereby independently initiating a non-trivial sensitivity analysis.

Although the resulting output is not perfect and contains some visual flaws (e.g., overlapping text), it is important to note that the agent used was neither trained nor equipped with RAG, and the conversation was brief and based on a simple prompt. In this context, the results are promising and illustrate the potential for further development of such interfaces.

Another example is the *Explainer agent*, specialized in interpreting and explaining the obtained results to the user. In particular, it can conduct sensitivity or what-if analyses, and thus collaborates closely with the Modeler and Coder agents.

Figures 6 and 7 present two short sessions with the Explainer agent for base problem one. In the first session, the user seeks reassurance that the solution proposed by the C-DSS is valid and asks about the impact of reducing the number of small containers produced. The agent responds with a detailed explanation; Figure 6 shows only the opening sentence and final conclusion.

In the second session, the user wants to understand why increasing the production of small containers is not feasible. As before, the agent provides a longer explanation, and Figure 7 shows the first and final statements. In both cases, the Explainer agent offers a fairly detailed and accurate analysis of the user's proposed changes.

Assuming the Modeler is a general-purpose agent, it is reasonable to enable communication with a domain-specific agent that provides expert knowledge in a given field. While

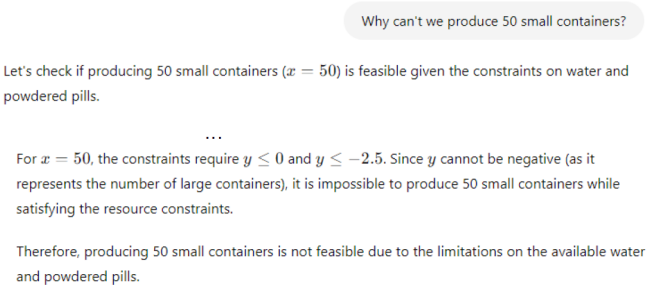


Fig. 7. Exemplary session 2 with Explainer agent

the implementation details remain open, introducing a Critic agent—responsible for verifying and validating the developed mathematical model (i.e., a Model Verifier)—appears well justified. This agent could replace or support the function of a human expert, which proved to be a bottleneck in our experiments.

In the area of code development, the current Code Verifier only checks for syntactic correctness by testing whether the code runs in a given environment. However, a dedicated *Tester agent* should also be envisioned—capable of validating the encoded model using predefined unit tests or scenario checks.

Finally, the entire agent ecosystem requires workflow management, a task that would fall to an *Orchestrator agent*. This agent would be responsible for dynamically arranging tasks, coordinating agent communication, and preventing loops or deadlocks in the system.

V. Summary

This paper introduces the concept of Conversational Decision Support Systems (C-DSS), which leverage Large Language Models (LLMs) to enhance Operations Research (OR) methodology, particularly during the modeling and coding stages. We propose a novel framework in which LLMs serve as proxies across various phases of decision support system development, with a focus on mathematical modeling and AMPL code generation.

By identifying 20 typical modeling artifacts for Linear Programming (LP) and Mixed-Integer Linear Programming (MILP), we evaluate LLM performance across four agent-based architectures: Monolith (M/C), M/C with Verifier (M/C+V), Modeler and Coder with Retrieval-Augmented Generation (M+C^R+V), and both Modeler and Coder enhanced with RAG (M^R+C^R+V). Using two base problems—one from the NL4Opt dataset and a custom-designed sewage discharging problem—we observe that simpler LP artifacts yield near-perfect success rates. In contrast, complex MILP artifacts involving logical constraints or non-convex functions remain significantly more challenging.

Our experiments reveal patterns in LLM behavior, suggesting that targeted solutions to specific artifacts offer great potential for improvement. We demonstrate that even a very short RAG for the Coder agent and a slightly longer, yet focused, RAG for the Modeler agent can significantly improve performance—particularly for piecewise functions, minimum-of-minima formulations, and indicator constraints.



Fig. 8. Potential role of LLM in OR methodology

We also envision a more comprehensive C-DSS architecture, incorporating agents such as the Modeler, Coder, Verifier, Visualizer, Explainer, and Orchestrator. This multi-agent design offers a unified approach to integrating LLMs into OR and opens new avenues for research in automating decision support systems. A Decision Support System is a product developed within the broader framework of the OR methodology, as referenced in Figure 1. As shown, the various steps in the methodology rely on different forms of language to facilitate communication among stakeholders—making it inherently well-suited for enhancement through LLMs. Therefore, we propose considering a revised OR methodology in which language models serve as core tools at each stage. Figure 8 illustrates this vision.

Despite rapid advances in large language models, there are still significant limitations. Our results show that for some quite typical cases, LLMs are not a reliable tool for modelling and coding. Since the real problems are typically more complex and may include more sophisticated aspects, such as complex nonlinearity, dynamics, or uncertainty, it must be admitted that the C-DSS concept is in its early stages of development. It requires a significant improvement for real applications. However, we believe that our proposed C-DSS concept not only advances the design of automated decision systems but also lays the foundation for further research into LLM-enhanced OR methodologies.

APPENDIX

A. Word problems used in the research

Base problem 1 (#64 from NL4Opt data set)

> There are two specialized containers, a small and large one, that are used to make a pharmaceutical paste. The small container requires 10 units of water and 15 units of the powdered pill to make 20 units of the paste. The large container requires 20 units of water and 20 units of the powdered pill to make 30 units of the paste. The pharmacy has 500 units of water and 700 units of powdered pills available. How many of each container should be used to maximize the amount of paste that can be made?

Base problem 2 (the sewage discharging problem)

> The two cities discharge sewage into two treatment plants with a capacity of 5,000 tons and 3,000 tons per day, respectively. The purified water is pumped further into the river. Each city can divide sewage in any proportion between both treatment plants. The daily operating cost of treatment plant 1 is USD 3.5/ton. The daily operating cost of treatment plant 2 is USD 6/ton. The operating cost of the clean water pumping station at treatment plant 1 is USD 2.5/ton, and at treatment plant 2 - USD 2.2/ton. Assuming that cities must discharge at least 4,000 and 3,500 tons of sewage per day, plan the system's operation to minimize the daily cost of its operation.

B. Content of Coder's RAG

> This GPT acts as an expert in linear programming and AMPL (A Mathematical Programming Language). It assists users in creating AMPL files for given linear programming problems by understanding their requirements and translating them into correct and optimized AMPL code. It ensures the AMPL files are syntactically accurate and efficient, adhering to best practices in linear programming and AMPL usage. It will guide users through the process of defining sets, parameters, variables, objective functions, and constraints in their linear programming models.
> Whenever you are asked to create an AMPL file, follow these rules:
> 1) Start with the declarations of sets followed by declarations of parameters and variables.
> 2) When declaring a set of parameters, do not assign any value. Add the section "data;" at the end of the AMPL file and put all assignments there.
> 3) Do not use sharp inequalities. Always apply "lower than or equal" or "greater than or equal".
> 4) Do not use double inequality constraints. Break the double inequalities into two separate constraints.
> 5) There can be only one objective in the AMPL file.
> 6) 'var' definitions should not include assignment from other variables directly.
> 7) Under the keyword "subject to", only one constraint can be defined. If more constraints are needed, each constraint definition should start with "subject to".

C. Content of Modeler's RAG

> **Piecewise functions**
> Remember that when dealing with easy piecewise function cases, there is NO NEED to use auxiliary binary variables. It is enough to add constraints for each segment of the piecewise function. For instance, if the problem is minimizing the convex piecewise function, we can add auxiliary continuous variable x and a constraint $x \geq f_i(x)$, where $f_i(x)$ is an i-th segment of the piecewise function. However, if you add binary variables indicating the segment of the function, these variables must be constrained by the original variable. If the i-th linear segment is denoted by $f_i(x)$ and it is valid in the range from x_{i-1} to x_i, then we need to set auxiliary variable v_i to 1 whenever x is in the range from x_{i-1} to x_i. We can model that as follows:
>
> $$x \leq x_i v_i + M(1 - v_i)$$
> $$x \geq x_{i-1} v_i - M(1 - v_i)$$
>
> For instance, let's consider the function consisting of two piecewise segments, $f_1(x)$ and $f_2(x)$. For x from 5 to 10, there is function $f_1(x)$, and for x in a range from 10 to 18, there is function $f_2(x)$. We introduce auxiliary variables v_1 and v_2 such that
>
> $$v_1 + v_2 = 1$$
>
> To constraint variable x by variables v_1 and v_2, we define inequalities as follows
>
> $$x \leq 10v_1 + M(1 - v_1)$$
> $$x \geq 5v_1 \breve{} M(1 - v_1)$$
> $$x \leq 18v_2 + M(1 - v_2)$$
> $$x \geq 10v_2 \breve{} M(1 - v_2)$$
>
> To calculate the value of the piecewise function, we introduce the auxiliary continuous variable z, and we define the following constraints:
>
> $$z \geq f_1(x) \breve{} M(1 - v1)$$
> $$z \leq f_1(x) + M(1 - v1)$$
> $$z \geq f_2(x) \breve{} M(1 - v2)$$
> $$z \leq f_2(x) + M(1 - v2)$$
>
> In the final version of the model, $f_1(x)$ and $f_2(x)$ must be substituted with linear formulas.

Min of min problem (minimization of the minimal value)

If the problem is to minimize the minimum of some values, we need to add auxiliary binary variables that indicate which value is the lowest. Let us consider variables x and y and the problem of minimizing the minimal value out of variables x and y. Then, we must introduce a binary variable v that is 0 if x is lower or equal to y, and v is 1 if y is lower than or equal to x. In the case of $x = y$, the value of v can be any. To model that, we need to add the following constraints:

$$x \leq y + Mv$$

$$y \leq x + M(1 - v)$$

Having variable v that indicates whether x or y is the minimal value we must introduce auxiliary continuous variable z that will be equal x if v is 0 or to y if v is 1. It is an indicator constraint that can be modeled as follows:

$$z <= x + Mv$$

$$z >= x^{\smile} Mv$$

$$z <= y + M(1 - v)$$

$$z >= y - M(1 - v)$$

Remember that z must be constrained from above and below. Since z is the minimal value of x and y, it can be directly minimized in the objective:

$$\min z$$

WARNING! Remember, YOU CAN NOT SIMPLY constraint z in this way:

$$z \leq x$$

$$z \leq y$$

and minimize z. This is MISTAKE! Variable z will always be 0. You must use an auxiliary binary variable. Watch out for this case!

Indicator constraints

An indicator constraint is a constraint that is controlled by a binary variable. If the constraint is "lower than", and the auxiliary variable that controls this constraint is z, then we can add Mz to the right-hand side of the constraint, where M is a big number. If the original constraint is as follows:

$$\sum_i a_i x_i \leq c$$

then it can be controlled with a binary variable y in the following way:

$$\sum_i a_i x_i \leq c + Mz$$

Indicator constraint can be used to model conditional constraint in which one variable depends on the value of another variable. For instance, if x must be greater or equal to A when y is greater or equal to B, we can add the indicator constraint for variable y as follows

$$y \geq B - M(1 - z)$$

where z is an auxiliary binary variable. So if z gets 1 then the constraint becomes active. Then z must be set to 1 if x is greater or equal A, which can be modeled as follows:

$$z \geq (x - A)/M$$

Remember that in the general case of conditional constraint, we need to use the concept of the indicator constraint, so an auxiliary variable is needed.

REFERENCES

[1] F. S. Hillier, *Introduction to operations research*, 9th ed. Boston [etc.]: McGraw-Hill, 2010. ISBN 9780071267670

[2] R. Ramamonjison, H. Li, T. Yu, S. He, V. Rengan, A. Banitalebi-dehkordi, Z. Zhou, and Y. Zhang, "Augmenting operations research with auto-formulation of optimization models from problem descriptions," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Y. Li and A. Lazaridou, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Dec. 2022. doi: 10.18653/v1/2022.emnlp-industry.4 pp. 29–62.

[3] R. Ramamonjison, T. Yu, R. Li, H. Li, G. Carenini, B. Ghaddar, S. He, M. Mostajabdaveh, A. Banitalebi-Dehkordi, Z. Zhou, and Y. Zhang, "Nl4opt competition: Formulating optimization problems based on their natural language descriptions," in *Proceedings of the NeurIPS 2022 Competitions Track*, ser. Proceedings of Machine Learning Research, M. Ciccone, G. Stolovitzky, and J. Albrecht, Eds., vol. 220. PMLR, 28 Nov–09 Dec 2022. doi: 10.48550/arXiv.2303.08233 pp. 189–203. [Online]. Available: https://proceedings.mlr.press/v220/ramamonjison23a.html

[4] H. Chen, G. E. Constante-Flores, and C. Li, "Diagnosing infeasible optimization problems using large language models," 2023.

[5] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large language models as optimizers," 2024.

[6] Q. Li, L. Zhang, and V. Mak-Hau, "Synthesizing mixed-integer linear programming models from natural language descriptions," 2023.

[7] Z. Xiao, D. Zhang, Y. Wu, L. Xu, Y. J. Wang, X. Han, X. Fu, T. Zhong, J. Zeng, M. Song, and G. Chen, "Chain-of-experts: When LLMs meet complex operations research problems," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=HobyL1B9CZ

[8] S. Wasserkrug, L. Boussioux, D. den Hertog, F. Mirzazadeh, I. Birbil, J. Kurtz, and D. Maragno, "From large language models and optimization to decision optimization copilot: A research manifesto," 2024.

[9] A. AhmadiTeshnizi, W. Gao, and M. Udell, "Optimus: Optimization modeling using mip solvers and large language models," 2023.

[10] ——, "Optimus: Scalable optimization modeling with (mi)lp solvers and large language models," 2024.

[11] T. Ahmed and S. Choudhury, "Lm4opt: Unveiling the potential of large language models in formulating mathematical optimization problems," *INFOR: Information Systems and Operational Research*, vol. 62, no. 4, pp. 559–572, 2024. doi: 10.1080/03155986.2024.2388452. [Online]. Available: https://doi.org/10.1080/03155986.2024.2388452

[12] R. Fourer, D. Gay, and B. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, ser. Scientific Press series. Thomson/Brooks/Cole, 2003. ISBN 9780534388096. [Online]. Available: https://books.google.com.hk/books?id=Ij8ZAQAAIAAJ

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 175–183
DOI: 10.15439/2025F0376
ISSN 2300-5963 ACSIS, Vol. 43

Slurm plugin for HPC operation with time-dependent cluster-wide power capping

Alexander Kammeyer*†, Florian Burger*, Daniel Lübbert* and Katinka Wolter†
0000-0002-7858-0354, 0000-0003-4745-5515, 0000-0003-3852-5665, 0000-0002-8630-0869
*Physikalisch-Technische Bundesanstalt, Abbestraße 2-12, 10587 Berlin, Germany
Email: {alexander.kammeyer, florian.burger, daniel.luebbert}@ptb.de
†Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany
Email: {a.kammeyer,katinka.wolter}@fu-berlin.de

***Abstract*—HPC systems are shared between many users. Managing their resources and scheduling compute jobs is a central task of these clusters. Scheduling also allows to control the workload and energy consumption of an HPC system. A Digital Twin of an HPC cluster can aid in the scheduling process by providing energy measurements about the system and predict scheduling decisions with a simulation. For a real-world use case, an integration of the Digital Twin with the scheduler is necessary. A possible use case are energy limitations as part of a demand response process between the HPC operator and energy supplier.**

Therefore, this paper introduces a plugin for Slurm, an open-source scheduler, that implements a scheduling algorithm for time-dependent cluster-wide power capping. It uses a node energy model to predict the energy consumption of jobs and can start jobs at different frequencies to stay below the configured power limit. The plugin interfaces with the Digital Twin that provides energy measurements for the compute nodes to track the system power consumption in real time and update the power limitations if necessary.

The plugin is tested on a cluster and compared against a scheduling simulation of the algorithm. The analysis compares the power profile of the simulation and the real system and the allocation of the jobs over time. Differences in the execution and the power trace are analysed and discussed.

I. Introduction

THE High-Performance Computing (HPC) community has focussed on achieving maximum performance for a long time. This doctrine has recently come under pressure by rising energy prices and a transformation of the energy market towards renewable energies while HPC systems continue to grow in size and energy demand. A trend that is further accelerated by the recent shift towards artificial intelligence.

An HPC system under full load has a relatively constant energy requirement. However, a shift towards renewable energy production means greater volatility in the energy production. HPC systems need a mechanism to respond to these fluctuations. Solely relying on fossil energy source is also not an option, as regulatory frameworks, such as the Blue Angel for data centres [1], define upper bounds for carbon-dioxide emissions.

Dynamically adapting the energy consumption of an HPC system has further use cases. It can help stabilize the energy grid by reducing the load in the grid or by increasing the load if enough energy is available. Additionally, such a mechanism allows the continued automatic operation of the cluster in case of energy shortages.

Reacting to limited energy availability is sometimes referred to as demand response. While it is common practice for large energy consumers, HPC centres have not yet widely adopted such mechanisms [2]. However, first mechanism have been proposed. One possible demand response mechanism was presented in [3]. A Digital Twin is used to collect data about the system energy consumption and together with information about energy availability, the scheduler can adapt the overall energy consumption to stay below a defined threshold. The algorithm in the original paper was validated with a simulation. This paper implements the algorithm as a plugin for the popular cluster scheduler Slurm [4].

This paper makes the following contributions:

- The algorithm uses a node energy model to predict the energy consumption of the jobs. This paper uses a new test platform for the evaluation of the plugin. Thus, an updated node energy model is necessary that has been created for the new test system.
- The algorithm from the original facilitates time-dependent cluster-wide power capping. This paper implements this algorithm as a plugin for the open-source scheduler Slurm.
- The algorithm was previously only evaluated in a simulation of the Digital Twin. This paper compares the simulation against the performance on a real cluster. The evaluation includes power traces of the real system and the node allocation to the jobs.

The remainder of this paper is structured as follows: Section II introduces Slurm in detail and presents relevant literature, especially regarding cluster operation under a power limitation. Preparatory work such as the updated node energy model and a recap of the scheduling algorithm is presented in Section III. In Section IV, the implementation of the Slurm plugin is discussed. The plugin is then evaluated in the real world and compared to simulation results in Section V.

II. Related Work

An HPC system consists of many individual components and is typically shared among users for running their compute jobs. These systems thus need a resource manager to avoid

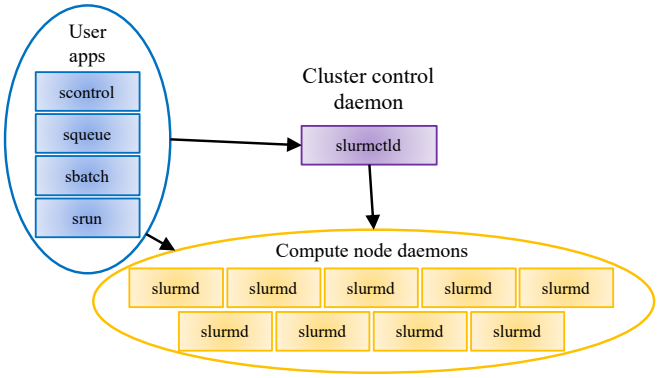


Figure 1: Overview of the Slurm components. The `slurmctld` is the central controller of the cluster while `slurmd` handles each compute node. Client components such as `scontrol`, `squeue`, `sbatch` and `srun` allow the user to interact with the cluster and submit and manage compute jobs.

resource conflicts between users as well as a strategy to assign the available resources to the requests of the users, referred to as scheduling strategy. Different resource managers and schedulers exist such as e.g. Moab, PBS, Grid Engine and Slurm [4].

Being open-source software written in the C programming language, Slurm has been widely adopted by the research community. Slurm works both as a resource manager and as scheduler and also starts and monitors the compute jobs on the cluster. Additionally, it is able to handle accounting of the resource usage and supports resource reservation in advance. The software consists of multiple components: the Slurm control daemon `slurmctld`, the Slurm daemon on the compute nodes `slurmd` and command-line utilities for the users to manage Slurm and submit (`sbatch`) and control (`scontrol`) their jobs. Slurm also handles the MPI configuration for the jobs and offers the launch program `srun`. These components are shown in Figure 1. Slurm has further components that are responsible for accounting and billing but they are not relevant in the context of this paper.

HPC systems are composed of individual compute nodes which might feature different hardware, e.g. CPU cores, memory and additional components such as storage, GPUs or TPUs. The nodes can be grouped into partitions for bundling nodes with identical hardware or for aiding the scheduling process, e.g. when certain nodes are reserved for short or long running jobs or similar. When users want to run a task on the HPC system, they submit a job. The job defines what the user wants to compute, the required resources to complete the computation and for how long the job will need those resources. Slurm can further divide jobs into job steps, that define tasks that may be executed in parallel or contain pre- and post-processing and do require different resources.

The specific scheduling algorithm and resource allocation can contribute to the energy efficiency and stability of the system and thus has been subject to many research studies. The surveys by Czarnul [5] and Kocot [6] provide an overview of tools for energy and power management and power-aware scheduling.

Simulating the scheduling process is a common approach in HPC research. One example is CQSim [7]. Based on trace data, the authors propose a scheduling approach for demand response, and test it in their simulation [8]. To retrieve such traces, the system needs to be actively monitoring the energy consumption of each job. Not every cluster is designed to do that because the hardware does not provide the necessary measurements, or because the scheduling environment does not track the data, or both. Another important aspect is taking the specifics of the different applications into account, when imposing a power cap on the cluster [9]. The scheduler may automatically change the frequencies and thus runtime of the jobs to dynamically distribute the power. Other demand response methods rely on power traces to create a node model and limit power consumption of the HPC clusters [10]. None of these approaches have been tested in the real world.

One approach in system management and system modelling has attracted increased interest over the last years: Digital Twins [11]. A Digital Twin is a virtual representation of a real-world object, in this context an HPC cluster. A real-world object is represented as a Digital Twin in the virtual world. Data about the object is collected and integrated into the Digital Twin. With their characteristic bi-directional data linkage, changes in the real world are reflected in the digital domain and vice versa. The Digital Twin models the behaviour in the real world and can test new parameters before they are applied to the real system. Simulations are often cheaper and faster and a negative impact on the real system can be avoided. In the HPC context, these simulations are generally scheduling simulations as they describe changes in the system state well.

Digital Twins are under development for HPC systems ranging from small clusters to one of the world's largest systems, Frontier [12]. As part of their smart campus initiative [13], the Physikalisch-Technische Bundesanstalt (PTB) is developing a Digital Twin for their HPC system [14]. It has been demonstrated that it can be used to reduce system emissions. One major use case of the Digital Twin is the operation of HPC systems under a time-dependent cluster-wide power cap. The power can be capped by the Digital Twin, depending on the currently available energy [3]. Precise energy measurements also allow the Digital Twin to track the data centre Power Usage Effectiveness (PUE) [15].

Similar to the node energy model in Section III-B, application power profiles are a common technique used in application modelling and have been done for a wide variety of applications, including WZ factorisation [16] and matrix factorisation [17]. An alternative to Dynamic Voltage and Frequency Scaling (DVFS) is the implementation of a power cap as a hardware driver [18].

III. Preliminary Work

This section introduces the test system used in this paper, and shows how Slurm has been configured for this specific system. Since the test system differs from the system in [3], a new node energy model has been created. Furthermore, a cap on the implemented scheduling algorithm is given.

A. Test System

For the practical test of the Slurm plugin, this paper relies on a cluster consisting of Raspberry Pis. While this setup is not typical for a commercial HPC system, it has a similar structure and allows the development of the plugin without interrupting the production HPC system of PTB. The setup is based upon a design found proposed in [19].

The system is equipped with a total of 10 Raspberry Pi 4B with 8 GB RAM each. They are connected via their $1\,\mathrm{Gbit/s}$ Ethernet ports. Each Pi also has the Power over Ethernet (PoE) HAT (Hardware Attached on Top) installed that allows the nodes to be supplied with power from the Ethernet switch. All run the Raspberry Pi OS Bookworm.

The switch, a Netgear GS316EPP, supplies power via Power over Ethernet (PoE) to the nodes and offers a gigabit interconnect. It has a rudimentary management interface that provides energy information for each of the ports and allows management and power cycling of individual ports. These features thus allow us to track the per-node energy usage and to turn nodes off and on when needed as the Raspberry Pi does not support Wake-On-LAN. The energy data is queried by a collector agent of the Digital Twin and provided via the Digital Twin database.

One node is designated as the head node while the remaining 9 nodes act as compute nodes. The head node provides a global file system via NFS to all compute nodes. For this purpose, a single SDD is connected via an SATA-to-USB adapter and provides the necessary storage space. It also contains scratch space that is shared between the nodes. The compute nodes themselves are disk-less and get their image via TFTP from the head node when they boot. An additional Ethernet-to-USB adapter establishes a connection to the outside world while a DHCP server provides the internal cluster network with addresses for the compute nodes. The head node uses NAT to forward the internet connection to the compute nodes. This is necessary because the Raspberry Pi does not have a real-time clock and the clocks of the compute nodes would be out of sync after a reboot.

As the scheduler and resource manager we have opted to use Slurm [4]. Being open source software, the Digital Twin can be integrated with the scheduler. Slurm is configured with a single partition spanning all 9 compute nodes. Since the cluster is not intended for production use with many users, only basic logging is configured.

The head node is running the `slurmctld` control daemon that manages the scheduling logic and interfaces with the Digital Twin. The compute nodes run the `slurmd` daemons that connect the nodes to the control daemon. The user connects to the head node and can use the Slurm utilities, e.g. `sbatch`, to control the cluster and submit batch jobs.

Slurm provides some power saving functionalities [20], most notably node suspension and resumption. The scheduler monitors the nodes. If a node has been idle for a certain amount of time, it is shut down. When the node is needed again, Slurm issues a restart. This is implemented by power-cycling the Ethernet port of the switch, resulting in the Raspberry Pi booting again. The Slurm daemon on the node then contacts the Slurm controller, signalling that the node is ready to receive the next job. For this experiment, Slurm has been configured to turn of a compute node after 10 minutes of inactivity.

B. Node Energy Model

The node energy model is an integral part of the scheduling algorithm. It allows the algorithm to estimate the energy consumption of a job. It is the part that makes the algorithm hardware specific. The remaining logic works independently from any underlying hardware characteristics. However, it needs to be updated if the algorithm should run on a different cluster with new workloads. Hence, this section presents the node energy model for the test system used in this paper. For this experiment, a total of 4 different benchmarks and parallel applications have been selected, namely the HPL and HPCG benchmarks as well as OpenFOAM and Geant4.

The High-Performance Linpack (HPL) [21] is used to create the TOP500 ranking of supercomputers. The benchmark is a linear equation solver and uses highly optimized vector instructions, thus serving as a kind of upper bound for the model. The second benchmark, the High Performance Conjugate Gradients (HPCG) [22], complements the HPL benchmark by adding more memory intensive operations.

The two benchmarks are supplemented by two numeric applications. The first, Open Field Operation And Manipulation (OpenFOAM) [23], is a computational fluid dynamics package and the second, Geometry and Tracking (Geant4) [24], is a Monte Carlo simulation toolkit.

To create the node energy model, each of the 4 jobs was run on the nodes on all supported frequencies. The switch provides an energy measurement that allows to monitor the jobs energy consumption. From these measurements, the Time-to-Solution (TtS) and Energy-to-Solution (EtS) can be derived. Figure 3 shows the results for all supported frequencies from $0.6\,\mathrm{GHz}$ to $1.8\,\mathrm{GHz}$. Pareto-optimal points [25], [26] are coloured in orange. For the model shown in Table I, three points have been selected: the highest frequency of $1.8\,\mathrm{GHz}$ and if available two pareto-optimal points. The scaling factor is used to adjust the job runtime for the increased runtime required if the job runs at a lower frequency setting.

This model shows a downside of the Raspberry Pis as a test platform. They have a relatively high offline energy consumption, especially compared to their idle consumption and the consumption under load [27]. This amounts to $22.1\,\mathrm{W}$ offline consumption for the compute nodes of the cluster. While $2.70\,\mathrm{W}$ is not much compared to other platforms, it is

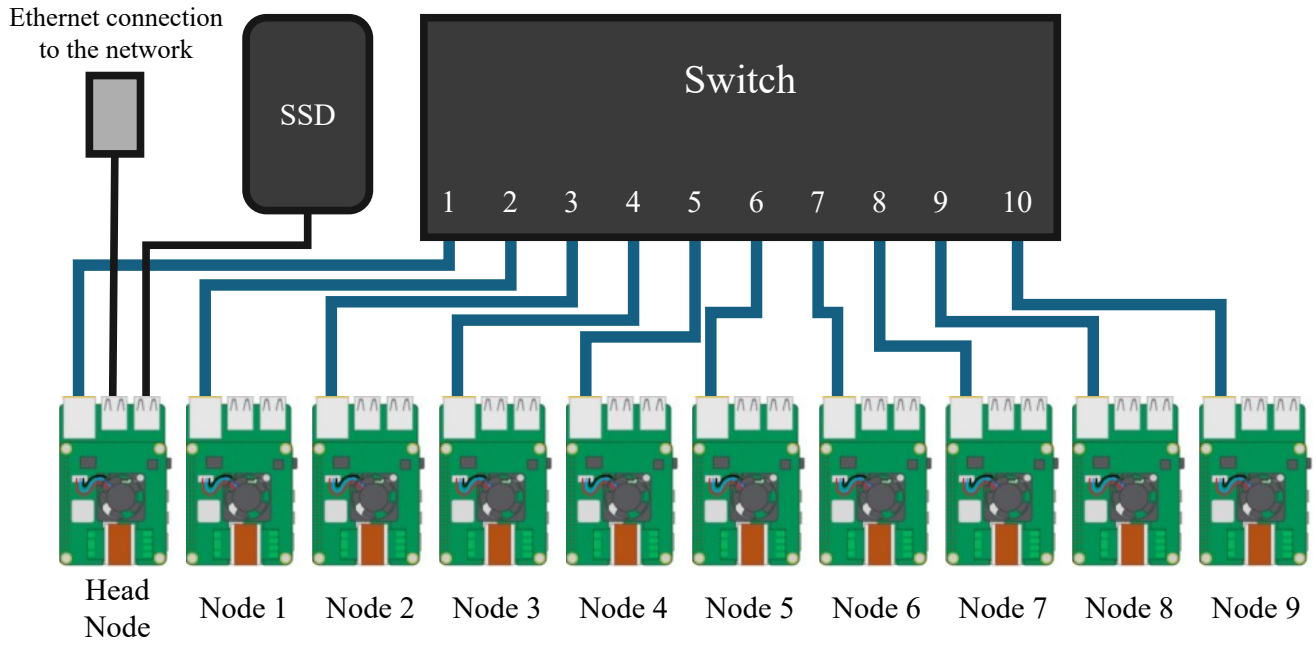


Figure 2: Schematic overview of the test cluster. It consists of one head node and a total of 9 compute nodes. They are connected via Ethernet and receive power via PoE. The head node has an additional Ethernet port to connect the cluster to a network. The SSD is also connected to the head node and provides storage for the compute nodes.

Table I: The node energy model

Job type	Frequency	Power	Scaling factor
HPL	1.8 GHz	6.20 W	1.000
	1.6 GHz	5.74 W	1.049
	1.5 GHz	5.47 W	1.055
HPCG	1.8 GHz	5.05 W	1.000
	1.3 GHz	4.25 W	0.911
	1.2 GHz	4.13 W	0.933
OpenFOAM	1.8 GHz	5.95 W	1.000
	1.4 GHz	5.17 W	1.015
	1.1 GHz	4.89 W	1.064
Geant4	1.8 GHz	6.62 W	1.000
	1.5 GHz	5.55 W	1.208
	1.1 GHz	4.92 W	1.625
Idle	-	3.12 W	-
Offline	-	2.70 W	-

almost half as high as the busy consumption. This behaviour can be changed by forcibly disabling PoE on the switch instead of only issuing a software shutdown.

C. Algorithm Recap

The goal is to run the HPC cluster under a variable power limitation while maximizing the system usage. The algorithm's pseudocode is shown in Algorithm 1. This section briefly describes the idea behind the algorithm while an in-depth description was given in [3]. The algorithm uses two main concepts to improve the throughput: DVFS, and turning off nodes. The node energy model provides the optimal frequency settings, while turning off unused compute nodes further decreases the system power consumption.

The scheduling algorithm receives a list of eligible jobs, and sorts them by time of arrival. The first "if" statement checks whether there are enough online and offline nodes available to start the job, and starts offline nodes if necessary.

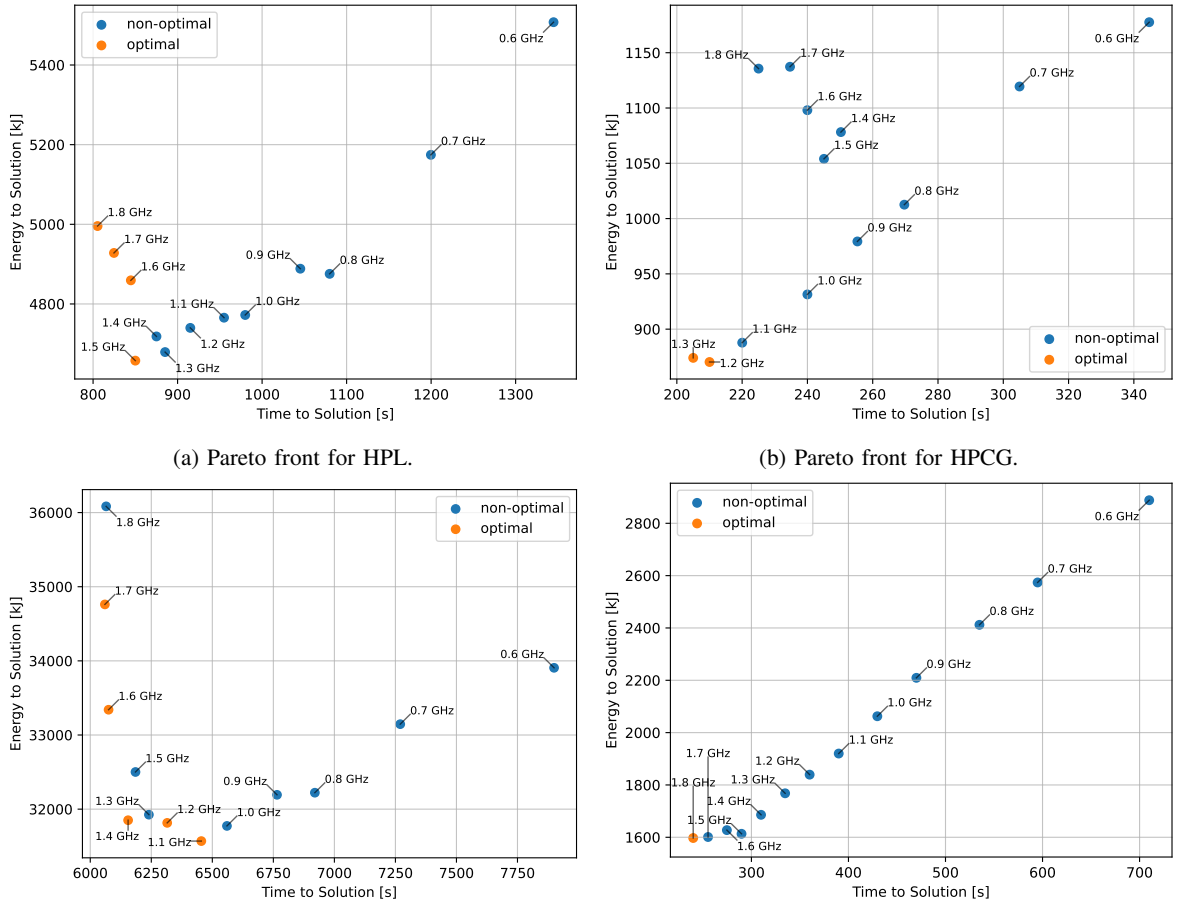
The second "if" statement checks whether enough online nodes are available to start the job. If that is the case, it uses the node energy model to estimate the job's power consumption. If starting the job would exceed the power limit, the job is deferred; otherwise, the job is started.

When a job is deferred, the routine moves on to the next job in the queue. This allows the algorithm to backfill jobs that might fit within the available resources.

IV. Slurm Plugin

The Slurm control daemon `slurmctld` is responsible for job scheduling. Two scheduling algorithms are included in the package as plugins. The first plugin is called `builtin` and implements a simple FCFS algorithm. The second plugin is called `backfill` and, as the name implies, implements FCFS with backfilling. This paper presents an additional, new scheduling plugin which implements the algorithm from Section III-C that can be used instead of the integrated plugins.

By default, Slurm schedules jobs automatically without consulting the scheduling plugin, as long as no resource conflicts arise. E.g. a job submitted to an empty cluster that fits in terms of compute nodes gets scheduled without the scheduling plugin. This circumvents the energy limit check of the scheduling plugin, and in our case could lead to the power consumption exceeding defined limits. To avoid such a



(a) Pareto front for HPL.

(b) Pareto front for HPCG.

(c) Pareto front for OpenFOAM.

(d) Pareto front for Geant4.

Figure 3: Pareto fronts for the four different applications in the node energy model. Pareto-optimal frequencies in terms of EtS and TtS are coloured orange. The remaining frequencies are coloured blue.

situation, all jobs must be submitted in the `hold` state with a priority of zero. This prevents Slurm from automatically scheduling jobs [28].

The Slurm control daemon periodically consults the scheduling plugin. The plugin first consults the database and updates the power limits and queries the current power consumption of the system from the database. It contains a set of functions that allow it to connect to the InfluxDB of the Digital Twin. InfluxDB does not offer a native C client implementation. However, since Slurm includes the cURL library, the plugin uses cURL to connect to the REST API of the database in order to query both the power consumption of the nodes and the power limitations. This provides a level of abstraction of the database calls from the rest of the code. Furthermore, helper functions convert the reply from csv to C-structs for further use.

During this periodic invocation, the plugin uses the core scheduling routine `_backfill_power()` which is shown in Algorithm 2. It does not use the usual scheduling queue but the entire job list. First, it checks if any jobs are pending but have not started yet in the first while loop (lines 13-21). Those are jobs that have been scheduled by the plugin but have not yet been started by Slurm. This avoids releasing to many jobs that later conflict with one another. If a job is waiting to start, the function is exited and waits for the periodic scheduling before making the next attempt.

The second while loop (starting in line 25) handles the actual scheduling. The plugin first consults `job_test_resv()` (lines 30.32), a Slurm function that tests whether enough nodes are available. Afterwards, `select_g_job_test()` (lines 31-49) tries to select the optimal nodes and tests whether the job can start immediately. Last but not least, `check_power_dvfs()` (lines 57-58) uses the node energy model to estimate the energy requirement of the job and whether the current power limitations allow the job to start. It is also responsible to set the frequency and adjust the runtime of the job according to the numbers in Table I. If enough energy is available, the priority of the job is increased which releases the job from the `hold` state. The `slurmctld` recognizes that and starts the job. It is also responsible to

Algorithm 1 Pseudocode of the core scheduling routine

```
backfill_power(eligible_jobs) {
  for (Job j : eligible_jobs) {

    // test if nodes need to be booted
    // check if enough nodes are offline
    if (j.nodes > online_nodes(j) &&
        j.nodes <= online_and_offline_nodes(j)) {
      // how many nodes need to be booted?
      toboot = j.nodes - online_nodes(j);

      // check if the nodes would
      // exceed the power limit
      if(!check_power(j, get_offline(j, toboot))){
        continue;
      }
      boot_nodes(j, toboot);
    }

    // test if enough nodes are online
    // and available
    if (j.nodes <= online_nodes(j)) {

      // check if the job and nodes would
      // exceed the power limit
      if (!check_power_dvfs(j)) {
        continue;
      }

      assign_nodes(j);
      j.wait_time = tick - j.submit_time;
      running.add(j);
      eligible_jobs.remove(j);

      // trigger an event in the simulation
      // on job completion
      e = new JobEvent(tick + j.run_time_scaled,
                j, JobState.COMPLETED);
      eventQueue.add(e);
    }
  }
}
```

select the nodes and start them if necessary. Slurm also sets the configured frequency.

So far, the implementation has been tested on a homogeneous cluster with a single partition. It does not yet consider multi-partition scheduling or other features such as multi-step jobs. Another aspect the algorithm cannot control yet is energy usage during the boot process of the nodes, which can lead to power spikes. One possible way to control this is by introducing a delay in the boot script that Slurm uses to power-cycle the ports on the switch. An early loaded power cap driver, as proposed in [18] might also help.

V. Evaluation

This paper presents a Slurm plugin based on a scheduling algorithm, evaluated previously only through simulation. To compare the simulated results with the real-world, both the simulation and the plugin are configured with the node energy model from Section III-B and an identical number of compute nodes.

The workload for the evaluation is generated with the improved Feitelson job model [29], [30]. The model generates a job trace in the Standard Workload Format (SWF) [31]. While the format encodes which application was run, the

Algorithm 2 Actual code: Core scheduling routine of the Slurm plugin

```
static void _backfill_power(void) {
  int j, rc = SLURM_SUCCESS, job_cnt = 0;
  job_record_t* job_ptr;
  bitstr_t *alloc_bitmap = NULL;
  bitstr_t *avail_bitmap = NULL;
  bool resv_overlap = false;
  resv_exc_t resv_exc = { 0 };
  time_t now = time(NULL);
  list_itr_t *iter = NULL;

  // check if a job has been scheduled
  // and has not started yet, defer further
  // scheduling in this case
  iter = list_iterator_create(job_list);
  while ((job_ptr = list_next(iter))) {
    if(IS_JOB_PENDING(job_ptr) == true
      && job_ptr->priority > 0) {
      list_iterator_destroy(iter);
      return;
    }
  }
  list_iterator_destroy(iter);

  // go through the list of held jobs
  iter = list_iterator_create(job_list);
  while ((job_ptr = list_next(iter))) {
    if(IS_JOB_PENDING(job_ptr) == true
      && job_ptr->priority == 0) {
      // Determine which nodes a job can use
      // based upon reservations
      j = job_test_resv(job_ptr, &now, true,
        &avail_bitmap, &resv_exc,
        &resv_overlap, false);
      if (j != SLURM_SUCCESS) {
        FREE_NULL_BITMAP(avail_bitmap);
        reservation_delete_resv_exc_parts(
          &resv_exc);
        continue;
      }
      // Select the "best" nodes for given job
      // from those available
      rc = select_g_job_test(job_ptr,
        avail_bitmap,
        job_ptr->details->min_nodes,
        job_ptr->details->max_nodes,
        job_ptr->details->max_nodes,
        SELECT_MODE_RUN_NOW,
        NULL, NULL,
        &resv_exc,
        NULL);
      FREE_NULL_BITMAP(avail_bitmap);
      reservation_delete_resv_exc_parts(&resv_exc);
      if (rc != SLURM_SUCCESS) {
        continue;
      }
      // We are certain that a job can start
      // check the power limit and adjust DVFS
      if(check_power_dvfs(job_ptr,
          get_req_nodes(job_ptr)) < 0) {
        continue;
      } else {
        // release the job
        // slurmctld will start the job
        job_ptr->priority = 1;
        job_ptr->state_reason = WAIT_NO_REASON;
        break;
      }
    }
  }
  list_iterator_destroy(iter);
}
```

Feitelson model does not generate this information. Hence, an equal distribution is used to generate the workload information with equal probability for each of the 4 different applications. Slurm only supports runtime limits with minute resolution for the job while the Feitelson model produces job lengths with seconds resolution. All job runtimes are therefore rounded up to the next full minute.

As a baseline, the job trace was simulated by the scheduling simulation component integrated in the Digital Twin. The power trace is shown in Figure 4a and the allocation of the jobs to the nodes in Figure 5a. In comparison, the same job trace was run with identical power limitations on the cluster. A script processed the job trace file and submitted the jobs accordingly. The power trace is shown in Figure 4b and the node allocation in Figure 5b.

In the simulation, the algorithm uses the gaps between the power limit while staying below the set limit. Larger jobs did not fit below the limit and were delayed beyond the limitation. This effect can also be seen in the node allocation in Figure 5a. The results are comparable to the first iteration of tests with the other node energy model [3].

In the experiment on the test system, looking at the power trace, the cluster mostly stays below the limit. Sometimes the limit is slightly exceeded in the order of $\leq 3\%$ which can be attributed to booting nodes and job initialization. In contrast to the simulated system where energy needed to start the nodes is not accounted for, the real system constantly tracks consumed energy also while booting additional nodes. When the operating system and Slurm are fully loaded, controlling the energy use is rather simple. However, especially during the early boot phase, the system draws a lot of power while not offering a mechanism to control it.

Another difference is that larger jobs start earlier in the real system test. This increases the system utilization and reduces the time required to complete the job trace. This again hints towards slight differences in the energy calculation between the plugin and the simulation. While the simulation considers the idle energy usage of the nodes, the plugin considers the energy usage of all running jobs and assumes that unused nodes are shut down by Slurm. The plugin only checks whether the running jobs allow an additional job to start. Furthermore, the node energy model upon which in both cases scheduling decisions rely on, represents the in average power consumption over time of jobs - not their maximum values. Therefore it is clear, that the real system will exhibit these maxima in power consumption by slightly surpassing the power limit. This could of course be cured in future by a more refined and time-resolved node energy model.

The time the scheduling and start of the job actually takes makes up a third difference between the simulation and the real world. In the simulated run, jobs are tightly packed as the simulation starts jobs instantly. In contrast, the Slurm plugin first releases them from the hold state and subsequently waits for the job to actually start before making the next scheduling decision. This causes a slight delay, thus creating small gaps between the jobs visible in Figure 5b.

To summarize, this experiment has demonstrated that the power-capping algorithm is feasible on a real system via a Slurm plugin, with a few initial limitations. The next step in the development will be to align the real-world and the simulation more closely with one another. The Slurm logic is complex and hard to replicate in a pure simulation. The simulation also does not simulate everything. This includes accurate power models for the boot process, and time required to start and stop jobs on the nodes.

The test system only has a single node type and has one partition configured. The plugin remains to be tested in a heterogeneous system configuration: Slurm allows to configure different hardware in a single partition, which will make the check more complex. Our plugin does not yet consider shared node allocation, but this can be complemented in a next step, as it is supported by Slurm.

In the current implementation, the node energy model is hard-coded in the plugin, which reduces portability of the code. For a broader use, the parameters will be moved to a configuration file. The accuracy of the algorithm depends on the accuracy of the node energy model. So far, no changes in the power consumption of the jobs have been considered. Additionally, the model is static and does not adapt to new workloads. Another possible improvement for the plugin are malleable jobs, which can be extended in their runtime when they are switched to a lower frequency during their runtime.

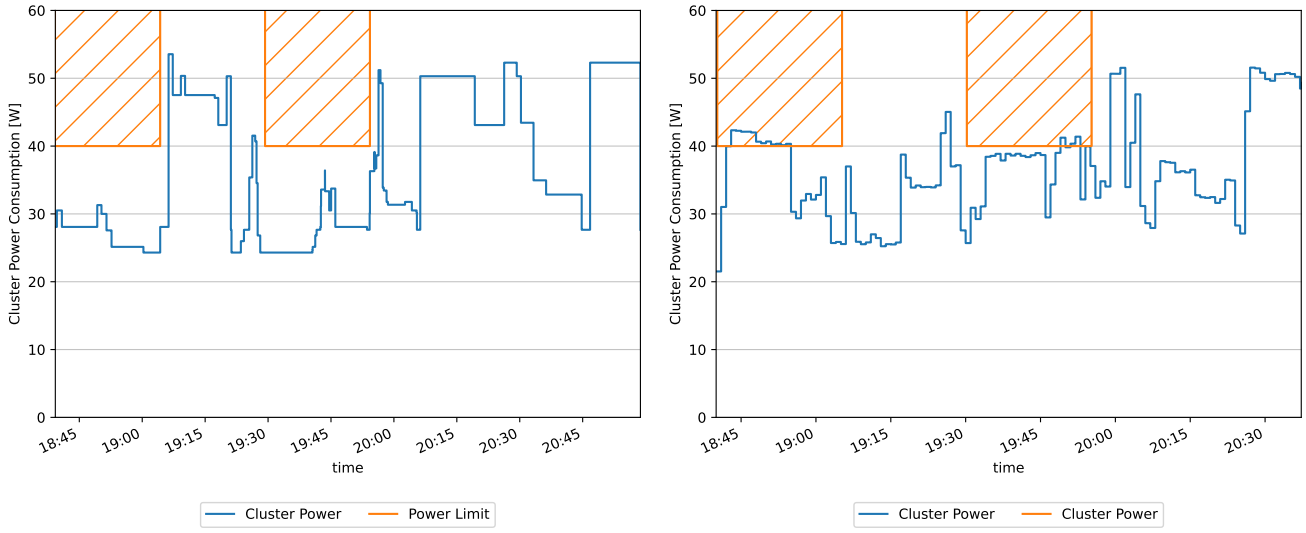
VI. Conclusion and Future Work

This paper presented a new scheduling plugin for the widely used scheduling and resource management software Slurm. The plugin allows an HPC cluster to operate under a system-wide power cap. The plugin works by holding jobs until enough power is available. Jobs can be configured by the plugin with a specific frequency according to the node energy model to reduce energy consumption. If the budget allows, the plugin releases the job from the hold state and sets the frequency. The Slurm infrastructure then handles the job start and node configuration.

The scheduling algorithm uses a node energy model to estimate the energy consumption of a compute job. For this paper, a new model has been created for the specific hardware of the test system. It has been demonstrated that this approach can be adapted to different systems, making the scheduling algorithm portable.

The plugin was demonstrated on the test system against a simulation with an identical job trace. Differences between the simulation and plugin have been discussed, especially job start times and energy calculation. While the plugin was able to stay below the power limit, booting compute nodes might briefly surpass the configured limit. Possible fixes have been discussed. The plugin was able to schedule jobs efficiently, and shows higher utilization than the simulation.

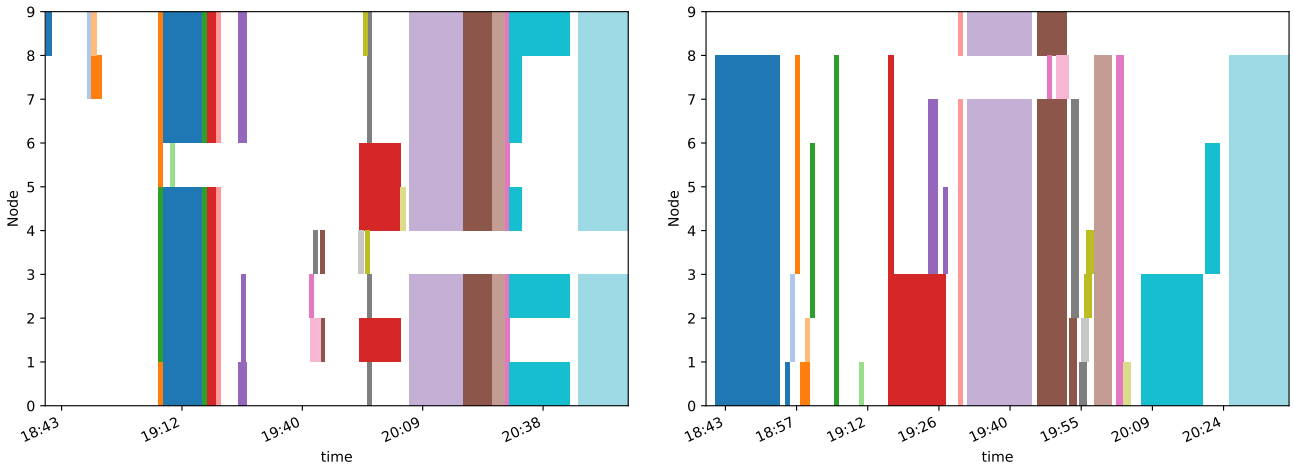
The Slurm plugin presented in this paper was tested with a homogeneous system configuration with identical nodes in a first step. Modern HPC systems often use heterogeneous system configurations with different node types and different



(a) Power trace of the simulation by the Digital Twin.

(b) Power trace of the test cluster.

Figure 4: Power traces of the experiment. A job trace was simulated and run on the test cluster to compare the performance of the algorithm in the simulation and the real world. The power trace is shown in blue and the limit is shown as the orange dashed surface.



(a) Node allocation of the simulation.

(b) Node allocation of the real system.

Figure 5: Node allocation of the experiments. The width of the rectangles corresponds to the runtime and the height shows which nodes where allocated to the job.

generations of CPUs. The test system could be extended with newer Raspberry Pi 5s or NVIDIA Jetson boards to create such a heterogeneous system.

Currently, the node energy model is hard-coded in the plugin. This would need to be shifted to a configuration file. The plugin has not been tested with heterogeneous or multi-partition cluster configurations. Since these are common nowadays, further development is necessary to support these configurations.

ACKNOWLEDGEMENT

We would like to thank our colleagues from department Q.13, first and foremost S. Thümmler, for their help with building the case for the Raspberry Pi cluster.

REFERENCES

[1] RAL UMWELT, *Rechenzentren DE-UZ 228*, 2nd ed., Fränkische Straße 7, 53229 Bonn, Feb. 2025. [Online]. Available: https://produktinfo.blauer-engel.de/uploads/criteriafile/de/DE-UZ-228-280225-de-Kriterien-V3.pdf

[2] T. Patki, N. Bates, G. Ghatikar, A. Clausen, S. Klingert, G. Abdulla, and M. Sheikhalishahi, "Supercomputing centers and electricity service providers: A geographically distributed perspective on demand management in europe and the united states," in *High Performance Computing*, J. M. Kunkel, P. Balaji, and J. Dongarra, Eds. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-41321-1_13. ISBN 978-3-319-41321-1 p. 243–260.

[3] A. Kammeyer, F. Burger, D. Lübbert, and K. Wolter, "HPC operation with time-dependent cluster-wide power capping," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39, 2024. doi: 10.15439/2024F1066 p. 385–393.

[4] M. A. Jette and T. Wickberg, "Architecture of the slurm workload manager," in *Job Scheduling Strategies for Parallel Processing*, D. Klusáček, J. Corbalán, and G. P. Rodrigo, Eds. Cham: Springer Nature Switzerland, 2023. ISBN 978-3-031-43943-8 p. 3–23.

[5] P. Czarnul, J. Proficz, and A. Krzywaniak, "Energy-aware high-performance computing: Survey of state-of-the-art tools, techniques, and environments," *Scientific Programming*, vol. 2019, p. 8348791, 2019. doi: 10.1155/2019/8348791. [Online]. Available: https://doi.org/10.1155/2019/8348791

[6] B. Kocot, P. Czarnul, and J. Proficz, "Energy-aware scheduling for high-performance computing systems: A survey," *Energies*, vol. 16, no. 2, 2023. doi: 10.3390/en16020890. [Online]. Available: https://www.mdpi.com/1996-1073/16/2/890

[7] X. Yang, Z. Zhou, S. Wallace, Z. Lan, W. Tang, S. Coghlan, and M. E. Papka, "Integrating dynamic pricing of electricity into energy aware scheduling for hpc systems," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. New York, NY, USA: Association for Computing Machinery, 2013. doi: 10.1145/2503210.2503264. ISBN 9781450323789. [Online]. Available: https://doi.org/10.1145/2503210.2503264

[8] S. Wallace, X. Yang, V. Vishwanath, W. E. Allcock, S. Coghlan, M. E. Papka, and Z. Lan, "A data driven scheduling approach for power management on hpc systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '16. IEEE Press, 2016. doi: 10.5555/3014904.3014979. ISBN 9781467388153

[9] D. Bodas, J. Song, M. Rajappa, and A. Hoffman, "Simple power-aware scheduler to limit power consumption by hpc system within a budget," in *2014 Energy Efficient Supercomputing Workshop*, 2014. doi: 10.1109/E2SC.2014.8 p. 21–30.

[10] K. Ahmed, J. Liu, and K. Yoshii, "Enabling demand response for hpc systems through power capping and node scaling," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018. doi: 10.1109/HPCC/SmartCity/DSS.2018.00133 p. 789–796.

[11] ISO Central Secretary, "Digital twin – concepts and terminology," International Organization for Standardization, Geneva, CH, Standard ISO/IEC 30173:2023, Nov. 2023. [Online]. Available: https://www.iso.org/standard/81442.html

[12] W. Brewer, M. Maiterth, V. Kumar, R. Wojda, S. Bouknight, J. Hines, W. Shin, S. Greenwood, D. Grant, W. Williams, and F. Wang, "A digital twin framework for liquid-cooled supercomputers as demonstrated at exascale," in *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024. doi: 10.1109/SC41406.2024.00029 p. 1–18.

[13] B. Jung, A. Kammeyer, V. Peltason, M. Ulbig, M. Wehming, and D. Hutzschenreuter, "Systems metrology in future cities – the example smart metrology campus (smc)," *Measurement: Sensors*, p. 101800, 2024. doi: 10.1016/j.measen.2024.101800. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917424007761

[14] A. Kammeyer, F. Burger, D. Lübbert, and K. Wolter, "Developing a digital twin to measure and optimise hpc efficiency," *Measurement: Sensors*, p. 101481, 2024. doi: 10.1016/j.measen.2024.101481. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917424004574

[15] ——, "Determining data centre pue with a digital twin," in *Sensor and Measurement Science International*, ser. SMSI 2025. AMA Service GmbH, May 2025. doi: 10.5162/SMSI2025/A7.1 p. 71–72.

[16] B. Bylina, J. Bylina, and M. Piekarz, "Impact of processor frequency scaling on performance and energy consumption for wz factorization on multicore architecture," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35, 2023. doi: 10.15439/2023F6213 p. 377–383.

[17] B. Bylina and M. Piekarz, "The scalability in terms of the time and the energy for several matrix factorizations on a multicore machine," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35, 2023. doi: 10.15439/2023F3506 p. 895–900.

[18] A. Krzywaniak, J. Proficz, and P. Czarnul, "Analyzing energy/performance trade-offs with power capping for parallel applications on modern multi and many core processors," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2018. doi: 10.15439/2018F177 p. 339–346.

[19] Raspberry Pi Ltd., "How to build a Raspberry Pi cluster," May 2025. [Online]. Available: https://www.raspberrypi.com/tutorials/cluster-raspberry-pi-tutorial/

[20] SchedMD LLC, "Slurm power saving guide," May 2025. [Online]. Available: https://slurm.schedmd.com/power_save.html

[21] A. Petitet, R. C. Whaley, J. Dongarra, and A. Cleary, "Hpl - a portable implementation of the high-performance linpack benchmark for distributed-memory computers," Dec. 2018, version 2.3. [Online]. Available: https://www.netlib.org/benchmark/hpl/

[22] J. Dongarra, M. A. Heroux, and P. Luszczek, "High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems," *The International Journal of High Performance Computing Applications*, vol. 30, no. 1, p. 3–10, 2016. doi: 10.1177/1094342015593158. [Online]. Available: https://doi.org/10.1177/1094342015593158

[23] H. G. Weller, G. Tabor, H. Jasak, and C. Fureby, "A tensorial approach to computational continuum mechanics using object-oriented techniques," *Computer in Physics*, vol. 12, no. 6, p. 620–631, 11 1998. doi: 10.1063/1.168744. [Online]. Available: https://doi.org/10.1063/1.168744

[24] S. Agostinelli, J. Allison, K. Amako *et al.*, "Geant4—a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, p. 250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900203013688

[25] N. Sudermann-Merx, *Fortgeschrittene Modellierungstechniken*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, p. 161–193. ISBN 978-3-662-67381-2. [Online]. Available: https://doi.org/10.1007/978-3-662-67381-2_7

[26] D. Kolossa and G. Grübel, "Evolutionary computation and nonlinear programming in multi-model-robust control design," in *Real-World Applications of Evolutionary Computing*, S. Cagnoni, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. ISBN 978-3-540-45561-5 p. 147–157.

[27] "Rpi 4 consumes 2.5w when shut down," May 2025. [Online]. Available: https://raspberrypi.stackexchange.com/questions/104944/rpi-4-consumes-2-5w-when-shut-down

[28] R. P. Becker, "Entwurf und implementierung eines plugins für slurm zum planungsbasierten scheduling," Bachelor's thesis, Freie Universität Berlin, Berlin, 2021.

[29] D. G. Feitelson, "Packing schemes for gang scheduling," in *Job Scheduling Strategies for Parallel Processing*, D. G. Feitelson and L. Rudolph, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. ISBN 978-3-540-70710-3 p. 89–110.

[30] D. G. Feitelson and M. A. Jettee, "Improved utilization and responsiveness with gang scheduling," in *Job Scheduling Strategies for Parallel Processing*, D. G. Feitelson and L. Rudolph, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. ISBN 978-3-540-69599-8 p. 238–261.

[31] S. J. Chapin, W. Cirne, D. G. Feitelson, J. P. Jones, S. T. Leutenegger, U. Schwiegelshohn, W. Smith, and D. Talby, "Benchmarks and standards for the evaluation of parallel job schedulers," in *Job Scheduling Strategies for Parallel Processing*, D. G. Feitelson and L. Rudolph, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999. doi: 10.1007/3-540-47954-6_4. ISBN 978-3-540-47954-3 p. 67–90.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 185–193
DOI: 10.15439/2025F1850
ISSN 2300-5963 ACSIS, Vol. 43

Assigning scientific texts to existing ontologies

Lukáš Korel
0000-0002-4071-0360
Faculty of Information Technology
Czech Technical University in Prague
Thákurova 2700/9,
160 00 Praha 6, Czech Republic
Email: Lukas.Korel@fit.cvut.cz

Martin Holeňa
0000-0002-2536-9328
Institute of Computer Science,
Czech Academy of Sciences,
Pod Vodárenskou věží 271/2,
182 00 Praha 8, Czech Republic
Email: martin@cs.cas.cz

***Abstract*—Humans try to help computers understand the properties of the real world, and ontologies can be used for this task. Manual enhancement of ontologies is highly time-consuming for domain experts. This paper proposes a solution to match a scientific text to the most relevant ontology using artificial neural networks. Our approach selects a paragraph or a sentence, uses representation learning to embed it into a vector space by some embedder, and measures its relevance to embedded textual properties from the selected ontology by a modified version of a Siamese neural network. We have considered different embedders, their quality has been evaluated on a use case with available ontologies from several application domains.**

***Index Terms*—ontology, neural network, embedding, text matching, LLM**

I. Introduction

Knowledge graphs and ontologies are graph-based structures for knowledge representation. They are defined as "a formal specification of a shared conceptualization" [1], provide a standardized description formalism to express knowledge, facilitating its exchange. Knowledge is typically related to some specific domain. Knowledge graphs and ontologies are commonly used to increase understanding of the related domain.

A domain ontology serves as a framework defining fundamental concepts, such as classes, attributes, and relationships, within a specific knowledge domain. Their definitions encapsulate information about their meaning and constraints. Classes can be defined through annotations or by interconnecting them with properties. Each domain ontology typically employs domain-specific terms to denote its primitives.

The manual construction of classes and their relations is too demanding for the time of domain experts, and automation could bring large amounts of time savings. The automated construction of an ontology consists of several steps:

1) Parse the source texts and extract only relevant content
2) Extract content from the ontologies defining its domain
3) Represent the extracted contents in the same vector space
4) Find the ontology most relevant to the source text
5) Locate the new knowledge in the structure of the ontology
6) Insert the new knowledge, and align it with the ontology internal structure, e.g., via removing duplicities

This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/208/OHK3/3T/18.

This paper deals with the issue of automatic ontology design only by assigning articles to ontologies, which corresponds to points 3 and 4. It represents both, the content extracted from ontologies, and the given text in the same vector space and uses an artificial neural network (ANN) to find the ontology most relevant to the given research text.

The next section recalls related work. Section III describes the methodology of text preprocessing and text representation learning, and our architecture of the neural network for the task of text-to-ontology matching. In Section IV, the proposed methodology is validated in a case study by applying it to scientific data.

II. Related Works

In connection with learning and extending ontologies, ANNs have been primarily used for the identification of concepts, relations, and attributes [2], [3], [4]. With respect to relations, some ANN-based methods have been developed specifically for subsumption relations needed for the construction of taxonomies [5], [6], [7], [8]. In connection with the integration of ontologies, they have been used primarily for ontologies matching [9], [10], [11], [12], [13]. The variety of employed ANN types is rather large. It includes traditional multilayer perceptrons [14], adaptive resonance theory networks [15] and associative memories [16], as well as the modern deep convolutional networks [9], [17], deep belief networks [2], long short-term memory (LSTM) networks together with their bidirectional variant (BiLSTM) [18] and gated recurrent units networks [19], [20].

The dependence of ontologies on texts led to using networks developed for text and natural language representation learning, most importantly BERT [21], [22], the bidirectional encoder representations from transformers, and word2vec [23], the most traditional network for embedding text into a Euclidean space. The close relationship of ontologies to knowledge graphs led also to the use of RDF2Vec [7], [20], which was originally proposed for knowledge graphs [24]. Based on similar principles as word2vec and RDF2Vec, OWL2Vec for embedding ontologies [25] and recently OWL2Vec4OA (OWL2Vec for ontology alignment) [26] were proposed. The

most related paper to our assigning scientific texts to existing ontologies solution is [27]. It presents an attempt to solve a text-to-ontology mapping problem by utilizing NLP tools and neural networks, and assess the final result through visualizing the latent space and exploring the mappings between an input text and ontology classes. Finally, the graph-like structure of ontologies brought the application of graph neural networks [11], [12].

Closest to the proposed approach is the way ANNs have been used in connection with translating into OWL [28], [19], with predicate chaining and restriction [16], and with taxonomy extraction from knowledge graphs [7]. In [28], ontology learning is tailored as a transductive reasoning task that uses two recurrent neural networks to translate text in natural language into OWL specifications in description logic. That approach was further developed in [19], resulting in a system based on a single recurrent network of GRU type. It uses an encoder-decoder configuration and translates a subset of natural language into the description logic language ALLQ through syntactic transformation. Moreover, the system generalizes over different syntactic structures and has the ability to tolerate unknown words through copying input words as extralogical symbols to the output, as well as the ability to enrich the training set with new annotated examples.

In [16], a mapping is established between ontologies and a pair of interacting associative memories. One of them stores assertions, and the other stores entailment rules. The recent work [7] describes a method for the specific task of extracting a taxonomy from an embedding of a knowledge graph. Over that embedding, which can be obtained for example with RDF2Vec, hierarchical agglomerative clustering is performed, first without using type information, and then injecting types into the hierarchical clustering tree.

III. Methodology

This section describes details of the employed methods. In the first part, we address processing textual files, splitting them into paragraphs, and keeping only paragraphs fulfilling minimal length and content relevant to the document's topic. The second part describes the textual embedders considered in assigning texts to ontologies. The final part describes the learning of similarity between an ontology and a given text.

A. Text Preprocessing

The texts used in the case study in Section IV have been extracted from scientific PDFs (Portable Document Files) and the considered ontologies. To read and transform ontologies, we have used the Protégé tool [29], which can understand many commonly used formats for ontology representation. We have chosen the Ontology Web Language (OWL) as the target format and extracted it using the Python package owlready2 [30]. As a representation of each ontology, we use textual content in natural language (e.g., label, comment, description, definition, and note) extracted from their formal representations. Contents shorter than five words were removed because of low context information.

B. Text Representation Learning

For typical downstream data analysis tasks, such as classification or clustering, it is suitable to embed words or other parts of text in a Euclidean space. This representation is mostly the result of representation learning using ANN-based embedders. In our research, we consider the following embedders:

1) Global Vectors for Word Representation (GloVe): [31] is an unsupervised learning algorithm that captures word semantics by leveraging global statistical information. It constructs a word-context co-occurrence matrix and then factorizes it to obtain word embeddings. Each embedding represents a word's position in the semantic space based on its statistical relationships with other words. However, to create embeddings for entire texts, an aggregation process is necessary, which relies on averaging the individual word embeddings based on the TF-IDF (Term Frequency-Inverse Document Frequency) weights.

2) InferSent: [32] is designed for creating universal sentence representations through supervised learning on natural language inference data. It approaches the task by using a carefully crafted training set that includes pairs of sentences with labeled relationships (entailment, contradiction, or neutral). InferSent's embeddings are learned by predicting relationships between sentences. This approach generates sentence embeddings that encapsulate semantic relationships, making it suitable for various downstream natural language processing tasks.

3) Doc2vec-DM: , i.e., the distributed memory version of Doc2Vec [33], is an extension of Word2Vec [34] tailored for document-level embeddings. It operates by learning distributed representations of documents, considering both word and document context. The model can provide either a single embedding per document or embeddings for individual words within the document. This flexibility makes it versatile for various applications, such as document similarity analysis or content-based document retrieval.

4) Bidirectional Encoder Representation from Transformers (BERT): [35] captures contextual information bidirectionally, considering both the left and right context of each word in a sentence. BERT embeddings are contextualized and can be used at both word and sentence levels. Its variant SciBERT is specialized for scientific literature and the variant Sentence-BERT focuses on generating embeddings tailored for entire sentences. Each version of BERT addresses specific use cases, offering state-of-the-art performance across a range of natural language processing tasks.

5) Llama 3.1 + LLM2vec: uses the large language model (LLM) Llama 3.1 [36] published in July 2024, with various model sizes. We have chosen the 8B model, which fits into commonly available GPUs. Llama 3.1 is an auto-regressive language model that uses an optimized transformer architecture. We use it for representation learning using LLM2Vec [37], which enables bidirectional attention, trains the model with masked next-token prediction, and performs unsupervised contrastive learning. The model can be further fine-tuned to achieve state-of-the-art performance in specific cases.

C. Network Learning

Our proposed architecture is inspired by Siamese neural networks [38], which are widely used for tasks involving the estimation of the similarity between two inputs, such as in face verification or semantic textual similarity. The key motivation behind this inspiration lies in the core functionality of Siamese networks: they encode two inputs into a shared embedding space using twin subnetworks with shared weights, and then compute a distance or similarity measure between these embeddings.

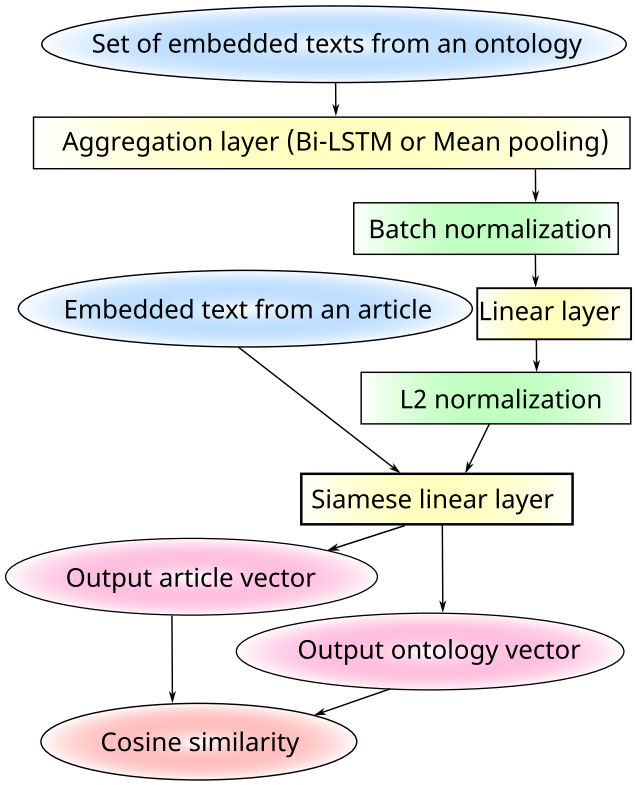


Fig. 1. Neural network schema illustrating used layers (represented with boxes) and flow of the embedded inputs from an ontology and an article to the final cosine distance between them.

In our case, the task requires assessing the semantic distance between a source text and an ontology concept. This parallels the role of Siamese networks, where the goal is to quantify how similar or different two representations are. By adopting this paradigm, our architecture learns to embed both the input text and ontology representation into a common space and returns a scalar distance that reflects their semantic alignment. Our neural network takes an embedded text as an anchor input and a list of embedded, randomly shuffled representative descriptions from an ontology as a representative input. As an output, it returns the distance of the embedded text and representative samples.

Figure 1 shows the internal architecture. The representative input dataset flows through the aggregator, which is represented by the Mean pooling layer or the bidirectional LSTM (long-short-term memory) with a training dropout equal to 0.2 layer in our experiments. The whole batch is then normalized, which is known to improve LSTM results [39]. The following linear layer reduces size and its output is L2 normalized. The following Siamese linear layer is the core of this architecture, which takes the anchor and the output from the second branch and returns for each of them a numeric vector, its size is a quarter compared to the anchor vector. Finally, the cosine similarity between those vectors is computed. More details are provided in Algorithm 1.

IV. Case Study

This section describes the data used in the case study, its experimental setting, and the evaluation of the obtained results.

A. Used Data

The ontologies used in our experiments come from chemistry for both training and testing, all of them being somehow related to catalysis, whereas for testing there are also ontologies from some more distant application domains. Table I shows the details of the considered ontologies.

TABLE I
THIS TABLE SHOWS THE CONSIDERED ONTOLOGIES, THEIR COUNTS OF TEXTUAL DESCRIPTIONS OF CLASSES AND RELATIONS, AND SCIENTIFIC AREAS TO WHICH THEY ARE RELATED.

Area	Ontology name	Count of items with textual definitions
Chemistry	Allotrope Foundation Ontology (AFO)	2894
Chemistry	Chemical Entities of Biological Interest (CHEBI)	176873
Chemistry	Chemical Methods Ontology (CHMO)	3084
Chemistry	Systems Biology Ontology (SBO)	694
Chemistry	National Cancer Institute Thesaurus (NCIT)	166212
Biology	Biological Collections Ontology (BCO)	671
Energy	Digital Construction Energy (DICES)	67
Environment	Environment Ontology (ENVO)	6605
Finance	Financial Industry Business Ontology (FIBO)	3362

B. Experimental Setting

Each anchor and representative input pair was extracted from the ontologies because we do not have other texts with known ground truth. Each extracted pair was required to fulfill the condition that the anchor text must not appear in the representative descriptions. For our experiment, we selected the considered ontologies and divided them into two parts. From the chemistry group, 80 % of the textual descriptions were randomly selected for training and the rest for testing on similar domains. The remaining ontologies, from more distant domains, were completely used for testing. From the training

Algorithm 1 Training the Neural Network (NN) with Texts from Ontologies and Anchor Texts

Require: Dataset $\mathcal{D} = \{(O_i, a_i, y_i)\}$, where O_i is a list of texts, a_i is the anchor text, and $y_i \in \{0, 1\}$ is the similarity label.
Require: Embedder EM, assigning to each input a vector of size Features.
Require: Aggregator A_ϕ (Bidirectional LSTM or Mean pooling) and Siamese model f_ψ, with trainable parameters ϕ and ψ, an optimizer minimizing a loss function $\mathcal{L}$.
Require: Hyperparameters: number of epochs E and batch size BS.
Initialize parameters ϕ and ψ of the A_ϕ and f_ψ.
Split the dataset $\mathcal{D}$: $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{tr}} \cap \mathcal{D}_{\text{val}} = \emptyset$
Set Siamese network output size: OutputSize $=$ Features$/4$.
Initialize bestLoss $\leftarrow \infty$ and bestModel $\leftarrow (\phi, \psi)$.
for epoch $= 1$ to E **do**
  Shuffle the dataset $\mathcal{D}_{\text{tr}}$ and divide it into batches of size BS.
  for each batch $\mathcal{B} \subset \mathcal{D}_{\text{tr}}$ **do**
    for each $(O_i, a_i, y_i) \in \mathcal{B}$ **do**
      Aggregate branch:
      for each text $o \in O_i$ **do**
        Compute embedding: $h_o = EM(o)$.
      end for
      Aggregate embeddings using selected aggregator: $h_{O_i} = \mathrm{A}_\phi(\{h_o\})$.
      Apply 1D batch normalization to h_{O_i}, yielding $h_{O_i}^{\text{norm}}$.
      Pass $h_{O_i}^{\text{norm}}$ through a fully connected layer, yielding $h_{O_i}^{\text{fc}}$.
      Apply L2 normalization: $h_{O_i}^{\text{L2}} = \frac{h_{O_i}^{\text{fc}}}{\|h_{O_i}^{\text{fc}}\|_2}$.
      Anchor branch:
      Compute an embedding for the anchor text: $h_{a_i} = EM(a_i)$.
      Siamese network:
      Compute the outputs: $z_i^{\text{agg}} = f_\psi(h_{O_i}^{\text{L2}})$, $z_i^{\text{anchor}} = f_\psi(h_{a_i})$.
      Compute the cosine similarity o_i of $z_i^{\text{agg}}, z_i^{\text{anchor}}$.
      Compute the loss: $\ell = \mathcal{L}(o_i, y_i)$.
    end for
    Update the parameters ϕ and ψ using the optimizer to minimize ℓ for the batch.
  end for
  Compute the validation loss ℓ_{val} for the current model on $\mathcal{D}_{\text{val}}$.
  if $\ell_{\text{val}} <$ bestLoss **then**
    Update bestLoss $\leftarrow \ell_{\text{val}}$ and bestModel $\leftarrow (\phi, \psi)$.
  end if
end for
return bestModel with parts A_ϕ and f_ψ corresponding to the least validation loss.

data, 20 % of textual descriptions were randomly selected as a validation dataset, and the rest as a training dataset. Both datasets were resampled to keep the same proportion of data from each ontology, to mitigate overfitting with texts from one ontology.

For the validation of our model, we have considered the hyperparameters listed in Table II. Marked bold is the combination selected on the validation dataset by the contrastive loss.

$$L_{\text{contrastive}}(x_1, x_2, y) = \frac{1}{2} y \cdot d_{\cos}(x_1, x_2)^2 + \frac{1}{2}(1 - y) \cdot \max(0, 1 - d_{\cos}(x_1, x_2))^2 \quad (1)$$

In the above definition:

- x_1 and x_2 are the outputs of the network for which the contrastive loss is computed, namely the x_1 for the embedded anchor text and the x_2 for the embedded representative list of texts from the selected ontology
- y denotes whether the anchor on the input belongs to the selected ontology ($y = 1$) or does not belong to it ($y = 0$)
- $d_{\cos}(x_1, x_2)$ is the cosine distance between x_1 and x_2

C. Results

All results in this case study were obtained using an independent testing dataset. Writing T for True, F for False, P for positive, N for negative, $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$, the definitions of the employed quality measures are as follows:

$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$,

TABLE II
CONSIDERED COMBINATIONS OF HYPERPARAMETERS BATCH SIZE (BS), OPTIMIZER (OPT), AND LEARNING RATE (LR), IN BOLD IS THE COMBINATION SELECTED ON THE VALIDATION DATASET BASED ON THE VALUE OF THE CONTRASTIVE LOSS (CL)

BS	OPT	LR	CL
8	Adam	0.1	0.126
8	Adam	0.01	0.126
8	Adam	0.001	0.126
8	SGD	0.1	0.125
8	SGD	0.01	0.126
8	SGD	0.001	0.125
16	Adam	0.1	0.129
16	Adam	0.01	0.081
16	Adam	0.001	0.089
16	SGD	0.1	0.066
16	SGD	0.01	0.063
16	SGD	0.001	0.065
32	Adam	0.1	0.129
32	Adam	0.01	0.124
32	Adam	0.001	0.125
32	SGD	0.1	0.047
32	**SGD**	**0.01**	**0.044**
32	SGD	0.001	0.046

$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, where $Precision = \frac{TP}{TP+FP}$ and $Recall = TPR$

$\text{AUC} = \sum_{i=1}^{n} (\text{FPR}[i] - \text{FPR}[i-1]) \left(\frac{\text{TPR}[i]+\text{TPR}[i-1]}{2} \right)$, where i indexes all the values of the prediction score, which is a numerical output of the model that represents the confidence that a sample belongs to the positive class. The prediction score is sorted and thresholds are applied to generate pairs of TPR and FPR values, spanning the range from the minimum to the maximum prediction score.

1) Statistics from ontologies for near and far domains:

The results in Table III for descriptions from ontologies with domains near the training data are best for embeddings obtained with the SciBERT in the model with LSTM aggregation. This may be due to SciBERT having been fine-tuned by scientific data. The worst embedder was Doc2Vec-DM. The results with Doc2Vec and GloVe have been expected because these solutions are based on simple textual embedders. The variant of the model with the mean pooling aggregation layer has even better results. The best results have been achieved with the Llama3.1+LLM2vec embedder without fine-tuning to a specific domain or task. Thanks to the fact that texts are the primary kind of data on which LLMs have been trained, our model had better performance in the considered quality measures.

These results were compared with our previous research [40], where we solved the text-to-ontology matching task. We used the *chemical-bert-uncased* [41] for embedding input texts. These embeddings were used as inputs to the following classifiers: Gaussian process, k-nearest neighbor, multilayer perception, random forest, and support vector machine. Among them, the multilayer perceptron classifier is the most similar to our ANN approach. The approach in [40] has not achieved as high values for the three considered quality measures as have been achieved in the experiments reported in this paper with the proposed ANN architecture, especially in combination with the SciBERT embedder. The best result in [40] were achieved by the Gaussian process classifier, compared to which the SciBERT-based solution reported here has accuracy better by 0.9, AUC by 5.2, and F1 by 8.9 percentile points in the LSTM variant, and in the mean pooling variant, accuracy is better by 1.8, AUC by 6.7, and F1 by 9.8 percentile points. The classification we performed in [40] was restricted to an a priori chosen set of chemical catalytic ontologies, whereas the approach we propose here is independent of the number of potential target ontologies.

TABLE III
RESULTS FOR ONTOLOGIES WITH DOMAINS NEAR TO THE TRAINING DATA COMPARED TO RESULTS FROM [40] BASED ON CLASSIFIERS

ANN LSTM solution	Accuracy	AUC	F1
Doc2Vec-DM	83.3 %	73.9 %	83.2 %
TF-IDF GloVe	84.8 %	76.3 %	84.7 %
InferSent	95.1 %	92.3 %	95.0 %
Llama3.1+LLM2vec	94.3 %	92.0 %	94.3 %
SciBERT	**98.2 %**	**97.1 %**	**98.1 %**
SentenceBERT	95.0 %	92.2 %	95.0 %
ANN Mean pooling solution	Accuracy	AUC	F1
Doc2Vec-DM	92.7 %	88.5 %	92.6 %
TF-IDF GloVe	95.9 %	93.6 %	95.8 %
InferSent	98.9 %	98.3 %	98.8 %
Llama3.1+LLM2vec	**99.3 %**	**98.9 %**	**99.3 %**
SciBERT	99.2 %	98.7 %	99.1 %
SentenceBERT	98.7 %	98.0 %	98.7 %
Classifiers-based solution	Accuracy	AUC	F1
Gaussian process	**97.5 %**	**92.2 %**	**89.5 %**
K-nearest neighbor	96.7 %	90.4 %	87.6 %
Multi-layer perception	97.0 %	91.0 %	87.8 %
Random forest	94.6 %	82.0 %	85.9 %
Support vector machine	97.2 %	91.6 %	88.7 %

TABLE IV
RESULTS FROM ONTOLOGIES WITH DOMAINS FAR FROM THE TRAINING DATA

ANN LSTM solution	Accuracy	AUC	F1
Doc2Vec-DM	67.0 %	55.9 %	66.8 %
TF-IDF GloVe	63.4 %	51.2 %	63.1 %
InferSent	**70.0 %**	**60.0 %**	69.6 %
Llama3.1+LLM2vec	60.2 %	50.9 %	59.3 %
SciBERT	63.5 %	51.3 %	63.2 %
SentenceBERT	69.9 %	59.9 %	**69.9 %**
ANN Mean pooling solution	Accuracy	AUC	F1
Doc2Vec-DM	67.5 %	56.6 %	67.3 %
TF-IDF GloVe	68.9 %	58.5 %	68.0 %
InferSent	69.7 %	59.6 %	**69.4 %**
Llama3.1+LLM2vec	62.7 %	50.3 %	62.2 %
SciBERT	**70.7 %**	**61.0 %**	67.8 %
SentenceBERT	63.7 %	51.6 %	62.9 %

The results in Table IV for texts from ontologies with

domains far from the training data are best for embeddings obtained with InferSent and SentenceBERT for the case of LSTM aggregation. In the case of aggregation by mean pooling, the best results were achieved by the model with the SciBERT and InferSent embedders. We did not expect this behavior because LLMs can handle long inputs and understand their semantics pretty well. These results show that embedders that are not based on LLMs are still very useful. This behavior may be caused by the fact that the LLMs are not primarily optimized to create textual embeddings.

2) Statistical significance tests:

The differences between the considered embedders were tested for significance by the Friedman test. The basic null hypotheses that each of the AUC, accuracy, and F1 are for all 6 embedders the same, were rejected for the LSTM aggregation, with the achieved p-values 1×10^{-2}, 3×10^{-2}, and 2×10^{-2}, respectively. In case of mean pooling based aggregation, the achieved p-values were 1×10^{-1}, 3×10^{-2}, and 1×10^{-2}, so we are not able to reject the hypotheses for the AUC metric. As to the post-hoc analysis, because we agree with the opinion of the authors of [42], we followed their proposal to use the Wilcoxon signed rank test with the two-sided alternative for all pairs of compared embedders. For the correction due to multiple hypotheses testing, we used the Holm method. The results are given in Table V and Table VI, the best results by the total score have embeddings from InferSent and both BERTs in the case of LSTM aggregation, and from InferSent and Llama3.1 in the case of mean pooling.

V. Conclusion

In this paper, we proposed an approach to assigning scientific texts to ontologies. This solution contributes to the complex process of automated knowledge processing and highlights the possibilities for research on enhancing existing ontologies. We used neural networks for text embedding and for comparison of obtained embeddings to automatically determine the ontologies most relevant to texts. To this end, we adapted a Siamese neural network in combination with representation learning by various embedders, namely GloVe with TF-IDF weighted aggregation, SentenceBERT, SciBERT, InferSent, Doc2Vec-DM and Llama 3.1 with LLM2vec framework. The embedding of given text serves as the input to one branch of the Siamese neural network, and the list of embeddings of textual descriptions from the considered ontology serves as the input to the second branch. The solutions based on Doc2Vec-DM and GloVE were unsuccessful. In this approach, the most successful combination in ontologies from highly diverse domains was the InferSent embedder, which takes into account contextual relations and produces embeddings with the highest dimension. It achieved good results in both test cases. Our results show that the proposed method surpasses an earlier solution based on a multilayer perceptron classifier. However, challenges remain in the extraction of knowledge from ontologies and in merging new knowledge with existing ontology. In our future research, we want to address these challenges using graph neural networks. Our motivation for such an approach is the graph structure of ontologies.

Acknowledgment

The research reported in this paper has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/208/OHK3/3T/18.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic and by the ELIXIR-CZ project (ID:90255), part of the international ELIXIR infrastructure.

References

[1] W. N. Borst, "Construction of engineering ontologies for knowledge sharing and reuse," Enschede Centre for Telematics and Information Technology (CTIT), Tech. Rep., 1997.

[2] F. Al-Aswadi, H. Chan, and K. Gan, "Extracting semantic concepts and relations from scientific publications by using deep learning," in *Proceedings of IRICT 2020*, 2021, pp. 374–383.

[3] N. Gupta, S. Podder, K. Annervaz, and S. Sengupta, "Domain ontology induction using word embeddings," in *ICMLA*, 2016, pp. 115–119.

[4] A. Katyshev, A. Anikin, M. Denisov, and T. Petrova, "Intelligent approaches for the automated domain ontology extraction," in *International Congress on Information and Communication Technology*, 2021, pp. 410–417.

[5] S. Althubaiti, S. Kafkas, M. Abdelhakim, and R. Hoehndorf, "Combining lexical and context features for automatic ontology extension," *Journal of Biomedical Semantics*, vol. 11, p. article no. 1, 2020.

[6] L. Espinoza-Anke, F. Ronzano, and H. Saggion, "Hypernym extraction: Combining machine-learning and dependency grammar," in *CICLing*, 2015, pp. 372–383.

[7] F. Martel and A. Zouaq, "Taxonomy extraction using knowledge graph embeddings and hierarchical clustering," in *SAC'21*, 2021, pp. 836–844.

[8] N.-A. R., J. R., and J. Castro, "Automated ontology extraction from unstructured texts using deep learning," in *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications*. Springer, 2020, pp. 727–755.

[9] A. Bento, A. Zouaq, and M. Gagnon, "Ontology matching using convolutional neural networks," in *LREC*, 2020, pp. 5648–5653.

[10] J. Chakraborty, B. Yaman, L. Virgili, K. Konar, and S. Bansal, "OntoConnect: Results for OAEI 2020," in *OM ISWC*, 2020, pp. 204–210.

[11] L. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan *et al.*, "MEDTO: Medical data to ontology matching using hybrid graph neural networks," in *KDD'21*, 2021, pp. 2946–2954.

[12] J. Wu, J. Lv, H. Guo, and S. Ma, "Daeom: A deep attentional embedding approach for biomedical ontology matching," *Applied Sciences*, vol. 10, p. article no. 7909, 2020.

[13] A. Benarab, G. Yin, and F. Rafique, "An embedding-based approach to the ontology matching problem," in *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, 2023. doi: 10.1109/PRML59573.2023.10348247 pp. 642–649.

[14] S. Kolozali, G. Fazekas, M. Barthet, and M. Sandler, "A framework for automatic ontology generation based on semantic audio analysis," in *Audio Engineering Society International Conference*, 2014, pp. 87–96.

[15] M. Hourali and G. Montazer, "Using ART2 neural network and bayesian network for automating the ontology constructing process," *Procedia Engineering*, vol. 29, pp. 3914–3923, 2012.

[16] C. Mercier, H. Chateau-Laurent, F. Alexandre, and T. ViĂ©ville, "Ontology as neuronal-space manifold: Towards symbolic and numerical artificial embedding," in *Workshop on Knowledge Representation for Hybrid and Compositional AI*, 2021, pp. 1–11.

[17] G. Li, "CNN based ontology learning algorithm and applied in PE data," *IAENG International Journal of Computer Science*, vol. 48, pp. 1–8, 2021.

[18] R. Mueller and S. Abdullaev, "Deep cause: Hypothesis extraction from information systems papers with deep learning for theory ontology learning," in *Annual Hawaii International Conference on System Sciences*, 2019, pp. 6250–6259.

TABLE V
COMPARISON OF AUC, ACCURACY, AND F1 RESULTS ON ALL 9 CONSIDERED ONTOLOGY DATASETS FOR THE EMBEDDERS COMBINED WITH LSTM AGGREGATION. THE VALUES IN THE TABLE ARE COUNTS OF DATASETS, IN WHICH THE EMBEDDER IN THE ROW HAS A HIGHER VALUE OF THE CONSIDERED PERFORMANCE MEASURE COMPARED TO THE EMBEDDER IN THE COLUMN. IF THE DIFFERENCE IS SIGNIFICANT ON THE 5 % LEVEL THEN THE COUNT IS IN BOLD. THE COLUMN "TOTAL" CONTAINS THE SUM OF THE COUNTS IN THE ROW. THE LAST COLUMN CONTAINS THE MEANS ± STANDARD DEVIATIONS OVER ALL TESTING DATASETS OF THE RESULTS FOR THE CONSIDERED PERFORMANCE MEASURE.

AUC	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		1	3	**1**	4	**1**	10	0.605 ± 0.105
Sentence_BERT	8		7	6	**8**	6	35	0.723 ± 0.167
Llama3.1	6	2		2	5	3	18	0.684 ± 0.185
InferSent	**8**	3	7		**8**	4	30	0.716 ± 0.173
Doc2Vec-DM	5	**1**	4	**1**		4	15	0.614 ± 0.140
SciBERT	**8**	3	6	5	5		27	0.729 ± 0.204
Accuracy	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		2	4	**1**	6	**2**	15	0.844 ± 0.016
Sentence_BERT	7		6	5	7	2	27	0.891 ± 0.061
Llama3.1	5	3		4	6	2	20	0.875 ± 0.065
InferSent	**8**	4	5		**9**	5	31	0.888 ± 0.049
Doc2Vec-DM	3	2	3	**0**		**2**	10	0.848 ± 0.056
SciBERT	**7**	7	7	4	**7**		32	0.893 ± 0.056
F1	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		**2**	4	**1**	6	**1**	14	0.840 ± 0.023
Sentence_BERT	**7**		7	6	7	3	30	0.890 ± 0.062
Llama3.1	5	2		4	5	2	18	0.871 ± 0.066
InferSent	**8**	3	5		**9**	5	30	0.884 ± 0.054
Doc2Vec-DM	3	2	4	**0**		2	11	0.847 ± 0.053
SciBERT	**8**	6	7	4	7		32	0.886 ± 0.063

TABLE VI

COMPARISON OF AUC, ACCURACY, AND F1 RESULTS ON ALL 9 CONSIDERED ONTOLOGY DATASETS FOR THE EMBEDDERS COMBINED WITH MEAN POOLING. THE VALUES IN THE TABLE ARE COUNTS OF DATASETS, IN WHICH THE EMBEDDER IN THE ROW HAS A HIGHER VALUE OF THE CONSIDERED PERFORMANCE MEASURE COMPARED TO THE EMBEDDER IN THE COLUMN. IF THE DIFFERENCE IS SIGNIFICANT ON THE 5 % LEVEL THEN THE COUNT IS IN BOLD. THE COLUMN "TOTAL" CONTAINS THE SUM OF THE COUNTS IN THE ROW. THE LAST COLUMN CONTAINS THE MEANS ± STANDARD DEVIATIONS OVER ALL TESTING DATASETS OF THE RESULTS FOR THE CONSIDERED PERFORMANCE MEASURE.

AUC	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		3	2	1	6	2	14	0.756 ± 0.172
Sentence_BERT	6		2	3	5	5	21	0.757 ± 0.217
Llama3.1	7	7		7	5	8	34	0.783 ± 0.221
InferSent	8	6	2		**7**	5	28	0.789 ± 0.190
Doc2Vec-DM	3	4	4	**2**		3	16	0.699 ± 0.158
SciBERT	7	4	1	4	6		22	0.764 ± 0.217
Accuracy	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		4	3	2	**8**	2	19	0.903 ± 0.041
Sentence_BERT	5		4	3	**7**	3	22	0.904 ± 0.062
Llama3.1	6	5		6	**8**	6	31	0.914 ± 0.070
InferSent	7	6	3		**8**	7	31	0.917 ± 0.053
Doc2Vec-DM	**1**	**2**	**1**	**1**		2	7	0.881 ± 0.058
SciBERT	7	6	3	2	7		25	0.907 ± 0.062
F1	GloVe	Sentence_BERT	Llama3.1	InferSent	Doc2Vec-DM	SciBERT	Total	$Mean \pm Std$
GloVe		5	4	2	8	3	22	0.898 ± 0.051
Sentence_BERT	4		3	2	6	3	18	0.898 ± 0.069
Llama3.1	5	6		5	**8**	7	31	0.910 ± 0.074
InferSent	7	7	4		**9**	7	34	0.912 ± 0.059
Doc2Vec-DM	1	3	**1**	**0**		**2**	7	0.880 ± 0.059
SciBERT	6	6	2	2	**7**		23	0.899 ± 0.067

[19] G. Petrucci, M. Rospocher, and C. Ghindini, "Expressive ontology learning as neural machine translation," *Journal of Web Semantics*, vol. 52–53, pp. 66–82, 2018.

[20] J. Potoniec, "Learning OWL 2 property characteristics as an explanation for an RNN," *Bulletin of the Polish Academy of Sciences, Technical Sciences*, vol. 68, pp. 1481–1490, 2020.

[21] A. Memariani, M. Glauer, F. Neuhaus, T. Mossakowski, and J. Hatings, "Automated and explainable ontology extension based on deep learning: A case study in the chemical domain," in *3rd International Workshop on Data Meets Applied Ontologies*, 2021, pp. 1–16.

[22] A. Oba, I. Paik, and A. Kuwana, "Automatic classification for ontology generation by pretrained language model," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2021, pp. 210–221.

[23] N. Teslya and S. Savosin, "Matching ontologies with Word2Vec-based neural network," in *ICCSA*, 2019, pp. 745–756.

[24] P. Ristoski and H. Paulheim, "Rdf2vec: Rdf graph embeddings for data mining," in *International Semantic Web Conference*, 2016, pp. 498–514.

[25] A. Ritchie, J. Chen, L. Castro, D. Rebholz-Schuhmann, and E. JimÃ©nez-Ruiz, "Ontology clustering with OWL2Vec," in *DeepOntoNLP*, 2021, pp. 54–61.

[26] S. Teymurova, E. Jiménez-Ruiz, T. Weyde, and J. Chen, "Owl2vec4oa: Tailoring knowledge graph embeddings for ontology alignment," in *Sixth International Knowledge Graph and Semantic Web Conference*. Springer International Publishing, October 2024. ISSN 0302-9743. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/34052/

[27] U. Yorsh, A. S. Behr, N. Kockmann, and M. Holeňa, "Text-to-ontology mapping via natural language processing models," in *Conference on Theory and Practice of Information Technologies*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253161694

[28] G. Petrucci, C. Ghindini, and M. Rospocher, "Ontology learning in the deep," in *EKAW*, 2016, pp. 480–495.

[29] M. A. Musen, "The protégé project: a look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015. doi: 10.1145/2757001.2757003. [Online]. Available: https://doi.org/10.1145/2757001.2757003

[30] J.-B. Lamy, "Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies," *Artificial Intelligence In Medicine*, vol. 80, pp. 11–28, 2017.

[31] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014. doi: 10.3115/v1/D14-1162 pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[32] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017. doi: 10.18653/v1/D17-1070 pp. 670–680. [Online]. Available: https://aclanthology.org/D17-1070

[33] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *ArXiv*, vol. abs/1507.07998, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:7364264

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019. doi: 10.18653/v1/N19-1423 pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[36] Meta, "Llama-models/models/llama3_1 at meta-llama/llama-models," Jul 2024. [Online]. Available: https://github.com/meta-llama/llama-models/tree/main/models/llama3_1

[37] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "LLM2Vec: Large language models are secretly powerful text encoders," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=IW1PR7vEBf

[38] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, p. 25, 08 1993. doi: 10.1142/S0218001493000339

[39] J. Whang and A. Matsukawa, "Exploring batch normalization in recurrent neural networks," *Stanford Center for Professional Development*, 2019, accessed on 29 February 2024. [Online]. Available: https://jaywhang.com/assets/batchnorm_rnn.pdf

[40] L. Korel, A. S. Behr, M. Holeňa, and N. Kockmann, "Using artificial neural networks to determine ontologies most relevant to scientific texts," in *Proceedings of the 22nd Conference Information Technologies – Applications and Theory (ITAT 2022)*, ser. CEUR Workshop Proceedings, vol. 3226. CEUR-WS.org, 2022. ISSN 1613-0073 pp. 44–54. [Online]. Available: https://ceur-ws.org/Vol-3226/paper5.pdf

[41] R. A. Company, "Bert for chemical industry," https://huggingface.co/recobo/chemical-bert-uncased, 2022.

[42] A. Benavoli, G. Corani, and F. Mangili, "Should we really use post-hoc tests based on mean-ranks?" *Journal of Machine Learning Research*, vol. 17, pp. 1–10, 2016.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 195–205
DOI: 10.15439/2025F0110
ISSN 2300-5963 ACSIS, Vol. 43

Estimating the entropy of covering-based rough set approximation operators

Wannes De Maeyer, Steven Van Overberghe, Chris Cornelis
0009-0005-1597-1023
0000-0002-0818-468X
0000-0002-6852-4041
Department of Mathematics, Computer Science and Statistics
Ghent University,
Ghent, Belgium
Email: {wannes.demaeyer, steven.vanoverberghe, chris.cornelis}@ugent.be

Mauricio Restrepo
0000-0003-1007-353X
Department of Mathematics
Universidad Militar Nueva Granada,
Bogotá, Colombia
Email: mauricio.restrepo@unimilitar.edu.co

***Abstract*—Rough set theory provides a robust framework for dealing with inconsistent data by utilizing equivalence relations to group indiscernible instances. A significant extension of this framework is the concept of covering-based rough sets, where equivalence relations are replaced with coverings. The effectiveness of covering-based rough sets in practical applications largely depends on the choice of an appropriate covering. To guide this selection, various metrics have been proposed to evaluate the quality of coverings, including entropy. While entropy serves as a valuable measure, its exact computation is often prohibitively expensive. In this paper, we propose an efficient method for estimating the entropy of covering-based rough set approximation operators, making the metric more feasible for practical use. We assess the accuracy of these estimates through experiments on both synthetic coverings and coverings for real-world datasets, the latter constructed using the Mapper algorithm, from the field of topological data analysis.**

I. Introduction

IN 1982 Pawlak introduced rough set theory as a framework to handle possibly inconsistent data [1]. Rough set theory revolves around the notion of an information system: a couple $(U, \mathcal{A})$, where U is a universe of instances (objects, data samples) and $\mathcal{A}$ is a set of attributes for which these instances take values. The set $\mathcal{A}$ naturally gives rise to an equivalence relation E on U, by partitioning U into equivalence classes that have the same values for all the attributes in $\mathcal{A}$. Instances in each equivalence class are called indiscernible. Rough set theory then proceeds to provide lower and upper approximations of concepts that are represented by a set $A \subseteq U$. The lower approximation is the union of all equivalence classes that are contained entirely within A. This can be viewed as the set of instances that certainly belong to A (every instance in the lower approximation is only indiscernible to elements of A). The upper approximation is the union of all equivalence classes that have non-empty intersection with A. Semantically this is the set of instances that possibly belong to U (every instance in the upper approximation is indiscernible with an element in A). Sets $A \subseteq U$ for which the lower approximation and upper approximation equal A itself are called consistent, or exact. It can be verified that the lower approximation is the largest consistent subset of A, while the upper approximation is the smallest consistent set that includes A.

Various researchers have proposed generalizations of classical rough sets by replacing the equivalence relation with a covering of U [2, 3, 4], resulting in a more flexible framework. Recall that a covering of U is any collection of subsets of U whose union is equal to U. By contrast to a partition, elements of a covering are allowed to overlap. This is useful in various situations, for example:

- handling missing data [5]: when the value of an instance for an attribute is unknown, it is considered indiscernible from any other (known or unknown) value.
- tolerance-based rough sets [6]: when dealing with numerical data, two objects are often considered indiscernible when their distance is lower than a given threshold.

Clearly, such indiscernibility relations are no longer transitive. These covering-based rough sets have been studied extensively from a theoretical perspective ([7, 8, 9, 10]) and used in many applications, in particular concerning attribute reduction [11, 12, 13]. However, selecting the right covering for an application can be hard because there may be multiple important factors to consider. One such factor is the granularity of the covering. This measures how fine the covering is and how detailed the rough set approximations are; it can also be interpreted as the amount of information that remains when the approximation operators are executed. To find coverings that have a certain granularity, the concept of entropy (which was originally introduced as a measure of information content) was extended to rough set approximations in [14]. Many applications in rough set theory use the notion of entropy [15, 16]. However, when it comes to covering-based rough set theory, it is computationally very expensive to calculate this naively since it takes into account all of the possible subsets of the universe U. For this reason, we propose an efficient way to estimate the entropy.

In [17] a promising method to generate coverings of datasets using the Mapper algorithm from the field of topological data analysis [18] was introduced. We use these coverings

to evaluate the convergence of our estimates. We choose this method because it produces coverings that closely resemble the original dataset topologically. Moreover, one can easily control several aspects of the coverings, such as the number of subsets in the covering, and the amount of overlap between them.

The remainder of this paper is structured as follows. We first recall the concepts of covering-based rough sets and entropy in Sections 2 and 3. Then, in Section 4, we outline a method to estimate the entropy of a given covering using the so-called strong granule-based approximations from covering-based rough set theory. To obtain the estimation, we first consider the method from [19] which allows us to estimate the entropy of any probability distribution. However, we observe that it is restrictively slow for our purposes. Therefore, we adapt this method using some properties of covering-based rough sets to find better estimates that are calculated more quickly. In Section 5, we evaluate the convergence of our estimates experimentally on a number of synthetic and benchmark datasets. Finally, in Section 6, we give a discussion and introduce some ideas for future exploration.

II. Covering-based rough sets

When working with real-world classification problems, inconsistencies often occur in datasets; intuitively, this happens when two instances are indiscernible but belong to different classes. Pawlak introduced rough sets in 1982 to handle such inconsistencies [1]. In Pawlak's seminal work, an approximation space is defined as an ordered pair (U, E), where E is an equivalence relation that groups indiscernible instances together over the universe U. Based on this setup, the lower and upper approximations of a set can be described in several equivalent ways. In this paper, we adopt the granule-based approach:

$$\underline{apr}(A) = \{[x]_E \in U/E \mid [x]_E \subseteq A\}$$
$$\overline{apr}(A) = \{[x]_E \in U/E \mid [x]_E \cap A \neq \emptyset\}$$

For a subset A of U, its rough set approximation is represented as the pair $(\underline{apr}(A), \overline{apr}(A))$. These approximation operators are dual in the sense that $\underline{apr}(A) = \overline{apr}(A^c)^c$, where ^{c} refers to the classical set complement. These approximations have been used in many applications ranging from classification [20] (especially rule extraction [21, 22]) to feature selection [23, 24, 25].

Classical rough set theory often struggles with continuous features, as the indiscernibility relation may no longer satisfy the properties of an equivalence relation. To overcome this limitation, various generalizations of classical rough sets have been proposed. One such generalization are covering-based rough sets, which replace the partition given by the indiscernibility relation with a covering. This defines a covering approximation space: an ordered pair $(U, \mathbb{C})$, where $\mathbb{C}$ is a covering of U, that is, a collection of subsets whose union equals U. While it is tempting to generalize Pawlak's definitions by replacing the equivalence classes $[x]_E \in U/E$ with elements $K \in \mathbb{C}$, this substitution does not preserve the duality between the approximations. To address this, Yao [2] introduced two dual pairs of granule-based approximations for covering approximation spaces.

$$\underline{apr'}_{\mathbb{C}}(A) = \bigcup\{K \in \mathbb{C} : K \subseteq A\} \tag{1}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \underline{apr'}_{\mathbb{C}}(A^c)^c \tag{2}$$
$$\underline{apr''}_{\mathbb{C}}(A) = \overline{apr''}_{\mathbb{C}}(A^c)^c \tag{3}$$
$$\overline{apr''}_{\mathbb{C}}(A) = \bigcup\{K \in \mathbb{C} \mid K \cap A \neq \emptyset\} \tag{4}$$

Equations (1) and (2) are called the strong approximation operators, while Equations (3) and (4) are referred to as the weak approximations. The rationale behind this terminology is explained by the following property:

$$\underline{apr''}_{\mathbb{C}}(A) \subseteq \underline{apr'}_{\mathbb{C}}(A) \subseteq A \subseteq \overline{apr'}_{\mathbb{C}}(A) \subseteq \overline{apr''}_{\mathbb{C}}(A)$$

Example 1. *Let $U = \{x_1, \ldots, x_9\}$ be a universe which has the following covering, $\mathbb{C} = \{\{x_0, x_1\}, \{x_1, x_2\}, \{x_2, x_5\}, \{x_3, x_4, x_7\}, \{x_6, x_7, x_8, x_9\}\}$ and $A = \{x_1, x_2, x_3, x_4, x_5\}$. A has the following approximations:*

$$\underline{apr'}_{\mathbb{C}}(A) = \{x_1, x_2, x_5\}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5\}$$
$$\underline{apr''}_{\mathbb{C}}(A) = \{x_2, x_5\}$$
$$\overline{apr''}_{\mathbb{C}}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5, x_7\}$$

Generating suitable coverings for a particular application is an important but challenging task. To assist with this process, we associate an ordering to the set of all coverings.

Definition 1. *[9] Let $\mathbb{C}$ and $\mathbb{C}'$ be two coverings of U. $\mathbb{C} \ll \mathbb{C}'$ if and only if for all $L \in \mathbb{C}'$ there exists a set $S \subseteq \mathbb{C}$ such that $L = \cup_{K \in S} K$. When $\mathbb{C} \ll \mathbb{C}'$, we also say that $\mathbb{C}$ is finer than $\mathbb{C}'$.*

We have the following proposition [9]:

Proposition 1. *Let $\mathbb{C}$ and $\mathbb{C}'$ be coverings of U. Then $\underline{apr'}_{\mathbb{C}'}(A) \subseteq \underline{apr'}_{\mathbb{C}}(A) \subseteq A \subseteq \overline{apr'}_{\mathbb{C}}(A) \subseteq \overline{apr'}_{\mathbb{C}'}(A)$ for all $A \subseteq U$, if and only if $\mathbb{C} \ll \mathbb{C}'$.*

Example 2. *To illustrate the previous proposition we continue Example 1. Let $\mathbb{C}' = \{\{x_0, x_1, x_2\}, \{x_2, x_5\}, \{x_3, x_4, x_6, x_7, x_8, x_9\}\}$. Then we clearly have that $\mathbb{C} \ll \mathbb{C}'$. If we now recalculate the approximations of A, using $\mathbb{C}$, we get:*

$$\underline{apr'}_{\mathbb{C}'}(A) = \{x_2, x_5\}$$
$$\overline{apr'}_{\mathbb{C}'}(A) = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$$

In [17] a promising new method was introduced to generate coverings for classification problems, using the Mapper algorithm that originated in the field of topological data analysis (TDA). It generates coverings of a universe U, as shown in Algorithm 1, which generally represent the topological structure of the dataset and are therefore useful for our purposes. The algorithm uses a lens function $f : U \to \mathbb{R}^d$, an input covering $\mathbb{I}$ of $f(U)$, and a clustering algorithm. Often

the lens function is chosen to be a projection on particular attributes or PCA components. The input covering $\mathbb{I}$ is in practice constructed using a number of (hyper-)cubes among each dimension in $\mathbb{R}^d$, such that consecutive cubes have a certain overlap percentage p.

Mapper proceeds in the following way: the lens function transforms the original dataset U to the lower-dimensional space $f(U)$, on which the covering $\mathbb{I}$ is defined. Then, for each I inside the input covering, the instances from U projected onto it by f (i.e., $f^{-1}(I)$) are partitioned by means of a clustering algorithm, another parameter of the algorithm. These clusters form the resulting covering.

In [26] we proved the following result which holds when the clustering algorithm is equal to the single-linkage clustering algorithm [27] (shown in Algorithm 2).

Proposition 2. *Let U be a dataset, $f : U \to \mathbb{R}^n$. Let K denote the single linkage clustering algorithm. Let $\mathbb{I}$ and $\mathbb{I}'$ be coverings of $f(U)$, and let $\mathbb{C}_{\mathbb{I},\varepsilon}$ and $\mathbb{C}_{\mathbb{I}',\varepsilon'}$ denote the coverings generated by the Mapper algorithm using lens function f, the single linkage algorithm with parameters ε and ε' and coverings $\mathbb{I}$ and $\mathbb{I}'$. If $\mathbb{I} \ll \mathbb{I}'$ and $\varepsilon \leq \varepsilon'$, then $\mathbb{C}_{\mathbb{I},\varepsilon} \ll \mathbb{C}_{\mathbb{I}',\varepsilon'}$.*

Algorithm 1: The Mapper algorithm

Input: a dataset U with an associated metric, a *lens* function $f : U \to \mathbb{R}$ (or $\mathbb{R}^m$), a covering $\mathbb{I} = \{I_j \mid j = 1, \dots n\}$ of $f(U)$, and a clustering algorithm cl on U.

for $j \in \{1, 2, \dots, n\}$ **do**

  $A_j = \{A_{j1}, \dots, A_{jk} \mid i = 1, \dots k\} \leftarrow cl(f^{-1}(I_j))$;

end

Result: $\cup A_j$

Algorithm 2: The Single Linkage algorithm

Input: a dataset U, with an associated metric d and a real number $\varepsilon > 0$

Construct a graph G_ε, with U as vertices and an edge between x and y when $d(x, y) \leq \varepsilon$. ;

$\{A_1, \dots, A_k\} \leftarrow$ The connected components of G_ε;

Result: $\{A_1, \dots, A_k\}$

III. Entropy

In order to be able to control the fineness of the constructed coverings, we require a measure that reflects granularity. A suitable candidate for this is entropy, which is classically defined for probability distributions:

Definition 2. *[28] Let $\Pr$ be a probability distribution over a finite universe U. The entropy of $\Pr$ is defined as:*

$$H(\Pr) := -\sum_{x \in U} \Pr(x) \log \Pr(x)$$

This definition can be naturally adjusted to partitions:

Definition 3. *[19] Let $\pi = \{A_1, \dots, A_k\}$ be a partition of U. We define the entropy of π as:*

$$H(\pi) := -\sum_{i=1}^{k} \frac{|A_i|}{|U|} \log \frac{|A_i|}{|U|}.$$

The entropy indeed reflects the fineness of a partition. When every element resides in its own equivalence class, the entropy is maximal and equal to $\log |U|$. When there is only one equivalence class, the entropy is minimal and equal to 0.

We also have the following property:

Proposition 3. *[14] Let $\pi \ll \pi'$ be two partitions of U, then $H(\pi) \geq H(\pi')$.*

Since we work with coverings instead of partitions, Definition 3 is not directly applicable. Instead, we adopt the proposal from [14] that computes the entropy associated with an arbitrary pair of covering-based approximation operators.

Definition 4. *[14]*

- *Let $(\underline{apr}, \overline{apr})$ be a pair of covering-based approximation operators over a finite universe U. We define $\Pi_{\underline{apr}}^{\overline{apr}}$ as the set of equivalence classes of $\mathcal{P}(U)$ of the equivalence relation $A \approx B \Leftrightarrow (\underline{apr}(A), \overline{apr}(A)) = (\underline{apr}(B), \overline{apr}(B))$.*
- *We define the entropy of $(\underline{apr}, \overline{apr})$ as follows:*

$$\mathcal{H}_{\underline{apr}}^{\overline{apr}} = H(\Pi_{\underline{apr}}^{\overline{apr}})$$

Again we observe that a high entropy occurs when approximation operators are very fine and a low entropy occurs when approximation operators are very rough. When we have for all $A \subseteq U$ that $(\underline{apr}(A), \overline{apr}(A)) = (A, A)$ (the approximation operator is maximally fine), the size of each class in $\Pi_{\underline{apr}}^{\overline{apr}}$ is equal to 1 and thus the entropy reaches its the maximal value, $|U|$. When an approximation operator is maximally rough, i.e. $(\underline{apr}(A), \overline{apr}(A)) = (\emptyset, U)$ for all A, $\Pi_{\underline{apr}}^{\overline{apr}}$ is equal to $\{U\}$ and thus the entropy is minimal and equal to 0.

Example 3. *Let $U = \{x_1, \dots, x_n\}$ and $\mathbb{C} = \{\{x_i, x_{i+1}\} \mid 1 \leq i < n\}$.*

We calculate the entropy of the strong approximations by first constructing $\Pi_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}$.

Let $\approx$ be the equivalence relation corresponding to $\Pi_{\underline{apr'}_{\mathbb{C}}}^{\overline{apr'}_{\mathbb{C}}}$. Then $A \approx B$ if either $A = B$ or A and B are equal to $\{x_i \mid 1 \leq i \leq n$ and $i \equiv 0$ (mod 2)$\}$ and $\{x_i \mid 1 \leq i \leq n$ and $i \equiv 1$ (mod 2)$\}$.

To see this, first note that

$$\underline{apr'}_{\mathbb{C}}(A) = \{x_i \mid x_i \in A \text{ and } (x_{i-1} \in A \text{ or } x_{i+1} \in A)\}$$
$$\overline{apr'}_{\mathbb{C}}(A) = \{x_i \mid x_i \in A \text{ or } (x_{i-1} \in A \text{ and } x_{i+1} \in A)\}$$

Suppose that $A \neq B$ but $A \approx B$. Let i be an integer such that $x_i \in A$ and $x_i \notin B$. Since $x_i \in \overline{apr'}_{\mathbb{C}}(A)$, $x_{i+1} \in B$ and $x_{i-1} \in B$. Since $x_i \notin \underline{apr'}_{\mathbb{C}}(B)$, $x_{i-1} \notin A$ and $x_{i+1} \notin A$. We can repeat the same procedure for x_{i-1} and x_{i+1} and in the end we conclude that A and B have the requested shape.

Thus, $\Pi^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}}$ has one equivalence class of size two and all other classes have size one. We can now calculate the entropy:

$$\begin{aligned}\mathcal{H}^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}} &= \frac{-\log 2^{-n+1}}{2^{n-1}} - \sum_{i=1}^{2^n-2} \frac{1}{2^n} \log \frac{1}{2^n} \\ &= \frac{n-1}{2^{n-1}} + \frac{(2^n-2)n}{2^n} \\ &= \frac{2n-2+n2^n-2n}{2^n} \\ &= n - \frac{1}{2^{n-1}}\end{aligned}$$

We also prove that entropy increases monotonically with increasing fineness in coverings. This is useful because, as mentioned in Proposition 2, we are able to construct chains of coverings under $\ll$. Therefore, we may exploit this monotonicity to guide the search for coverings that possess a certain entropy.

Proposition 4. *Let $\mathbb{C} \ll \mathbb{C}'$ be coverings of U. If $(\underline{apr'}_{\mathbb{C}}(A), \overline{apr'}_{\mathbb{C}}(A)) = (\underline{apr'}_{\mathbb{C}}(B), \overline{apr'}_{\mathbb{C}}(B))$ then also $(\underline{apr'}_{\mathbb{C}'}(A), \overline{apr'}_{\mathbb{C}'}(A)) = (\underline{apr'}_{\mathbb{C}'}(B), \overline{apr'}_{\mathbb{C}'}(B))$.*

Proof. First assume $\underline{apr'}_{\mathbb{C}}(A) = \underline{apr'}_{\mathbb{C}}(B)$. We now have for every $K \in \mathbb{C}$ that $K \subseteq A$ if and only if $K \subseteq B$. Assume by contradiction that $\underline{apr'}_{\mathbb{C}'}(A) \neq \underline{apr'}_{\mathbb{C}'}(B)$ then (w.l.o.g.) there exists a $K \in \mathbb{C}'$ such that $K \subseteq A$ but $K \nsubseteq B$. Because $\mathbb{C} \ll \mathbb{C}'$ we have that $K = \cup_{L \in S} L$ for some $S \subseteq \mathbb{C}$. However, for all $L \in S$, it holds that $L \subseteq A$ and thus $L \subseteq B$. This is a contradiction. Thus, $\underline{apr'}_{\mathbb{C}'}(A) = \underline{apr'}_{\mathbb{C}'}(B)$.

Because of the duality of the strong approximation operator, also $\overline{apr'}_{\mathbb{C}'}(A) = \overline{apr'}_{\mathbb{C}'}(B)$.

□

This implies the following theorem:

Theorem 1. *Let $\mathbb{C} \ll \mathbb{C}'$ be coverings of a universe U, then $\mathcal{H}^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}} \geq \mathcal{H}^{\overline{apr'}_{\mathbb{C}'}}_{\underline{apr'}_{\mathbb{C}'}}$.*

Proof. This follows from Proposition 4 and Proposition 3. □

There does not exist a similar result establishing monotonicity between the weak granule-based approximations and entropy. Because of this, we only consider the strong approximations in the remainder of this paper.

IV. Estimating entropy

Since our definition of entropy takes into account all subsets of U, it is computationally expensive to calculate its exact value. Because of this, we will estimate the value of the entropy. It is well-known that there is no unbiased estimator for the entropy of a probability distribution when there is no external knowledge about the probability distribution [29]. There exist, however, some good estimators that do not depend on external knowledge. We will also use our knowledge about covering-based rough sets to estimate and determine the exact size of certain equivalence classes of $\Pi^{\overline{apr}}_{\underline{apr}}$, which will lead to an unbiased estimator anyway.

We proceed in the following way: Let Y equal the multiset $\{-\frac{|\mathcal{A}_i|}{2^{|U|}} \log \frac{|\mathcal{A}_i|}{2^{|U|}} \mid \mathcal{A}_i \in \Pi^{\overline{apr}}_{\underline{apr}}\}$. Estimating the entropy is then equivalent to estimating the sum of Y.

Inspired by [19] and [30], we do this using the Horvitz–Thompson estimator [31]. This estimator is unbiased and counts the sum of all elements y_i in a universe Y by sampling a set S from Y without replacement. The estimator is equal to

$$\sum_{y_i \in S} \frac{y_i}{\Pr(y_i \in S)},$$

where $\Pr(y_i \in S)$ is the probability of a sample y_i belonging to S. However, this estimator is only unbiased if we know the size of each $\mathcal{A}_i$ exactly.

We proceed as follows: first, we generate n random subsets $\{B_1, \ldots, B_n\}$ of U. Let M be the set of approximations $\{(\underline{apr'}_{\mathbb{C}}(B_i), \overline{apr'}_{\mathbb{C}}(B_i)) \mid 1 \leq i \leq n\}$. We observe that the probability of B_i belonging to a certain equivalence class $\mathcal{A}_l$ of $\Pi^{\overline{apr}}_{\underline{apr}}$ is equal to the proportion of the size of the equivalence class to the number of possible subsets: $\frac{|\mathcal{A}_l|}{2^{|U|}}$.

We define $p_k = \frac{|\mathcal{A}_k|}{2^{|U|}}$ for every $1 \leq k \leq |M|$ (note that $|M| \leq n$), where $\mathcal{A}_k = \{B \subseteq U \mid (\underline{apr'}_{\mathbb{C}}(B), \overline{apr'}_{\mathbb{C}}(B)) = (\underline{A}_k, \overline{A}_k)\}$, and $(\underline{A}_k, \overline{A}_k)$ is the k-th element of M. The probability of not generating a set from $\mathcal{A}_k$ is equal to $(1-p_k)^n$ and thus the probability of generating a set inside $\mathcal{A}_k$ is equal to $1-(1-p_k)^n$. Therefore, the result from the Horvitz-Thomson estimator is:

$$\hat{\mathcal{H}}^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}} = -\sum_k \frac{p_k \log p_k}{1-(1-p_k)^n} \tag{5}$$

We also have an estimator for the variance of the Horvitz-Thomson estimator [31]:

$$\begin{aligned}&\hat{V}\left(\hat{\mathcal{H}}^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}}\right) \\ &= \left(\hat{\mathcal{H}}^{\overline{apr'}_{\mathbb{C}}}_{\underline{apr'}_{\mathbb{C}}}\right)^2 - \sum_k \frac{(p_k \log p_k)^2}{1-(1-p_k)^n} \\ &\quad - \sum_{k \neq j} \frac{(p_k \log p_k)(p_j \log p_j)}{1-(1-p_k)^n-(1-p_j)^n+(1-p_k-p_j)^n}\end{aligned} \tag{6}$$

When we have exact values for p_k available, both of these estimators are unbiased [31] and can thus be used to estimate the entropy and the corresponding confidence interval. A new problem presents itself: determining p_k. We can do this by empirically estimating p_k (Section IV-A) or we can use properties of the covering itself to determine it exactly (Section IV-B).

A. Coverage-adjusted estimator

The coverage-adjusted estimator was introduced in [19] and uses the Horvitz-Thomson estimator to estimate the entropy of a probability distribution. According to the observations of the authors of [19], the most intuitive way to estimate p_k empirically is by letting $\hat{p}_k$ equal the number of generated sets B_i that have $(\underline{A}_k, \overline{A}_k)$ as their approximations divided by n. However, they introduced a more useful approximation, where

the estimator converges faster to the true value of the entropy. Here $\hat{C} = 1 - \frac{f_1}{n}$, where f_1 equals the number of approximations $(\underline{A}_k, \overline{A}_k)$ such that there is only one B_i which has $(\underline{A}_k, \overline{A}_k)$ as its approximation. We will call approximations which are reached more than once collisions. We now define

$$\hat{p}_k = \frac{\hat{C}}{n} |\{B_i \mid (\underline{apr'}_{\mathbb{C}}(B_i), \overline{apr'}_{\mathbb{C}}(B_i)) = (\underline{A}_k, \overline{A}_k)\}| \quad (7)$$

This estimation of entropy based on Equation (5) and Equation (7) is called the coverage-adjusted estimator. It converges to the true value of entropy as follows. Let $\hat{\mathcal{H}}$ be the coverage-adjusted entropy estimate of $\mathcal{H}$ generated using n samples. We now have that $|\hat{\mathcal{H}} - \mathcal{H}| = O(1/\log(n))$ [19]

The main downside of this estimator is that we need to choose a large enough n to ensure that there is an approximation $(\underline{A}_k, \overline{A}_k)$ that is reached at least twice. When there are no such sets, $\hat{C} = 0$ and thus all $p_k = 0$ which will lead to an entropy of 0 (convention states that $0 \log 0 = 0$). When the entropy is high this often means that n needs to be unfeasibly large. When the entropy is equal to e, we expect this to happen on average after generating $O(2^{e/2})$ sets [32]. To mitigate this problem, the authors of [19] replaced the definition of $\hat{C}$ with $1 - \frac{f_1}{n+1}$. Now $\hat{C}$ is never equal to 0 and still has the same asymptotic behavior. However, when the true value of the entropy is too high, this can still cause problems. When no collisions occur the coverage-adjusted estimator is equal to:

$$\begin{aligned}
\hat{\mathcal{H}} &= -\sum_{i=1}^{n} \frac{\hat{C}/n \log(\hat{C}/n)}{1 - (1 - \hat{C}/n)^n} \\
&= \sum_{i=1}^{n} \frac{\log(n(n+1))}{n(n+1)(1 - (1 - \frac{1}{n(n+1)})^n)} \\
&= \frac{\log(n(n+1))}{(n+1)(1 - (\frac{n(n+1)-1}{n(n+1)})^n)} \\
&= O(\log(n))
\end{aligned}$$

We would thus need exponentially many samples to estimate an entropy accurately when no collisions occur. This can be intractable when the entropy is large.

B. Horvitz-Thompson Exact Estimator

In this section we introduce a different way of estimating the entropy by determining the exact size of $\mathcal{A}_k$ so that we can use Equation (5) to get an unbiased estimator.

To do this, we introduce the following definition: $P_{\mathbb{C}}$ is the partition of U given by the equivalence relation $\sim$, defined as follows:

$$(\forall x, y \in U)(x \sim y \Leftrightarrow \{K \in \mathbb{C} \mid x \in K\} = \{K \in \mathbb{C} \mid y \in K\})$$

In other words, points that are equivalent always occur together in the covering. Also note that every element K of $\mathbb{C}$ can be written as the union of elements in $P_{\mathbb{C}}$.

We will call elements of $P_{\mathbb{C}}$ blocks, and we define a tripartition of $P_{\mathbb{C}}$ for every $A \subseteq U$:

$$\begin{aligned}
P_1(A) &:= \{X \in P_{\mathbb{C}} \mid X \subseteq A\} \\
P_2(A) &:= \{X \in P_{\mathbb{C}} \mid \emptyset \neq X \cap A \neq X\} \\
P_3(A) &:= \{X \in P_{\mathbb{C}} \mid X \cap A = \emptyset\}
\end{aligned}$$

The strong approximations are fully characterized by P_1, P_2 and P_3 because:

$$\begin{aligned}
\underline{apr'}_{\mathbb{C}}(A) &= \cup \{K \in \mathbb{C} \mid (\forall X \in P_{\mathbb{C}})(X \subseteq K \Rightarrow X \in P_1(A))\} \\
\overline{apr'}_{\mathbb{C}}(A) &= \cup \{X \in P_{\mathbb{C}} \mid (\forall K \in \mathbb{C})(X \cap K \neq \emptyset \\
&\qquad \Rightarrow (\exists X' \in P_1(A) \cup P_2(A))(X' \subseteq K))\}
\end{aligned}$$

For a given tripartition (P_1, P_2, P_3) of $P_{\mathbb{C}}$ the number of sets $A \subseteq U$ that have $P_i(A) = P_i$ (for $i \in \{1, 2, 3\}$) is equal to:

$$\prod_{X \in P_2} \left(2^{|X|} - 2\right)$$

Because of this we can count all $A \in \mathcal{A}_k$ by generating all possible assignments (P_1, P_2, P_3) that have the correct lower and upper approximations.

For each $A \in \mathcal{A}_k$, every block $X \in P_{\mathbb{C}}$ is in some $P_i(A)$; we call i the type of X with respect to A. We will count all sets $A \in \mathcal{A}_k$, by deciding for each block $X \in P_{\mathbb{C}}$ which type it has. This needs to be done in such a way that every $Y \in \mathbb{C}$ is 'satisfied', meaning:

- If $Y \subseteq \underline{A}_k$, all blocks $X \subseteq Y$ must be in $P_1(A)$
- If $Y \cap \overline{A}_k = \emptyset$, all blocks $X \subseteq Y$ must be in $P_3(A)$
- Else, at least one block $X \subseteq Y$ should be in $P_1(A) \cup P_2(A)$ and at least one block $X \subseteq Y$ should be in $P_2(A) \cup P_3(A)$.

It is clear that every A for which these conditions hold has $(\underline{A}_k, \overline{A}_k)$ as its approximations.

Based on the first two cases above, for some blocks X we can immediately establish their type. For others (blocks that do not appear in the first two cases) we need to perform a backtracking search. Note that for every unsatisfied $Y \in \mathbb{C}$, $Y \cap \overline{A}_k \neq \emptyset$ and $Y \not\subseteq \underline{A}_k$.

The process proceeds as follows: as long as there are unsatisfied $Y \in \mathbb{C}$, we recursively pick one such Y arbitrarily and decide how it can be satisfied. For this, we arbitrarily order the blocks $X \subseteq Y$ which are still undecided. Then we vary the type of the first X among all types that are consistent with the requested lower and upper approximation. We repeat this process until Y is satisfied. Notice that this implies we can leave some $X \subseteq Y$ undecided. We also take into consideration that a block with only one element can never be in $P_2(A)$. Once all $Y \in \mathbb{C}$ are satisfied, we count the number of sets $A \subseteq U$ for which the type of each $X \in P_{\mathbb{C}}$ coincides with the assignment and add this to a running counter. That is:

$$\Pi_{X \in P_2}(2^{|X|} - 2) \cdot \Pi_{X \in N}(2^{|X|})$$

where N denotes the set of blocks for which no decision was made. By using recursion, we exhaustively generate all

possible assignments that satisfy every Y in $\mathbb{C}$ and thus we count all sets A with the right approximations.

The pseudo-code of this approach can be found in Algorithm 3. Below, we detail some optimizations that can be performed:

- When deciding the type of a block X for a $Y \in \mathbb{C}$ that already has some elements in P_1, it is enough to decide whether X is in P_1 or in $P_2 \cup P_3$, instead of deciding if it is in P_1, P_2 or P_3. Therefore the branching factor of the recursion is reduced from 3 to 2 in such cases.
- Instead of picking an arbitrary unsatisfied $Y \in \mathbb{C}$, we can choose a more specific Y in order to speed up the algorithm. A common heuristic is to make the next decision about the object with the fewest possibilities. In this case, we choose the Y with the least amount of undecided blocks left.
- The partition of U into blocks can be redefined dynamically during the runtime of the algorithm. Particularly if a $Y \in \mathbb{C}$ gets satisfied but has some undecided blocks left, it is no longer relevant that they are in Y, and therefore they could possibly be merged with other blocks. We chose not to implement this in general because it requires a lot of 'bookkeeping', but we used a limited variant. After the initial assignment of types to blocks that can be made without guessing, we make the partition coarser just once.

We will call the estimator of entropy that uses this approach to determine $|\mathcal{A}_k|$ the Horvitz-Thompson-exact estimator (HTEE).

Example 4. *Let U be a universe and $\mathbb{C}$ be the covering with three elements, shown in Figure 1, such that $P_{\mathbb{C}}$ has 5 blocks: $X_1, \ldots, X_5$ and $\mathbb{C} = \{X_1 \cup X_2, X_2 \cup X_3 \cup X_4, X_4 \cup X_5\}$.*

We want to calculate the number of sets $A \subseteq U$ so that $\underline{apr'}_{\mathbb{C}}(A) = X_4 \cup X_5$ and $\overline{apr'}_{\mathbb{C}}(A) = U$. We do this by exhaustively generating all possible assignments[1] of P_1, P_2, P_3 and N, using Algorithm 3 . These are shown in Table I. For example, for the first assignment, we see that even though the type of X_3 is undecided, all elements in $\mathbb{C}$ are satisfied. The amount of sets that are realizations of assignment 5 are:

$$\texttt{count}(\textit{assignment 5}) = \left(2^{|X_1|} - 2\right)\left(2^{|X_2|} - 2\right)2^{|X_3|}$$

If we add up the amount of realizations of all possible assignments we get the following result:

$$2^{|X_3|}\left(2^{(|X_1|+|X_2|)} - 2\right) - 2^{|X_1|} + 1$$

If we set the size of all blocks equal to 10 we have that the probability of a random subset of U having $(X_4 \cup X_5, U)$ as approximations is equal to:

$$\frac{2^{|X_3|}\left(2^{(|X_1|+|X_2|)} - 2\right) + 1}{2^{50}} = \frac{1073738753}{1125899906842624} \approx 9.537 \times 10^{-7}$$

[1]We assume that all blocks have at least two elements such that the assignment of type P_2 makes sense for each block.

Algorithm 3: Exact approximation counting

Input: covering $\mathbb{C}$; approximation $(\underline{A}, \overline{A})$
`// Preprocessing`
Check which $Y \in \mathbb{C}$ lie in $\underline{A}$, and which are disjoint with $\overline{A}$
Assign their elements the required type
$\mathbb{C}' \leftarrow \{Y \in \mathbb{C} \mid Y \text{ is still unsatisfied}\}$
Identify the 'blocks' with respect to $\mathbb{C}'$
$total \leftarrow 0$
`assign()`
Result: $total$ contains the number of sets with the given approximation

Function `assign()`**:**
  if *all $Y \in \mathbb{C}'$ are satisfied* **then**
    $total \leftarrow total + \Pi_{X \in P_2}(2^{|X|} - 2) \cdot \Pi_{X \in N}(2^{|X|})$
  else
    $Y \leftarrow$ unsatisfied $Y \in \mathbb{C}'$ with fewest amount of unassigned blocks
    `satisfy` (Y)
  end

Function `satisfy(`Y`)`**:**
  if Y *is satisfied* **then**
    `assign()`
  else
    $X \leftarrow$ first unassigned block in Y
    for $j \in \{1, 2, 3\}$ **do**
      $type_X \leftarrow P_j$
      `satisfy` (Y)
    end
    $type_X \leftarrow N$
  end

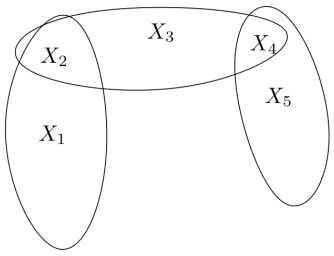


Fig. 1. $\mathbb{C}$ and its blocks from Example 4

Example 5. *We take Example 3 and inspect how the HTEE estimator will evaluate the entropy of $\mathbb{C}$. If we generate k sets of the universe U, we can safely assume that no collisions occur and no set is in the class of size two if k is small enough, since it is statistically very unlikely. Using the method described above, we can calculate the exact size of the equivalence class of each approximation that is generated*

TABLE I
ALL ASSIGNMENTS OF $P_{\mathbb{C}}$ GENERATED BY ALGORITHM 3 WHERE $\underline{apr'}_{\mathbb{C}}(A) = X_4 \cup X_5$ AND $\overline{apr'}_{\mathbb{C}}(A) = U$.

Assignments	1	2	3	4	5	6	7	8	9
P_1	X_1, X_4, X_5	X_1, X_4, X_5	X_2, X_4, X_5	X_2, X_4, X_5	X_4, X_5	X_4, X_5	X_2, X_4, X_5	X_2, X_4, X_5	X_4, X_5
P_2	X_2	$\emptyset$	X_1, X_3	X_1	X_1, X_2	X_1	X_3	$\emptyset$	X_2
P_3	$\emptyset$	X_2	$\emptyset$	X_3	$\emptyset$	X_2	X_1	X_1, X_3	X_1
N	X_3	X_3	$\emptyset$	$\emptyset$	X_3	X_3	$\emptyset$	$\emptyset$	X_3

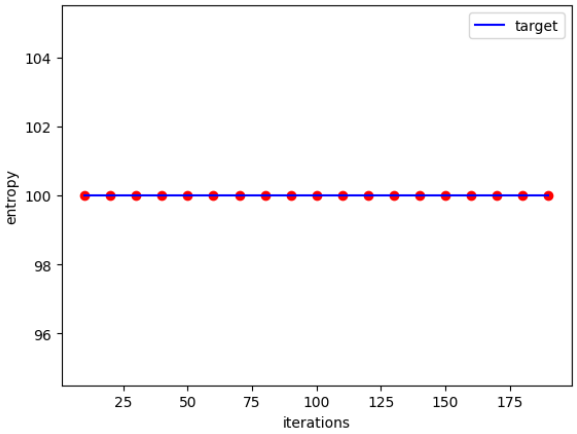


Fig. 2. The result of HTEE for the covering $\mathbb{C}$ with $n = 100$

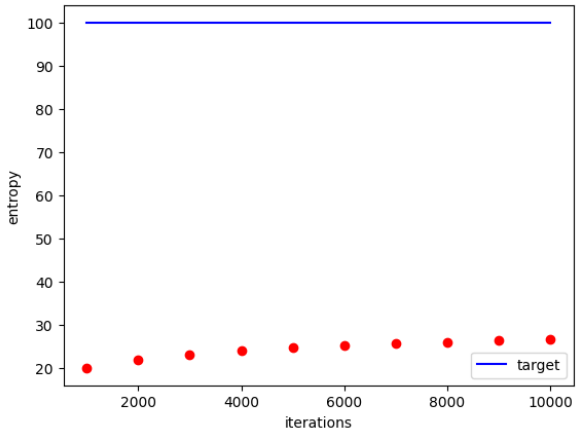


Fig. 3. The result of the coverage-adjusted estimator for the covering $\mathbb{C}$ with $n = 100$

and we get the following estimate:

$$\begin{aligned}\hat{\mathcal{H}} &= -\sum_{i=1}^{k} \frac{\log(2^{-n})}{2^n(1-(1-2^{-n})^k)} \\ &= \frac{kn}{2^n(1-(1-2^{-n})^k)} \\ &\approx \frac{kn}{2^n k 2^{-n}} \\ &= n\end{aligned}$$

The third line follows from:

$$\left(1-\frac{1}{2^n}\right)^k = \sum_{i=0}^{k}\binom{k}{i}(-2)^{ni} \approx 1-\frac{k}{2^n},$$

which holds when k is significantly smaller than 2^n.

V. EXPERIMENTS

In this section we will compare the previously discussed estimations.

We first evaluate the entropy for $n = 100$ for the covering from Example 3. When we use HTEE we see that, even after generating only 2 sets we already get an estimate that is less than 10^{-30} removed from $100 - \frac{1}{2^{-99}}$, as can be seen in Figure 2. The variance estimated by Equation (6) is practically zero. This confirms our findings from Example 5.

We may also observe that the coverage-adjusted estimator runs into problems because the value of the entropy is too large. Because of this there will be no collisions, which will cause $\hat{C}$ to be too far off. For completeness we provide a plot of the results of the coverage-adjusted estimator in Figure 3.

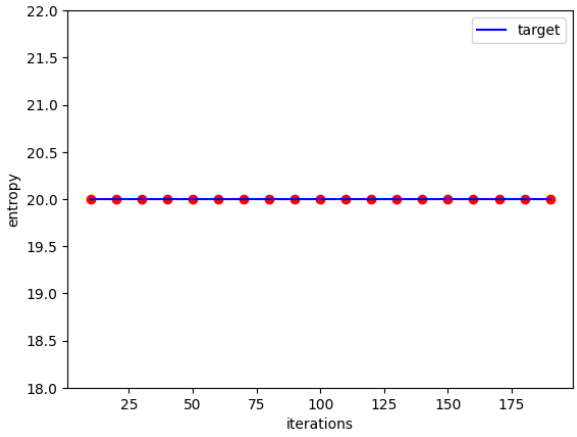


Fig. 4. The result of HTEE for the covering $\mathbb{C}$ with $n = 20$

When n is reduced to 20, both estimators produce accurate results, as seen in Figures 4 and 5.

Next, we consider real-world datasets from the UCI repository [33] and use coverings generated by Mapper. We use the following parameters for Mapper: the number of cubes is equal to 4, the overlap percentage is equal to 0.2, we use the projection onto the first PCA component as the lens function, and as a clustering algorithm we use the single-linkage algorithm with distance cutoff ε. The parameter ε will depend on the dataset we use since the average distance between points varies between datasets. We use the datasets with corresponding ε as displayed in Table II. The size of the

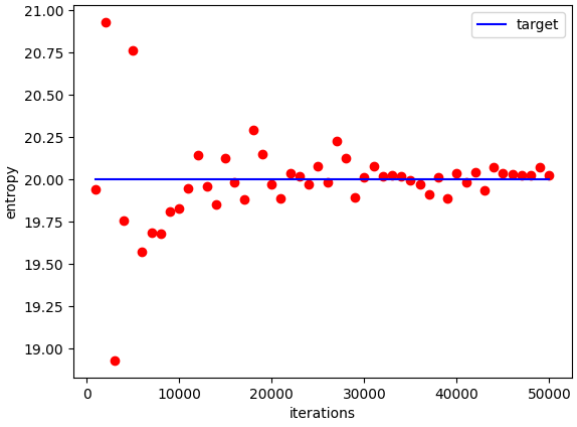


Fig. 5. The result of the coverage-adjusted estimator for the covering $\mathbb{C}$ with $n = 20$

TABLE II
THE DATASETS USED IN SECTION V

	Attributes	Instances	ε	$\lvert\mathbb{C}\rvert$
Iris	5	150	0.6	25
Vertebral Column	6	310	20	43
Wine	14	178	17.5	64

resulting coverings can also be found in this table.

The results obtained from the Iris dataset are presented in Figures 6 and 7. These figures show that the HTEE estimator converges rapidly after a small number of iterations. Additionally, the coverage-adjusted estimator consistently yields values lower than the actual entropy, which can be reasonably approximated as 15 based on the variance estimate of the HTEE estimator.

For the Vertebral Column and Wine datasets the entropy is again too high, meaning that we can only consider the HTEE estimator.

We see promising results where the variance of the estimator

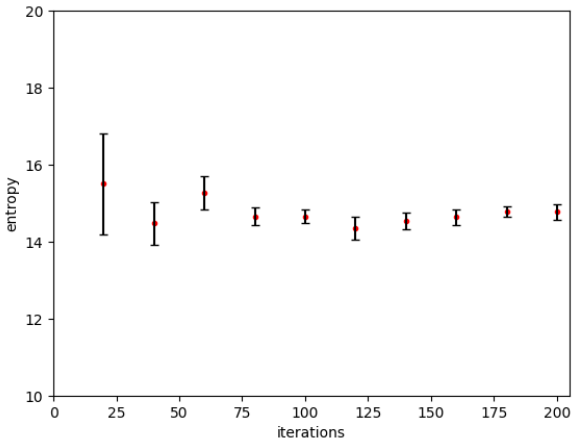


Fig. 6. The result of HTEE for the Iris dataset with $\varepsilon = 0.6$.

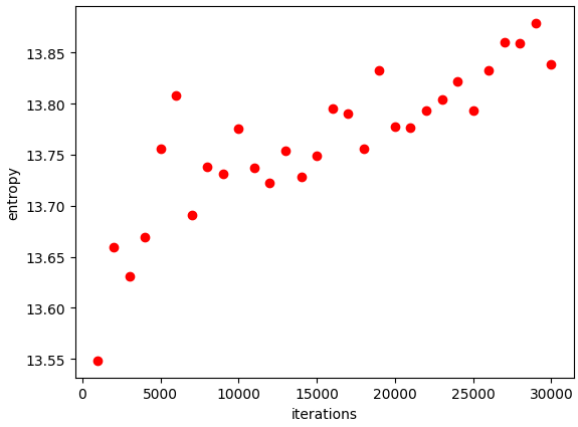


Fig. 7. The result of the coverage-adjusted estimator for the Iris dataset with $\varepsilon = 0.6$.

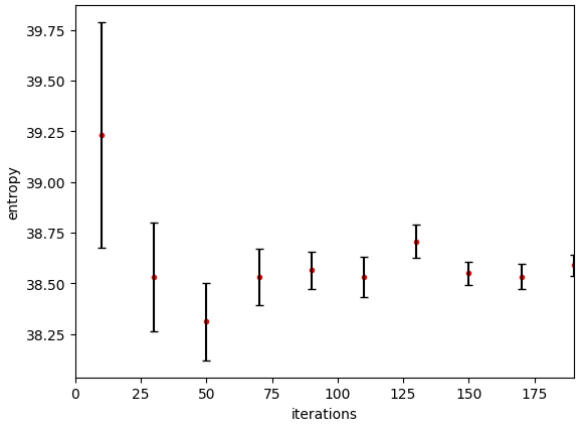


Fig. 8. The result of HTEE for the Vertebral Column dataset

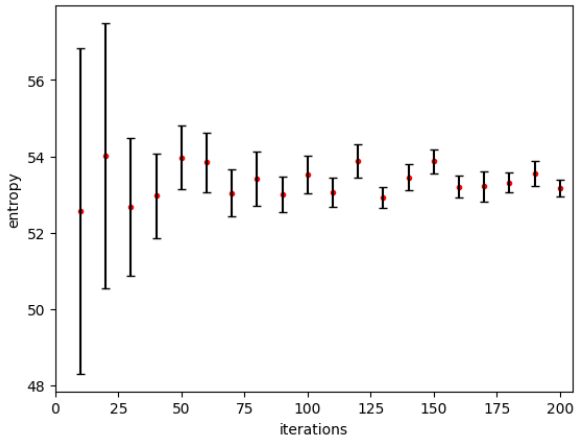


Fig. 9. The result of HTEE for the Wine dataset

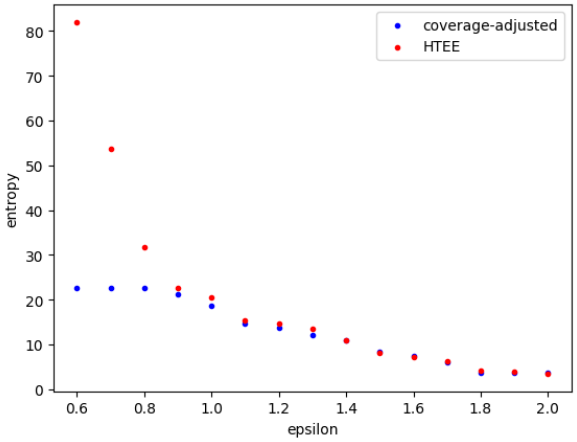


Fig. 10. The values of the HTEE and coverage-adjusted estimators for the iris dataset with a variable ε value.

quickly approaches zero as the number of iterations increases as shown in Figures 8 and 9.

We further analyze the Iris dataset by varying the value of ε, which controls the scale parameter in the Mapper algorithm. Adjusting ε generates a sequence of coverings under $\ll$, as mentioned in proposition 2. We vary ε from 0.6 to 2.0 in increments of 0.1. For the HTEE estimator, we perform 100 iterations, while for the coverage-adjusted estimator, we use 20,000 iterations—chosen to ensure comparable execution times for both methods. The variance of the HTEE estimator remains below 0.30 across all experiments. Based on this setup, Figure 10 presents the resulting comparisons. As shown, both estimators yield similar results when ε is large and the entropy is low. However, as entropy increases, the coverage-adjusted estimator consistently underestimates the true entropy. In cases of very high entropy, its estimates deviate significantly from the expected values.

VI. Discussion and Future Work

In this paper, we proposed an efficient method for estimating the entropy of the strong granule-based approximation operator in covering-based rough sets. Our approach involves generating random subsets of the universe U, determining the number of subsets that share the same rough set approximations, and applying the Horvitz-Thompson estimator to obtain an unbiased entropy estimate. We evaluated the performance of our estimator against the coverage-adjusted entropy estimator from the literature and found that our method not only yields greater accuracy but also performs more effectively in scenarios involving high-entropy coverings.

In future work, we plan to investigate estimates to calculate the size of $\mathcal{A}_k$ using properties of the covering and heuristics. A possible start to do this can be found in the following equation:

$$\Pi_{K\in P_{\mathbb{C}}, K\subseteq \overline{A}_k, K\not\subseteq \underline{A}_k}(2^{|K|}-2) \leq |\mathcal{A}_k| \leq 2^{|\overline{A}_k|-|\underline{A}_k|}$$

which can be proved by observing that

$$P_1 = \{K \in P_{\mathbb{C}} \mid K \subseteq \underline{A}_k\}$$
$$P_2 = \{K \in P_{\mathbb{C}} \mid K \subseteq \overline{A}_k, K \not\subseteq \underline{A}_k\}$$
$$P_3 = \{K \in P_{\mathbb{C}} \mid K \not\subseteq \overline{A}_k\}$$

may be used to assign types to all blocks.

When $\mathbb{C}$ is a partition, $\Pi_{K\in P_{\mathbb{C}}, K\subseteq \overline{A}_k, K\not\subseteq \underline{A}_k}(2^{|K|}-2)$ is exactly equal to $|\mathcal{A}_k|$. For general coverings, this does not hold any longer but we can use this as our approximation for $|\mathcal{A}_k|$.

Whenever $|K|$ is large enough for all K in $P_{\mathbb{C}}$, the difference between $2^{|K|}$ and $2^{|K|}-2$ is negligible. However in practice, most blocks in the used coverings are not that large. Because of this, we did not include this approach in the experimental analysis Section V. We still mention this estimate as it can probably be improved by using more heuristics to an even better estimate of $|\mathcal{A}_k|$.

Further research should also focus on constructing methods to estimate the entropy of (covering-based) rough set approximations other than the strong approximations. Finally, researchers and practitioners may also use these results to construct coverings with a given entropy. This is in particular useful for applications that require a certain degree of granularity.

Acknowledgments

This work was partially supported by Universidad Militar Nueva Granada, under project CIAS 3144-2020.

References

[1] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982. doi: https://doi.org/10.1007/BF01001956

[2] Y. Yao, "On generalizing rough set theory," in *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 44–51.

[3] Z. Shi and Z. Gong, "The further investigation of covering-based rough sets: Uncertainty characterization, similarity measure and generalized models," *Information Sciences*, vol. 180, no. 19, pp. 3745–3763, 2010. doi: https://doi.org/10.1016/j.ins.2010.06.020

[4] T. Yang and Q. Li, "Reduction about approximation spaces of covering generalized rough sets," *International Journal of Approximate Reasoning*, vol. 51, no. 3, pp. 335–345, 2010. doi: https://doi.org/10.1016/j.ijar.2009.11.001

[5] I. Couso and D. Dubois, "Rough sets, coverings and incomplete information," *Fundam. Inform.*, vol. 108, pp. 223–247, 01 2011. doi: 10.3233/FI-2011-421

[6] J. Järvinen and S. Radeleczki, "Rough sets determined by tolerances," *International Journal of Approximate Reasoning*, vol. 55, no. 6, pp. 1419–1438, 2014. doi: https://doi.org/10.1016/j.ijar.2013.12.005

[7] D. Chen, X. Zhang, and W. Li, "On measurements of covering rough sets based on granules and evidence

theory," *Information Sciences*, vol. 317, pp. 329–348, 2015. doi: https://doi.org/10.1016/j.ins.2015.04.051

[8] L. D'eer and C. Cornelis, "Notes on covering-based rough sets from topological point of view: Relationships with general framework of dual approximation operators," *International Journal of Approximate Reasoning*, vol. 88, pp. 295–305, 2017. doi: https://doi.org/10.1016/j.ijar.2017.06.006

[9] M. Restrepo, C. Cornelis, and J. Gómez, "Partial order relation for approximation operators in covering based rough sets," *Information Sciences*, vol. 284, pp. 44–59, 2014. doi: https://doi.org/10.1016/j.ins.2014.06.032

[10] ——, "Duality, conjugacy and adjointness of approximation operators in covering-based rough sets," *International Journal of Approximate Reasoning*, vol. 55, no. 1, Part 4, pp. 469–485, 2014. doi: https://doi.org/10.1016/j.ijar.2013.08.002 Rough Sets and Logic.

[11] C. Wang, M. Shao, B. Sun, and Q. Hu, "An improved attribute reduction scheme with covering based rough sets," *Applied Soft Computing*, vol. 26, pp. 235–243, 2015. doi: https://doi.org/10.1016/j.asoc.2014.10.006

[12] L. Fachao, R. Yexing, and J. Chenxia, "Attribute reduction method of covering rough set based on dependence degree," *International Journal of Computational Intelligence Systems*, vol. 14, pp. 1419–1425, 2021. doi: 10.2991/ijcis.d.210419.002

[13] Y. Yang, D. Chen, X. Zhang, and Z. Ji, "Covering rough set-based incremental feature selection for mixed decision system," *Soft Comput.*, vol. 26, no. 6, pp. 2651–2669, 2022. doi: https://doi.org/10.1007/s00500-021-06687-0

[14] P. Zhu and Q. Wen, "Entropy and co-entropy of a covering approximation space," *International Journal of Approximate Reasoning*, vol. 53, no. 4, pp. 528–540, 2012. doi: https://doi.org/10.1016/j.ijar.2011.12.004

[15] Z. Li, S. Wei, and S. Liu, "Outlier detection using conditional information entropy and rough set theory," *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 1, pp. 1899–1918, 2024. doi: 10.3233/JIFS-236009

[16] L. Zhou and F. Jiang, "A rough set approach to feature selection based on relative decision entropy," in *Rough Sets and Knowledge Technology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: doi.org/10.1007/978-3-642-24425-4_17 pp. 110–119.

[17] M. Restrepo and C. Cornelis, "Mapper-based rough sets," in *Rough Sets: International Joint Conference, IJCRS 2024, Halifax, NS, Canada, May 17–20, 2024, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2024. doi: https://doi.org/10.1007/978-3-031-65665-1_1 p. 3–17.

[18] A. D. Smith, P. Dłotko, and V. M. Zavala, "Topological data analysis: Concepts, computation, and applications in chemical engineering," *Computers & Chemical Engineering*, vol. 146, 2021. doi: https://doi.org/10.1016/j.compchemeng.2020.107202

[19] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage-adjusted entropy estimation," *Statistics in Medicine*, vol. 26, no. 21, pp. 4039–4060, 2007. doi: https://doi.org/10.1002/sim.2942

[20] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski, *Rough Set Algorithms in Classification Problem*. Heidelberg: Physica-Verlag HD, 2000, pp. 49–88.

[21] W. Ziarko and N. Shan, "Discovering attribute relationships, dependencies and rules by using rough sets," in *Proceedings of the Twenty-Eighth Annual Hawaii International Conference on System Sciences*, vol. 3, 1995. doi: 10.1109/HICSS.1995.375608 pp. 293–299 vol.3.

[22] H. Alqaheri, A. E. Hassanien, and A. Abraham, "Discovering stock price prediction rules using rough sets," *Neural Network World*, vol. 18, 01 2008.

[23] M. S. Raza and U. Qamar, *Understanding and using rough set based feature selection: Concepts, techniques and applications*, 2nd ed. Singapore, Singapore: Springer, Sep. 2019.

[24] K. Thangavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: A review," *Applied Soft Computing*, vol. 9, no. 1, pp. 1–12, 2009. doi: https://doi.org/10.1016/j.asoc.2008.05.006

[25] A. K. Das, S. Sengupta, and S. Bhattacharyya, "A group incremental feature selection for classification using rough set theory based genetic algorithm," *Applied Soft Computing*, vol. 65, pp. 400–411, 2018. doi: https://doi.org/10.1016/j.asoc.2018.01.040

[26] W. De Maeyer, C. Cornelis, and M. Restrepo, "The granular structure of covering-based rough sets generated by mapper," in *Proceedings of 16th European Symposium on Computational Intelligence and Mathematics (ESCIM 2025)*, in press.

[27] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed., ser. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, 2011.

[28] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[29] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003. doi: https://doi.org/10.1162/089976603321780272

[30] A. Chao and T.-J. Shen, "Nonparametric estimation of shannon's diversity index when there are unseen species in sample. environ ecol stat 10: 429-443," *Environmental and Ecological Statistics*, vol. 10, pp. 429–443, 2003. doi: https://doi.org/10.1023/A:1026096204727

[31] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952. doi: https://doi.org/10.2307/2280784

[32] M. Skorski, "Improved estimation of collision entropy in high and low-entropy regimes and applications to

anomaly detection," *IACR Cryptol. ePrint Arch.*, vol. 2016, p. 1035, 2016.

[33] M. Kelly, R. Longjohn, and K. Nottingham, the UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 207–217
DOI: 10.15439/2025F6392
ISSN 2300-5963 ACSIS, Vol. 43

Simultaneous pursuit of accountability for regulatory compliance, financial benefits, and societal impacts in artificial intelligence (AI) projects

Gloria J. Miller, DBA
0000-0003-2603-0980
maxmetrics, Heidelberg
Email: g.j.m@ieee.org

***Abstract*—This exploratory study, grounded in agency theory, employs quantitative analyses to investigate the simultaneous pursuit of accountability for regulatory compliance, financial benefits, and societal impacts within artificial intelligence (AI) projects. An agent-principal matrix was developed, synthesizing knowledge from the AI stakeholder model into 11 accountability indicators. These indicators establish a standard of responsibility among project actors for regulatory compliance, ethical practices, and financial benefits. Using quantitative methods, we analyzed survey data on accountability and defined the scope of AI systems under development. We identified two clusters of AI systems—autonomous and non-autonomous—based on seven features. We then examined how these two types of systems, as well as the importance of sustainability and fairness, impact the promotion of accountability. Results indicate that accountabilities shift based on the scope of the AI system and the project role. Regulatory compliance, financial benefits, and societal impacts are not mutually exclusive project goals and coexist. The findings quantify subjective and theoretical speculation about accountabilities within AI projects. Additionally, the study contributes empirical data to the literature on AI, ethics, and project management.**

***Index Terms*—Accountability, Artificial intelligence, AI, Project management, Ethics, Agency Theory**

I. Introduction

THE launch of ChatGPT in 2023 dramatically increased interest, enthusiasm, and concerns regarding artificial intelligence (AI); ChatGPT is a generative AI chatbot based on a large language model that produces conversational outputs. Recognizing the potential impact of AI systems on the public and the environment, the European Commission and numerous countries have proposed frameworks, laws, regulations, and rules to responsibly manage these impacts, emphasizing interdependent values, principles, and actions [1, 2, 3]. At the same time, organizations prioritize achieving the target financial and business benefits from their AI investments [4, 5, 6].

This exploratory study investigated the relationship between accountability and the simultaneous pursuit of regulatory compliance, financial benefits, and societal impacts in AI projects. An AI project refers to a temporary organization that develops AI systems before deploying them for productive use or putting them into the market [7]. In the public's opinion, accountability is, on average, perceived as one of the most important ethical principles before fairness, security, privacy, and accuracy [8]. The present study addresses the following research question: *Who within an AI project is accountable for regulatory compliance, financial benefits, and societal impacts?*

The study employed a questionnaire-based survey and quantitative analysis to measure AI project teams' expectations and perceptions regarding accountability. Its theoretical framework is grounded in agency theory and builds upon an existing AI stakeholder accountability model [7] that maps project success factors and AI ethical principles. These factors establish an accountability standard governing the relationship between project actors, regulations, societal impacts, and financial benefits. A quantitative analysis of survey data then identifies what individuals involved in AI projects are held accountable for.

This paper begins with a literature review, followed by a description of the study's theoretical background, hypothesis development, and research methodology, including data collection and analysis. The findings are then presented and discussed, alongside the study's contributions, implications, limitations, and conclusions.

II. Literature Review

A. AI Projects

AI refers to a wide range of efforts to replicate complex human capabilities, such as language use, vision, and autonomous action, using computational models. This encompasses various system categories, including chatbots, humanoid robots, behavioral software, and predictive systems [7]. AI systems function as artificial agents, capable of making decisions and acting with little or no human involvement. The complex data and models in AI systems are often referred to as "black-boxes" because the end-users, and sometimes the developers, are unable to explain or interpret the model decisions [9]. Consequently, concerns have been raised regarding civil and criminal liabilities, data protection, security,

trustworthiness, and transparency related to moral decision-making without human involvement [9, 10, 8].

An AI project is a temporary organization dedicated to developing AI systems before their deployment for productive use or commercialization [7]. Project teams comprise the natural or legal people who are best placed to understand the system's design and its potential impacts on individuals and society [7, 11]. Heaton et al. [12] reinforces the pivotal role of the team, arguing that developers should design transparent, explainable AI systems. However, AI projects often involve tensions between financial targets, ethics, internal controls, and compliance [13].

Specifically, ethical and community-oriented AI projects should engage stakeholders, such as end-user groups or affected communities, as participatory collaborators in the development of AI systems [12, 11]. However, participatory development is time-consuming and often faces systemic barriers, including a lack of support and resources for brokering relationships, as well as short project timelines. AI end-user groups and affected communities are vastly underrepresented in AI research relative to developers, designers, and organizational leaders [14].

B. AI Societal Impacts

Within AI projects, project actors, including project sponsors, managers, and team members, function as moral agents [15]. Their in-project decisions can significantly impact the lives of individuals and their liberty, human rights, or civil rights [12]. The actions of project actors can also result in environmental, financial, reputational, and political harm. AI systems are artificial agents that embody a moral code, as the outcomes of their decisions or recommendations can benefit or harm individuals or society. Consequently, project actors bear a moral responsibility to apply fair, ethical, and transparent processes to the AI systems they develop Heaton et al. [12].

AI projects may also have societal and environmental impacts. Training large models consumes significant amounts of energy and water, and produces carbon emissions, resulting in high energy costs and negative environmental impacts [10, 16]. Moreover, data representativeness has power, value, and cost implications. Determining who is included in or excluded from AI datasets is valuable. Hence, data curation involves tradeoffs between costs, accuracy, quality, completeness, representativeness, and equality [16, 11].

Ryan and Stahl [10] described 11 ethical categories that users and developers should consider when fulfilling their moral responsibilities within AI projects: beneficence, dignity, freedom and autonomy, justice and fairness, non-maleficence, privacy, responsibility, solidarity, sustainability, transparency, and trust. While agencies and governing bodies, including UNESCO and the EU High-Level Expert Group, have established ethical principles that define an acceptable moral code for AI systems, these principles are non-binding for developers, users, and platforms [1]. Furthermore, the principles alone are limited in their impact on AI design and governance, as it can be challenging to translate concepts, theories, and values into practice [10, 17].

C. AI Regulations

Regional legislators have encoded ethical principles into laws and regulations relating to AI. For instance, the EU introduced the AI Act, China the Chinese Internet Information Service Algorithm Recommendation Management Regulations, and Korea the AI Basic Act [1, 2, 3]. Conversely, in May 2025, the United States (US) House of Representatives proposed legislation that would ban federal, state, and local governments from enforcing laws or regulating AI for a 10-year period [18]; this sweeping moratorium was unsuccessful, but its proposal demonstrates the political appetite for regulating AI at the federal level in the US. DePaula et al. [19] found that the 50 US states are addressing AI in an uneven, piecemeal fashion. "While several legislations regulate an aspect of AI, a common response was to create task forces and commissions to study AI, prohibit potentially discriminatory practices of AI, and provide funding to capture potential economic or educational benefits from AI" [19, p. 822]. Regardless, research by Hopkins and Booth [20] found that regulations are ineffective in changing the behaviors of technology professionals.

The EU AI Act defines an AI system by its capability features (such as levels of autonomy, adaptiveness, and inferential abilities). It classifies such systems according to the risks they pose to society and the environment [1, 9]. The Act assigns the roles of deployers and providers to those working in the field of AI system development and use. The deployer is the natural or legal person responsible for the system during its development. This role is responsible for completing risk and compliance assessments, as well as registration, before proceeding to the next stage. The provider is the natural or legal person accountable for the system's operations.

D. AI Financial Benefits

AI 'projects aim to realize their target benefits and enhance organizational performance…"target" benefits are intentionally sought by project funders and strategically set before project commencement' [5, p. 656]. AI systems utilize data and algorithms to either benefit an organization's value proposition or monetize data. Generative AI applications such as ChatGPT are improving the performance in tasks that were previously unaffected by automation technology; empirical studies show performance gains between 14% and 50% in several business contexts [4].

"Value propositions define an organization's position in a market, its core skills and how these are applied for the benefit of others" [6, p. 168]. Consequently, introducing AI systems presents a business and technical challenge that should yield target benefits and tactical goals, such as increased revenue or improved performance efficiency. Organizations expect projects to be executed efficiently, delivering results within a given timeline and budget. They must therefore balance meeting target objectives with consideration of societal and environmental impacts, as well as the need to act ethically and comply with regulations.

III. Theoretical Framework and Hypothesis Development

This section outlines the theoretical framework and the development of hypotheses.

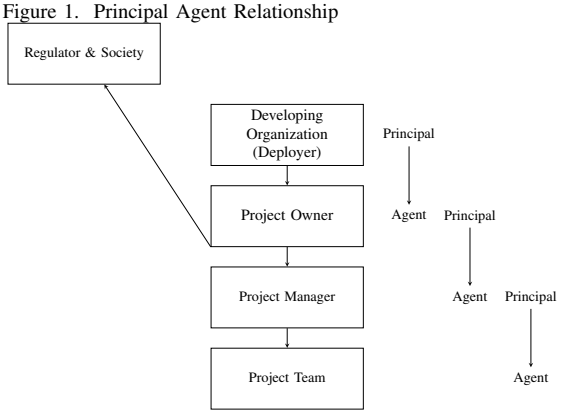
A. Theoretical Framework

Jensen and Meckling [21, p. 308] defined agency relationships as contracts "under which one or more persons (the principal(s)) engage another person (the agent) to perform some service on their behalf which involves delegating some decision-making authority to the agent." Within the principal–agent relationship, agents are accountable to principals for their conduct. The agents have obligations to fulfill and should face consequences for their failure to perform. Moreover, agents must have the capacity to act intentionally, make choices, and exert influence with some degree of autonomy in the decision-making process [12].

Relating the roles in the EU AI Act to actors in the AI stakeholder accountability model [7] and agency theory, the deployer represents the developing organization, while the provider represents the operating organization. Within the deployer's role, the project owner acts as an agent for the principal, the developing organization. Organizational strategic goals and policies flow downward from the principals to agents [14, 22].

When building AI systems, the project owner functions as the principal, with the project manager serving as the agent. Project management standards stipulate that the project manager is responsible for planning and managing the project, directing the performance of planned activities and overseeing the project's technical, administrative, and organizational interfaces [23]. The team thus serves as an agent of the project manager. Fig. 1 illustrates the principal–agent relationships concerning the developing organization and the regulators and society.

Figure 1. Principal Agent Relationship



The AI stakeholder accountability model specifies and defines the complex relationships between actors and AI success factors, as well as the conduct and obligations that characterize them [7]. For this study, the model serves as the basis for defining accountability, i.e., the obligations the agent has to the principal; the success factors describe the conduct, events, acts, or deliverables in that relationship.

Table I synthesizes the knowledge from the AI stakeholder accountability model into a matrix that consolidates the 17 success groups into 11 accountability indicators, organized across the columns. The rows represent the agent–principal relationships and signify whether the project agent has an obligation to the principal for a given indicator. Success factors encompass ethical principles and activities mandated by AI regulations, representing deliverables, acts, or events that an AI project must include [7].

The synthesis identifies ethical and regulatory deliverables in the agent–principal relationship for AI projects, providing a basis for quantitatively evaluating AI accountabilities in practice. We selected the AI stakeholder accountability model as a theoretical basis for its alignment of project success with the AI ethical principles from Ryan and Stahl [10], regulatory and financial accountabilities, and project actors. Success factors identify the deliverables necessary for achieving project objectives. They can be used to establish an accountability standard that governs the relationship between project actors and stakeholders. Furthermore, the model determines what actions the project actors should be held accountable for and identifies the project owner as accountable to the regulators on behalf of the developing organization.

B. Hypothesis Development

Table I defines the obligations of project agents in achieving regulatory compliance, developing ethical AI systems, and delivering financial benefits using the 11 accountability indicators. Thus, in practice, accountabilities should vary by project role as defined in the matrix. This leads to the formulation of the following hypothesis:

Hypothesis 1: The (A) project manager, (B) owner, and (C) team will be responsible for their subset of the 11 accountability indicators in the agent–principal matrix.

In the context of AI projects, [7, p. 475] notes that "[f]irst and foremost, the scope established by the project owner is an essential artifact in designing AI systems." The project scope is therefore an independent factor that determines the tasks and accountabilities of individual project actors. Similarly, the EU AI Act requires different obligations depending classification of AI system according to its use case and technical features [1, 9]. For example, AI use cases for social scoring or manipulation are prohibited. Furthermore, system features such as model type, adaptiveness after deployment, personal data use, or inferring input data outputs would affect the system classification, computing costs, and environmental impact [1, 9, 16]. Therefore, the following hypothesis is proposed:

Hypothesis 2: Depending on the system scope, there will be a significant difference in accountability for (A) regulatory compliance, (B) energy costs, and (C) sustainability.

Within AI projects, hidden information and actions create challenges for the agent–principal relationship. Hidden information refers to the principal's difficulties in determining whether the agent has the necessary qualities and skills to perform a task in the principal's interest [24]. Furthermore,

Table I
ACCOUNTABILITY BY AGENT-PRINCIPAL RELATIONSHIP AFTER [7]

		Accountabilities										
		Energy Cost **EC**	Ethic Practices **EP**	Financial Benefits **FB**	Intellect Property **IP**	Societal Impacts **SI**	Prj Efficiency **PE**	Privacy Protect **PP**	Product Quality **PQ**	Reg Compliance **RC**	Transparency Usability **STU**	Sustainability **SY**
Agent	Principal											
Prj Mgr	Owner	X	X		X		X	X		X	X	
Owner	Organization	X		X	X		X		X	X		
	Regulator		X					X		X		
	Society		X			X		X	X		X	X
Team	Prj Mgr	X					X					
	Owner		X		X			X		X	X	

Abbreviations: Prj-Project, Mgr-Manager; Reg-Regulatory

some AI systems are black-boxes with hidden features caused by complex data and models [9]. Nevertheless, the AI systems are the result of actions by the project teams, influenced by their choices, biases, values, and intentions [12]. Hidden actions refer to activities that may be in the agent's interests but not the principal's. Agency theory explains the issue of moral hazards, where within the principal–agent relationship, individuals may pursue their personal interests even when they conflict with the team's objectives [21, 24, 25].

Hypothesis 3: For an autonomous system scope, agents are more likely to be accountable for (A) system transparency and usability, and (B) protecting intellectual property or financial gains.

The performance of project actors is based on what their organization considers important and what they will be held accountable for. Thus, if the organization considers the development of fair, understandable, and sustainable systems important, project agents will be accountable for tasks that deliver responsible AI systems [7, 8, 10].

Hypothesis 4: If the organization emphasizes the importance of creating fair and understandable algorithms: (A) the project manager is more likely to be responsible for implementing ethical practices; (B) the owner is more likely to be responsible for societal impacts; and (C) the team is more likely to be responsible for system transparency and understandability, product quality, and societal impacts.

Hypothesis 5: If the organization emphasizes the importance of environmental sustainability, the (A) project manager, (B) owner, and (C) team are more likely to be responsible for sustainability, energy cost, or both.

Organizations undertake projects to achieve target benefits and enhance organizational performance [4, 5, 6].

Hypothesis 6: (A) A majority of (a) owners and (b) teams will be involved in achieving financial benefits, intellectual property, or both. (B) A majority of (a) project managers and (b) owners will be accountable for the project's efficiency.

The AI stakeholder accountability model does not associate achieving financial benefits, including revenue generation, cost efficiency, or project efficiency, with any ethical principle. Furthermore, the project team must make several tradeoffs when building AI systems, including accuracy, energy cost, financial costs, project efficiency, and regulatory compliance, as well as transparency and intellectual property rights, and legal safeguards and system flexibility [7, 13, 25].

Hypothesis 7: The following will be negatively correlated: (A) financial benefits and regulatory compliance responsibility, (B) financial benefits and privacy protections responsibility, and (C) intellectual property and system transparency and understandability.

IV. RESEARCH METHODOLOGY

This section describes the research methodology, including the data collection methods, measurement instrument, analysis methods, and reliability and validity measures. The unit of analysis for the study is the project.

A. Data Collection

This study collected data on accountability within an AI project using a web-based survey hosted on SurveyMonkey, a global survey platform that utilizes a proprietary panel of survey respondents as its audience. Data were collected in November 2023 and November 2024 using a SurveyMonkey audience panel for a US audience.

Respondents provided consent to participate. No personally identifiable data were collected, and no compensation was offered. Respondents who selected "None" as an AI project role were disqualified. The survey targeted respondents in consulting, information technology, management, and project management job roles.

For the 2023 survey, 228 invitations were sent to potential participants. Of these, 121 (47%) were disqualified based on their AI project role, providing 107 (53%) usable responses. SurveyMonkey reported an error rate of +/- 10%. For the 2024 survey, 151 invitations yielded 8 (5%) disqualifications and 143 (95%) usable responses, and an error rate of +/- 8%.

Of the combined 2023 and 2024 responses, 57 of the usable respondents had not completed at least three AI projects, and 29 entries had invalid data in the experience field. These were excluded, resulting in a final total of 164 valid responses.

B. Measurement Instrument

We employed a three-stage process of content, clarity, and expert review to develop the measurement instrument, which was in English. The expert review was performed by three professionals with experience in at least three AI projects in the last ten years. All stages resulted in adjustments to the text.

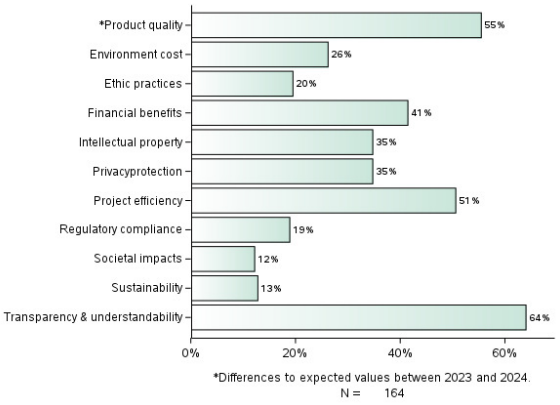


Figure 2. Accountability Indicator Demographics

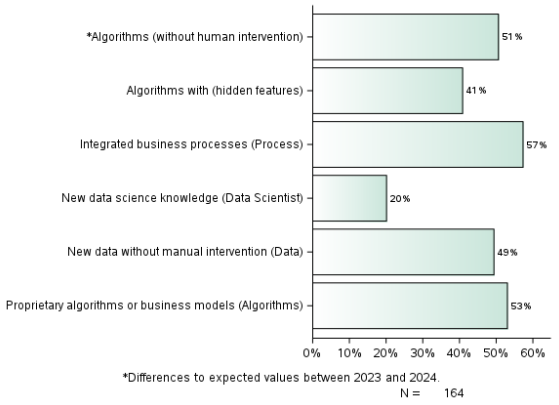


Figure 3. AI System Scope Demographics

This instrument defined algorithms, AI systems, and AI projects as follows:

> "An algorithm is a defined, repeatable process and outcome based on data, processes, and assumptions. An Artificial Intelligence (AI) system is a computer system that incorporates algorithms that learn from data and support human decision-making or autonomously make decisions. An AI project is a temporary organization (from few months to many years) that develops one or more AI systems" (author's definitions).

Respondents selected their *project role* from seven options and were grouped into composite role measures for the project owner, manager, and team. The roles and groups were taken from Miller [7].

Table I, which was also synthesized from Miller [7], was used as a bases for defining the measures for individual project *accountability* and *agent–principal relationship* accountabilities. For individual project *accountability*, respondents selected from 11 options or entered "other"; these options are shown in Table II under "Accountability Questions." Each selection created a binary measure with a value of one if it was selected and zero if not. Fig. 2 visualizes the statistics for the responses to this question. For the *agent–principal relationship* accountability combinations, a dummy variable was created with an indicator for each agent-principal pair that had an expected accountability in Table I.

Respondents selected one or more of six characteristics to describe a specific *AI system scope* they developed as part of their project [1, 9]. Fig. 3 presents the statistical responses regarding AI system characteristics. The respondents were also asked *"Of the 10 levels of automation. . . [i]n one of my last 3 AI projects, AI was at level. . . ?"* A table based on Sheridan et al. [26] provided 10 levels of automation to guide the participants' responses, allowing them to select answers from 1 to 10. A dummy variable was created where levels 1 to 6 were considered human decisions and levels 7 to 10 were considered AI decisions. The responses to questions about *AI system scope* and the *level of automation* were used to classify the AI system as autonomous or non-autonomous systems using latent class analysis (LCA).

For *experience*, respondents entered the number of AI projects they had completed in the last decade; those reporting fewer than three were not included in the analysis. To ascertain the importance of *fairness* and *sustainability*, two seven-point Likert scale questions were included, and binary measures were created for each.

Table II
ACCOUNTABILITY QUESTION

What aspects of project performance were you accountable for in the AI project referenced in question Q2? (check all that apply)

	Response Options	**Analysis Identifier**	**ID**
AI system scope	Product quality (training data, algorithm, user interface, system configuration)	Product quality	PQ
	System transparency and understandability, usage controls, decision quality	Transparency & understandability	STU
Regulatory	Ethic practices, training, policies, oversight	Ethic practices	EP
	Regulatory and legal compliance	Regulatory compliance	RC
	Protecting data privacy and confidentiality	Privacy Protections	PP
Societal Impacts	Social impacts of system usage (civil and human rights protections, etc.)	Societal impacts	SI
	Environmental sustainability of system usage	Sustainability	SY
Benefits	New revenue generation, cost savings, license fees	Financial benefits	FB
	Project time, cost, and quality performance—iron triangle	Project efficiency	PE
	Environmental costs, energy costs environmental impacts	Energy cost	EC
	Intellectual property protections (trade secrets, proprietary algorithms, obtaining patents, etc.)	Intellectual property	IP

C. Data Analysis

The SAS Studio Release 3.8 (Enterprise Edition) was utilized for quantitative analysis, statistical tests, and bias checks. The accountability data were binomial and non-normally distributed. Thus, parametric measures were suitable for the analysis. The Wilcoxon test was used to compare the means of variables and determine the significance of the comparison. Kendall's tau-b correlation coefficient was

used to evaluate the strength, direction, and significance of the relationship between measurement items. The correlation coefficient does not indicate which variable causes a change in another variable, nor does it imply causality. The Wilcoxon scores and correlations with a ρ-value of less than 0.05 were considered significant with a 95% confidence interval [27]. The rank shows the relevant importance of the indicator.

LCA clustering was employed to identify homogeneous clusters within the scope of the AI system. The dimensions were modeled in R version 4.0.2 using poLCA for the analysis. Seven items for the system scope and level of automation were specified as indicators in the model. LCA was executed with 50,000 iterations for 1–10 classes and 10 repetitions per model. The class model was selected by evaluating the Bayesian information criterion (BIC) fit statistics, log-likelihood, entropy, and the model's theoretical interpretability [28].

D. Reliability and Validity

This study relied on the quality of demographics and professional data from SurveyMonkey for sample construction. SurveyMonkey conducts periodic quality checks using surveys and panel calibration studies to help ensure the quality of open-ended responses and identify poor respondent behavior [29]. Moreover, Bentley et al. [30] conducted a comparative study on several audience panels and found the SurveyMonkey audience panel had an error margin that was "statistically indistinguishable" from more expensive panels. Thus, the sample construction was considered valid for the study.

The survey included a question to establish whether respondents had participated in both 2023 and 2024 to account for the effects in repeated surveys; 31% indicated they took both surveys. This repetition impacts the uniqueness of the responses and the effect size of the results. First, our unit of measure is the project. A person is likely to change projects between years; machine learning or AI projects are from a few weeks to several months [16]. Thus, the project characteristics may vary for the repeated responses, which is valuable to our analysis. Second, effect sizes must be evaluated based on the average number of repeated measures, as additional repeated measures provide less novel information [31]. Analysis of effect sizes was not relevant to our study; nevertheless, this factor must be considered, as it decreases the sample size, which impacts the generalizability of the study data.

For further sample validity, we used a screening question and performed quantitative checks on the survey responses to identify missing data, extreme responses (where the same response was recorded for all questions), and irregular responses to critical questions, including the entry of odd responses to open-ended questions regarding years of experience. Consequently, we removed 29 entries that were deemed invalid.

We conducted validity checks for common method bias and response bias. Response bias was assessed between survey years (2023 and 2024) to determine whether there were significant differences between the responses. Responses to the question on product quality in the accountabilities section and algorithm use without human intervention in the system scope section both exhibited significantly higher scores in 2024 than in 2023. These differences were managed in the analysis and reported in the results to avoid spurious interpretations.

Table III
DEMOGRAPHICS

Title	Label	N	Percent
Age	18 – 29	15	9.15
	30 – 44	84	51.2
	45 – 60	59	36.0
	> 60	6	3.66
Gender	Male	103	62.8
	Female	61	37.2
Experience	< 5 projects	47	28.7
	5 projects	41	25.0
	6 – 7 projects	14	8.54
	8 – 15 projects	30	18.3
	> 15 projects	32	19.5
Owner	Chief Executive Officer, Business Leader	31	18.9
	Chief Information Officer, IT Manager	53	32.3
Project Manager	Project Leader/Manager, Product Owner	40	24.4
Team	Data Scientist, Computer Vision, AI Specialist	14	8.54
	Data Engineer, Data Custodian, Curator	6	3.66
	Architect, Developer, Computer Scientist, IT	7	4.27
	Business User/ Analyst, Subject Matter Expert	13	7.93
Source	SurveyMonkey 2023	54	32.9
	SurveyMonkey 2024	110	67.1
AI system	Autonomous system	79	48.2
	Non−autonomous system	85	51.8
Automation	Level 1 – 6 Human Decisions	85	51.8
	Level 7 – 10 AI Decisions	79	48.2
Fairness	Not important	9	5.49
	Important	155	94.5
Sustainability	Not important	24	14.6
	Important	140	85.4

Given that all variables were collected using the same survey instrument from a single source, Harman's single-factor test was performed to check for common method bias [32], using unrotated factor analysis to analyze the scope and accountability variables. A single factor explained 16.53% of the variance, which is well below the 50% heuristic for indicating bias. Thus, common method bias was not considered an issue.

We applied the recommendations presented by Weller et al. [28] to prepare the LCA clusters.

V. FINDINGS

This section describes the study's findings. Table IV provides a summary of the hypotheses and findings.

A. Demographics

Table III, Fig. 2, and Fig. 3 present the demographics of the respondents. Based on the AI system scope, less than half of the participants (42%) indicated that they had developed *autonomous* systems. Fig 4 shows the mean score for the features by AI system scope classification. The majority of respondents reported that *fairness* (95%) and sustainability (85%) were important to their organizations.

B. Agent-Principal Accountabilities

The data did not support the accountability expectations from the AI stakeholder accountability model. Table V displays the basic statistics for the ratio of accountabilities

Table IV
SUMMARY OF HYPOTHESES FINDINGS

ID	Hypothesis	Supported
1	The (A) project manager, (B) owner, and (C) team will be responsible for their subset of the 11 accountability indicators in the agent–principal matrix.	(A) No (B) Yes (C) Yes
2	Depending on the system scope, there will be a significant difference in accountability for (A) regulatory compliance, (B) energy costs, and (C) sustainability.	(A) No (B) Yes (C) Yes
3	For an autonomous system scope, agents are more likely to be accountable for (A) system transparency and usability, and (B) protecting intellectual property or (C) financial gains.	(A) No (B) No (C) Yes
4	If the organization emphasizes the importance of creating fair and understandable algorithms: (A) the project manager is more likely to be responsible for implementing ethical practices;	No
	(B) the owner is more likely to be responsible for societal impacts; and	No
	(C) the team is more likely to be responsible for system transparency and understandability, product quality, and societal impacts.	No
5	If the organization emphasizes the importance of environmental sustainability, the (A) project manager, (B) owner, and (C) team are more likely to be responsible for sustainability, energy cost, or both.	No
6	(A) A majority of (a) owners and (b) teams will be involved in achieving financial benefits, intellectual property, or both.	No
	(B) A majority of (a) project managers and (b) owners will be accountable for the project's efficiency.	No
7	The following will be negatively correlated: (A) financial benefits and regulatory compliance responsibility,	No
	(B) financial benefits and privacy protections responsibility, and	No
	(C) intellectual property and system transparency and understandability.	Yes

Table V
AGENT-PRINCIPAL RELATIONSHIP STATISTICS

Agent	Principal	N	MIN	MEAN	MAX	STD
All	All Indicators	164	0.00	3.71	11.0	2.06
Owner	Organization	84	0.00	2.40	6.00	1.18
	Regulator	84	0.00	0.70	3.00	0.89
	Society	84	0.00	2.79	7.00	1.67
Prj Mgr	Owner	40	0.00	2.48	5.00	1.22
Team	Prj Mgr	40	0.00	0.65	2.00	0.74
	Owner	40	0.00	1.70	5.00	1.30

Legend: Roles: Prj-Project, Mgr-Manager
Statistics: N-Number, Std-Standard Deviation

in the agent–principal relationship. For example, the project manager and owner relationship is expected to include seven accountabilities, as shown in Table 1. However, the maximum was five, the mean was 2.48, and the standard deviation was 1.22; thus, no project manager claimed all the accountabilities suggested by the model. The project owner and teams fared better since the maximum was achieved; nevertheless, the means were relatively low. Table VI presents the percentage of accountability by respondent role. Therefore, hypotheses 1B and 1C are supported for the owner and team, but 1A is rejected for the project manager.

Notably, the respondents' experience level influenced several accountabilities. Respondents with 15 or more projects were accountable for intellectual property (IP) and product quality (PQ) more than other experience groups; those with fewer than five projects were responsible for ethical practices (EP), and those with fewer than six projects were more responsible for project efficiency (PE).

C. Impact of AI System Scope

We assessed accountability for each project role, where the AI system scope was the independent variable and the

Table VI
AGENT ACCOUNTABILITY STATISTICS

	Accountabilities										
	EC	EP	FB	IP	SI	PE	PP	PQ	RC	STU	SY
Agent	Percent										
Owner	13	9	21	18	7	29	17	32	10	35	8
PM	7	5	10	9	4	12	8	15	4	16	1
Team	6	5	10	9	1	10	10	9	5	13	4

Legend: Roles: PM-Project Manager;Accountabilities: EC-Energy Cost, EP-Ethic Practices, FB-Financial Benefits, IP-Intellectual Property, SI-Societal Impacts, PE-Project Efficiency, PP-Data Privacy, PQ-Product Quality, RC-Regulatory, STU-Traceability & Understandability, SY-Sustainability
Statistics: Owner=84, Prj Mgr=40, Team=40

accountabilities were the dependent variables. The scope had a significant influence on the accountabilities. There were significant differences in several indicators: financial benefits (FB) (H=8.55, ρ =0.00), energy cost (EC) (H=13.28, ρ =.00), and sustainability (SY) (H=5.19, ρ =.02). Consequently, agent–principal accountabilities for the owner–organization (H=14.69, ρ =0.00) and team–project manager (H=15.26, ρ =0.00) relationships were impacted. Hypothesis 2A is therefore rejected, and hypotheses 2B and 2C are supported. Table VII shows the Kruskal-Wallis mean rank statistics for each accountability and agent-principal pair by AI system scope.

System transparency and understandability (STU) ranked highest for both autonomous and non-autonomous scopes. The Kendall correlation presented in Table VIII revealed a negative, insignificant correlation between STU and the AI system scope (τ = −.06, ρ >.05). Thus, possessing hidden features did not affect transparency requirements or move the agent away from STU. Nevertheless, the EC and FB accountabilities were ranked significantly higher for the autonomous system scope. Intellectual property (IP) had no significant difference by scope. Therefore, hypothesis 3C is supported, while hypotheses 3A and 3B are rejected.

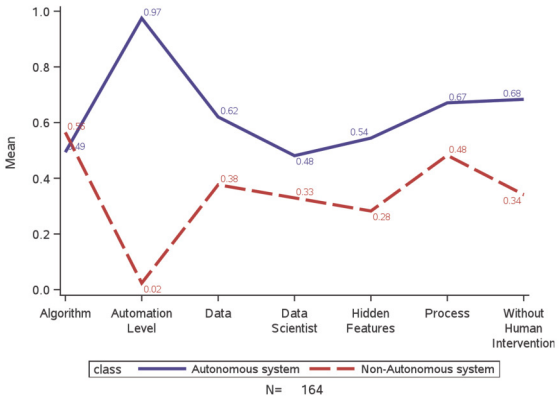


Figure 4. AI system scope classification by features for the survey population.

D. Regulatory Accountabilities

No significant difference was found in regulatory compliance (RC) accountability in terms of the AI system scope. Since many types of laws and regulations, such as those related to healthcare and disabilities, apply in the US regardless of

Table VII
ACCOUNTABILITY MEAN RANK STATISTICS BY AI SYSTEM SCOPE

			Autonomous system (N= 79)			Non-Autonomous system (N= 85)			Overall (N= 164)			Kruskal Wallis	
Category	**Label**	**ID**	**Mean**	**Std**	**Rank**	**Mean**	**Std**	**Rank**	**Mean**	**Std**	**Rank**	**H**	**ρ Value**
Agent-Principal													
Agent – Principal	Owner – Society		2.75	1.62	1	2.52	1.69	1	2.63	1.66	1	0.29	0.59
	Prj Mgr – Owner		2.71	1.53	2	2.28	1.44	2	2.49	1.49	2	2.33	0.13
	Owner – Organization		2.65	1.20	3	1.93	1.18	3	2.27	1.24	3	14.69	0.00
	Team – Owner		1.73	1.30	4	1.71	1.23	4	1.72	1.26	4	0.04	0.85
	Team – Prj Mgr		0.97	0.64	5	0.58	0.62	6	0.77	0.66	5	15.26	0.00
	Owner – Regulator		0.67	0.89	6	0.79	0.93	5	0.73	0.91	6	0.77	0.38
Accountabilities													
AI system scope	Transparency & Usability	STU	0.67	0.47	1	0.61	0.49	1	0.64	0.48	1	0.62	0.43
	Product Quality	PQ	0.57	0.50	3	0.54	0.50	2	0.55	0.50	2	0.13	0.72
Regulatory	Privacy Protection	PP	0.33	0.47	7	0.36	0.48	4	0.35	0.48	6	0.23	0.63
	Ethic Practice	EP	0.16	0.37	10	0.22	0.42	7	0.20	0.40	8	0.90	0.34
	Regulatory Compliance	RC	0.18	0.38	9	0.20	0.40	8	0.19	0.39	9	0.14	0.71
Societal Impacts	Sustainability	SY	0.19	0.39	8	0.07	0.26	11	0.13	0.34	10	5.19	0.02
	Social Impact	SI	0.15	0.36	11	0.09	0.29	10	0.12	0.33	11	1.27	0.26
Benefits	Project Efficiency	PE	0.58	0.50	2	0.44	0.50	3	0.51	0.50	3	3.52	0.06
	Financial Benefit	FB	0.53	0.50	4	0.31	0.46	6	0.41	0.49	4	8.55	0.00
	Intellectual Property	IP	0.39	0.49	6	0.31	0.46	6	0.35	0.48	6	1.34	0.25
	Energy Cost	EC	0.39	0.49	6	0.14	0.35	9	0.26	0.44	7	13.28	0.00

Abbreviations: N=Number observations, Std=standard deviation, KW=Kruskal-Wallis, ρ=significance, Prj=Project, Mgr=Manager

the technology, the results are not entirely unexpected. Approximately 10% of the owners, 4% of the project managers, and 5% of the team members claimed RC accountability. RC accountability was significantly correlated with FB (τ =.16, ρ <.05), privacy protections (PP) (τ =.27, ρ <.0001), ethical practices (EP) (τ =.27, ρ <.0001), and societal impacts (SI) (τ =.20, ρ <.05). In sum, both hypothesis 2A, positing that the scope will impact RC, and hypothesis 7A, proposing that there will be negative tradeoff between RC and FB, are rejected.

E. Societal Accountabilities

When controlling for interaction for the importance of sustainability or fairness, there was no significant difference in agent–principal accountabilities. First, all owners indicated that fairness was important. Second, there were no significant differences for any agents regarding the accountabilities of EP, STU, PQ, or SI when controlling for fairness or sustainability.

Both SI (τ =.22, ρ <.01) and SY (τ =.16, ρ <.05) were significantly correlated with FB. This could imply that organizations see their projects as benefiting society and considering sustainability goals. The AI stakeholder accountability model attributes these accountabilities exclusively to the owner. However, only a small percentage of owners (8%) claimed them, and the other roles also claimed these responsibilities. Hypotheses 4 and 5 are therefore rejected.

F. Financial Accountabilities

Using the accountabilities of FB or IP to represent the financial gains from the project, a minority of agents had these responsibilities: project owners (31%), project managers (15%), and team members (14%). Surprisingly, only 12% of project managers and 29% of owners claimed responsibility for PE, a standard metric of project management performance including budget, time, and quality. Thus, hypotheses 6A and 6B are rejected.

EC was significantly correlated with (τ =.29, ρ <.0001) and impacted by (SS=2.33, df(1)=13.05, ρ =<.0001) the AI system scope, but it was not affected by the sustainability importance measure as expected.

G. Tradeoffs Between Accountabilities

The tradeoff between IP and STU (τ = −.01, ρ >.05) was negative, but this was not significant. However, the tradeoffs anticipated by the AI stakeholder accountability model between FB and RC (.16, ρ <.01) and FB and PP (.01, ρ >.05) were not realized. Thus, hypotheses 7A and 7B are rejected, while hypothesis 7C is supported. Fig. 5 visualizes the Kendall Tau correlations that were significant with an effect size greater than .20.

VI. DISCUSSION AND CONCLUSIONS

This study employed the AI stakeholder accountability model, grounded in agency theory, to establish an accountability standard governing the relationships among project actors, regulation, ethical principles, and benefits. A quantitative analysis of survey data examined individuals' accountabilities in practice within AI projects. Additionally, we identified two clusters of AI systems—autonomous and non-autonomous—based on seven features. The *autonomous* cluster describes systems with a high level of machine-based decision-making that allow for limited human intervention, while the *non-autonomous* excludes these features.

This study examined how AI systems (both autonomous and non-autonomous) influence accountability, as well as the importance of sustainability and fairness in this context. As expected, we found that accountabilities shift based on the scope of the AI system; unexpectedly, we also determined that most expected accountabilities are not undertaken in practice.

The intersection between target benefits, societal impacts, and regulatory compliance depends on the project owner's role. Our findings suggest that the owner's accountability for regulatory compliance remains relatively unchanged, regardless of the scope of the AI system. However, their accountability towards the organization does change.

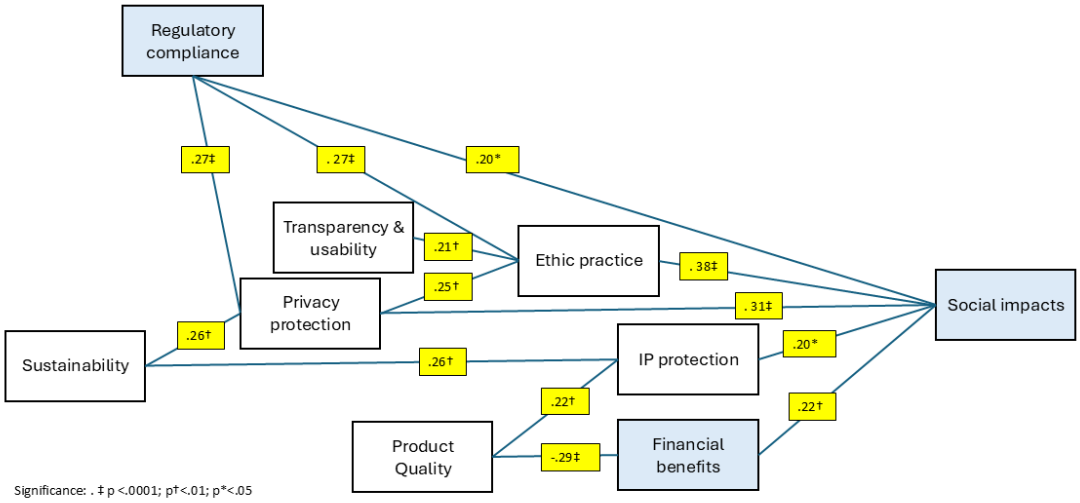


Figure 5. Kendall Tau Correlation Visualization

Most respondents practiced accountability regarding system transparency and understandability. This finding is consistent with Heaton et al. [12], who suggest that developers should design transparent and explainable AI systems to promote trust and mitigate any negative consequences that could arise from system use. Moreover, the transparency and understandability indicator is significantly correlated with ethical practice.

Ethical practice plays a central role in regulatory compliance and social impacts, and is further correlated with financial performance. This suggests that ethical practices may be the driving force behind the achievement of all other benefits. The analysis across experience levels further supports the argument for centering ethical practices. It showed that respondents with four or fewer projects indicated a greater responsibility for ethical practices than any other experience group. Taken together with the other indicators, this suggests that novices rely on documented practices (ethical practice) or traditional project performance measures (project efficiency) in their job performance. At the same time, more experienced individuals focus on creating content (intellectual property) and producing quality results (product quality).

We examined the fluidity of accountabilities, utilizing owner, project manager, and team roles from the sponsoring organization (i.e., the deployer, as defined by the EU AI Act). The accountabilities change based on the AI system's scope; if the project includes a scope typical of building autonomous decision-making tools, there is greater accountability, with a keen focus on the business impacts of product quality, intellectual property, and energy costs.

The lack of some expected tradeoffs may be attributed to the study's design. Wachnik [25] notes that many moral hazards in projects are due and the dangers linked to opportunistic supplier behaviour in a client-supplier relationship. The study lacked information on the respondents' organizational relationships, so it did not permit such analysis. Nevertheless, some tradeoffs between compliance and obtaining target benefits appear to be based on role assignments. Not all accountabilities are observed by a single agent, which helps avoid conflicts between personal and organizational interests [21, 24, 25]. Preventing moral hazards should incentivize the optimization of project organizations, including their contractual structures, team size, and role assignments.

Despite the robustness of the AI stakeholder accountability model, individuals hold less than 50% of the expected accountabilities in practice. However, we anticipate that this will change over time. First, the monetary and environmental cost of developing new and tuning existing AI models becomes extremely expensive, quickly [16]. Second, the EU AI Act comes into full effect by 2027. In response to non-compliance, the EU can impose "administrative fines of up to 35 000 000 EUR. . . or up to 7% of its total worldwide annual turnover. . . whichever is higher" [1, p. 115]. Third, generative AI capabilities extended with digital tools or integrated with robotics are estimated to produce "around 0.4-0.9 p.p. [percentage point] contribution to annual labour productivity growth, when assuming a standard capital multiplier of 1.5" [4, p. 10]. Finally, there are strong incentives for firms to implement one or more AI systems for economic benefits or to address competitive pressures [6].

This study shows that for ethical AI development, it is necessary to establish a project role assignment matrix; this finding is consistent with Herrera [9] and Kuehnert et al. [14]. Here, we propose that the AI stakeholder accountability model evaluated in the current study could provide a baseline standard for assigning AI project responsibilities.

A. Theoretical Contributions

Our study's findings quantify subjective and theoretical speculation about who is accountable for the financial, regulatory, and societal impacts of AI projects. This study identifies the relationships among system features, automation levels, and accountability indicators in AI projects by empirically investigating the practices of experienced project participants. The results contribute to the literature on AI, ethics, and

Table VIII
KENDALL TAU B CORRELATION COEFFICIENTS

Kendall Tau b Correlation Coefficients											
	AIS	**PQ**	**STU**	**PE**	**FB**	**EC**	**SY**	**IP**	**RC**	**PP**	**EP**
PQ	-.03										
STU	-.06	-.03									
PE	-.15	.07	-.03								
FB	-.23†	-.29‡	.12	.04							
EC	-.29‡	-.05	-.04	-.02	.06						
SY	-.18*	.05	.06	.09	.16*	.10					
IP	-.09	.22†	-.01	.03	.06	.18*	.26†				
RC	.03	-.01	.10	-.02	.16*	.00	.14	.11			
PP	.04	.19*	.09	-.07	.01	.15	.26†	.17*	.27‡		
EP	.07	.07	.21†	.18*	.15	.16*	.18*	.09	.27‡	.25†	
SI	-.09	.11	.12	.14	.22†	.16*	.19*	.20*	.20*	.31‡	.38‡

Significance: ‡p<.0001, †p < .01, *p < .05; Abbreviations: AISo=AI System Scope, Exp=Experience, PQ=Product Quality, STU=System Transparency and Usability, PE=Project Efficiency, FB=Financial Benefits, EC=Energy Cost, SY=Sustainability, IP=Intellectual Property, RC=Regulatory Compliance, PP=Privacy Protections, EP=Ethic Practices, SI=Social Impacts

project management by aligning regulatory requirements with a project success model. Although research on ethical principles in AI is widely available, this research addresses the criticism that organizational leaders are underrepresented in AI research [14].

B. Practical Implications

Project sponsors should understand the interplay between various project assignments and define performance measures to avoid conflicts that may create moral hazards. Furthermore, the scope of the AI system should not be underestimated when defining a project organization. Thus, when staffing decisions require a tradeoff in role assignments, it is crucial to consider ethics, regulations, and performance efficiency.

C. Limitations and Future Research

One limitation of the present study was its data collection strategy. The survey was distributed to randomly selected individuals in the US, did not consider a single project, and relied on self-report measures. First, relying on a sample consisting exclusively of respondents from the US may limit the generalizability of the findings, especially given that regions such as Europe have more advanced AI regulations in place. Second, the study did not assess different people from the same project team. Thus, it limited the ability to investigate team assignments and moral hazards. Third, the survey relied on self-reports by the participants. Thus, bias may have been introduced by individual memories or recalls.

To mitigate the limitations of this study, future research could focus on a case study to quantify the engagement of separate roles, or employ strategies to assess a single project environment. Future studies should investigate AI performance measures at a granular level by project role. Finally, this analysis was conducted from the deployer's perspective and should also be conducted from the promoter's perspective. The system configuration (SC), user interface (UI), and usage control (UC) qualities were not included in the survey as standalone items and were therefore excluded from the analysis. This was an acceptable exclusion, as the AI stakeholder accountability model identified these qualities as important for operations, which were not analyzed in this study. Although risk and quality management are important components of the EU AI Act, our survey did not include an independent question on these elements, so this aspect could not be independently analyzed. These are interesting avenues for future research to explore.

D. Conclusion

In conclusion, regulatory compliance, financial benefits, and societal impacts are not mutually exclusive project goals and coexist as shown in Fig. 5.

REFERENCES

[1] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[2] M. Theresia, "Newly enacted law sets basis for nat'l development of AI," Dec 27, 2024 2024.

[3] M. Sloane and E. Wüllhorst, "A systematic review of regulatory strategies and transparency mandates in AI regulation in europe, the united states, and canada," *Data & Policy*, vol. 7, 2025. https://dx.doi.org/10.1017/dap.2024.54

[4] F. Filippucci, P. Gal, and M. Schief, "Miracle or myth? assessing the macroeconomic productivity gains from artificial intelligence," OECD Publishing, Report, 2024.

[5] O. Zwikael, Y.-Y. Chih, and J. R. Meredith, "Project benefit management: Setting effective target benefits," *International Journal of Project Management*, vol. 36, no. 4, pp. 650–658, 2018. https://doi.org/10.1016/j.ijproman.2018.01.002

[6] C. F. Breidbach and P. Maglio, "Accountable algorithms? the ethical implications of data-driven business models," *Journal of Service Management*, vol. 31, no. 2, pp. 163–185, 2020. https://doi.org/10.1108/JOSM-03-2019-0073

[7] G. J. Miller, "Stakeholder-accountability model for artificial intelligence projects," *Journal of Economics and Management*, vol. 44, no. 1, pp. 446–494, 2022. https://doi.org/10.22367/jem.2022.44.18

[8] K. Kieslich, B. Keller, and C. Starke, "Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence," *Big Data & Society*, vol. 9, p. 205395172210929, 2022. https://dx.doi.org/10.1177/20539517221092956

[9] F. Herrera, "Attentiveness on criticisms and definition about explainable artificial intelligence," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024, Conference Proceedings, pp. 45–52. https://dx.doi.org/10.15439/2024F0001

[10] M. Ryan and B. C. Stahl, "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications," *Journal of Information, Communication and Ethics in Society*, vol. 19, no. 1, p. 61–86, 2021. https://dx.doi.org/10.1108/JICES-12-2019-0138

[11] L. H. Ajmani, N. A. Abdelkadir, and S. Chancellor, "Secondary stakeholders in AI: Fighting for, brokering, and navigating agency," p. 1095–1107, 2025. https://dx.doi.org/10.1145/3715275.3732071

[12] D. Heaton, J. Clos, E. Nichele, and J. E. Fischer, "The social impact of decision-making algorithms: Reviewing the influence of agency, responsibility and accountability on trust and blame," p. Article 11, 2023. 10.1145/3597512.3599706

[13] P. S. Scoleze Ferrer, G. D. A. Galvão, and M. M. de Carvalho, "Tensions between compliance, internal controls and ethics in the domain of project governance," *International Journal of Managing Projects in Business*, vol. 13, no. 4, p. 845–865, 2020. https://dx.doi.org/10.1108/IJMPB-07-2019-0171

[14] B. Kuehnert, R. Kim, J. Forlizzi, and H. Heidari, "The "who", "what", and "how" of responsible AI governance: A systematic review and meta-analysis of (actor, stage)-specific tools," p. 2991–3005, 2025. https://dx.doi.org/10.1145/3715275.3732191

[15] T. M. Jones, "Ethical decision making by individuals in organizations: An issue-contingent model," *Academy of Management Review*, vol. 16, no. 2, p. 366–395, 1991. https://doi.org/10.5465/amr.1991.4278958

[16] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *arXiv preprint arXiv:1906.02243*, 2019. https://doi.org/10.48550/arXiv.1906.02243

[17] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, p. 501–507, 2019. https://doi.org/10.1038/s42256-019-0114-4

[18] 119th Congress (2024-2025), "Committee print: Providing for reconciliation pursuant to h. con. res. 14, the concurrent resolution on the budget for fiscal year 2025," 2025.

[19] N. DePaula, L. Gao, S. Mellouli, L. F. Luna-Reyes, and T. M. Harrison, "Regulating the machine: An exploratory study of us state legislations addressing artificial intelligence, 2019-2023," p. 815–826, 2024. https://dx.doi.org/10.1145/3657054.3657148

[20] A. Hopkins and S. Booth, "Machine learning practices outside big tech: How resource constraints challenge responsible development," in *AIES 2021: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2021, Conference Proceedings, p. 134–145. https://dx.doi.org/10.1145/3461702.3462527

[21] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, vol. 3, no. 4, pp. 305–360, 1976. https://doi.org/10.1016/0304-405X(76)90026-X

[22] R. Derakhshan, R. Turner, and M. Mancini, "Project governance and stakeholders: a literature review," *International Journal of Project Management*, vol. 37, no. 1, pp. 98–116, 2019. https://doi.org/10.1016/j.ijproman.2018.10.007

[23] International Organization for Standardization, *Project, programme and portfolio management—Guidance on project management (ISO Standard No. 21502:2020-12)*. ISO, 2020.

[24] T. Guggenberger, L. Lämmermann, N. Urbach, A. Walter, and P. Hofmann, "Task delegation from AI to humans: A principal-agent perspective," in *ICIS 2023 Convention, Hyderabad, India*, 2023, Conference Paper.

[25] B. Wachnik, "Moral hazard in it project completion. a multiple case study analysis," pp. 1557–1562, 2015. http://dx.doi.org/10.15439/2015F68

[26] T. Sheridan, W. Verplank, and T. L. Brooks, "Human and computer control of undersea teleoperators," 1978.

[27] J. F. J. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis Seventh Edition*. Pearson College Division, 2014.

[28] B. E. Weller, N. K. Bowen, and S. J. Faubert, "Latent class analysis: A guide to best practice," *J. Black Psychol.*, vol. 46, no. 4, p. 287–311, 2020. https://doi.org/10.1177/0095798420930932

[29] SurveyMonkey, "SurveyMonkey answers to the ESOMAR questions to help buyers of online samples," 2024. [Online]. Available: https://www.surveymonkey.com/mp/legal/esomar-37/

[30] F. Bentley, K. O'Neill, K. Quehl, and D. Lottridge, "Exploring the quality, efficiency, and representative nature of responses across multiple survey panels," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, Conference Proceedings, pp. 1–12. https://dx.doi.org/10.1145/3313831.3376671

[31] A. Gordon and K. Thorson, *Dealing with Repeated Measures: Design Decisions and Analytic Strategies for Over-Time Data**. Cambridge University Press, 2024, pp. 532–564. https://dx.doi.org/10.1017/9781009170123.023

[32] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies." *J. Appl. Psychol.*, vol. 88, no. 5, pp. 879–903, 2003. https://dx.doi.org/10.1037/0021-9010.88.5.879

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 219–230
DOI: 10.15439/2025F1666
ISSN 2300-5963 ACSIS, Vol. 43

Deep Differentiable Logic Gate Networks Based on Fuzzy Łukasiewicz T-norm

Chan Duong Nguy
0009-0006-1339-9841
Vistula University
3 Stokłosy, 02-787, Warsaw, Poland
Email: chanduong1999@gmail.com

Piotr Wasilewski
0000-0003-0027-1102
Systems Research Institute,
Polish Academy of Sciences
6 Newelska, 01-447 Warsaw, Poland
Faculty of Computer Science,
Dalhousie University
6050 University Avenue, Halifax,
Nova Scotia, Canada
Email: pwasilew@ibspan.waw.pl

***Abstract*—Differentiable Logic Gate Networks (DLNs) offer a compelling framework for symbolic interpretability and reducing inference cost. Building on prior works using Menger and Zadeh T-norms, we investigate the Łukasiewicz T-norm as an alternative relaxation for classical logic gates. While it provides strong gradients in some regions, its flat areas result in vanishing gradients that hinder training. To address this issue, we use an initialization strategy that is analogous to residual connection in Neural Networks to encourage error signal propagation during training. Our empirical results show that Łukasiewicz based DLNs, though slightly less accurate, benefit from faster inference and lower memory requirements compared to Neural Networks, giving the opportunity of practical application in, e.g., resource constrained devices. Due to the structural clarity, DLNs facilitate direct inspection and tracing of information flow, which makes them suitable for application in explainable artificial intelligence (XAI).**

I. Introduction

DIFFERENTIABLE logic gate networks (DLNs) provide a promising framework for combining the interpretability of classical logic circuits with the scalability of gradient-based learning. Prior work has demonstrated that continuous relaxations of binary logic gates, such as those based on the Zadeh and Menger T-norms enable the use of standard backpropagation techniques while allowing conversion back to classical logic post training [19], [28], [15], [29]. Building on this foundation, we extend the study of T-norm-based relaxations by investigating the Łukasiewicz T-norm as a candidate operator for differentiable logic networks.

In production environments, inference efficiency directly affects user experience, operational costs, and energy consumption, making it a critical optimization target, especially for edge devices and real time systems [2], [21] with strict latency and resource constraints. Common strategies such as reduced-precision computation [3], [7], binary networks [22], and sparsity exploitation [8], [16] offer partial solutions but often come with trade-offs in accuracy or generality. DLNs offer an alternative path, where logic gate networks were trained using a continuous relaxation based on the fuzzy logic functions to enable gradient descent. Each computational unit in DLNs learns a probability distribution over 16 of these differentiable fuzzy logic functions using softmax. Post training, these units are assigned the most probable gate, resulting in networks that are naturally sparse and require no weights, leading to exceptional inference speeds. In our previous work [28], we followed this approach using the Zadeh T-norm for continuous relaxation. Although it is not differentiable at $a = b$, we assume the output to be a in those cases to preserve gradient flow. In this paper, we extend this line of research by exploring the Łukasiewicz T-norm as a relaxation operator for logic gate networks. Although similar to the Zadeh T-norm, the Łukasiewicz operator provides strong gradients but only in certain regions; however, it suffers from wide, flat areas with vanishing gradients. To address this, we use initialization strategies [20] that bias the network toward gate configurations that better propagate gradients, especially during the early stages of training.

Recent research on medical applications delving into large-data processing such as MRI [27], [14], EEG [6], [11] data, X-ray images [4] and clinical notes [5] which requires high computational resources. The introduction of DLNs open a new opportunity for processing relatively small data with low latency and small memory footprint e.g. pulsimeter, oxygen blood saturation measuring and non-invasive glucose monitoring.

II. Differentiable Logic Gate Networks

Hardware implementations utilizing logic gates exhibit exceptional performance characteristics nanosecond execution speeds, true parallelism, and deterministic reliability. However, designing complex systems at the gate level presents formidable challenges. Development with Verilog or VHDL lacks convenient debugging tools, requires specialized equipment for verification, and involves lengthy development cycles [23]. These constraints necessitate novel approaches such as symbolic learning that can select optimal logic gate configurations, dramatically reducing design cycles. On the other hand, binary logic gates are inherently discontinuous, making

gradient based training infeasible. Inspired by this, Differentiable Logic Gate Networks was introduced by [19] and further investigated by [28], offering an elegant solution to this challenge by incorporating principles from Differentiable Fuzzy Logics (DFL) [10].

Instead of operating on discrete values $\{0, 1\}$, DLNs allow inputs and outputs to take continuous values in the range $[0, 1]$. This enables smooth transitions between logical states and creates differentiable pathways through which gradients can flow during training.

Architecturally, DLNs differ from standard neural networks in their connection patterns. Each computational unit (analogous to a neuron) processes a pair of randomly assigned inputs, rather than computing a weighted sum across all inputs. This design results in inherently sparse connections, in contrast to the dense layers typical in traditional neural networks.

The core learning mechanism of DLNs involves dynamically selecting the most appropriate fuzzy logic operation at each unit to minimize the overall loss function. Unlike neural networks, which learn weights, DLNs learn which logical operations best model the underlying patterns in the data.

A Boolean operation can be defined as a function $f : \{0, 1\} \times \{0, 1\} \to \{0, 1\}$, leading to a total of 16 possible Boolean operations (Table I). Each of these can be extended to fuzzy logic. For example, the Boolean AND ($\wedge$) operation can be defined as:

$$a \wedge b = \begin{cases} 1 & \text{if } a = b = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The fuzzy logic equivalent to Boolean AND using Menger's T-norm [15] is:

$$T_M(a, b) = a \cdot b.$$

To represent the fuzzy logic operation probabilistically, each computational unit can be represented as:

$$r = \sum_{i=0}^{15} p_i \cdot f_i(a, b), \quad \text{for } i = 0, 1, \ldots, 15,$$

where the probabilities p_i are derived from a softmax over a learnable logit vector $\mathbf{w}$:

$$p_i = \frac{e^{w_i}}{\sum_{j=0}^{15} e^{w_j}}, \quad \text{for } i = 0, 1, \ldots, 15.$$

Each f_i denotes a fuzzy logic function corresponding to one of the 16 Boolean operations, table I.

Each layer in a DLN consists of several such units, and their outputs propagate forward to support complex decision-making processes. The choice of T-norm is a crucial hyperparameter, influencing both accuracy and inference speed, which we explore in the following subsections.

A. Menger Based Differentiable Logic Gate Networks

The original DLN formulation [19] used Menger's T-norm ($T_M(a, b) = a \cdot b$), also known as the probabilistic T-norm. This choice is intuitive due to its smooth and easily computable derivatives:

$$\frac{\partial T_M(a, b)}{\partial a} = b, \quad \frac{\partial T_M(a, b)}{\partial b} = a,$$

which are well-suited for gradient based optimization algorithms [24], [10].

Empirical results show that models based on the Menger T-norm achieve competitive accuracy compared to neural networks on many benchmarks while offering significantly lower memory usage and computational complexity due to their reliance on binary logic operations. Compared to other T-norms [28], Menger based DLNs (DLNs_M) demonstrate robustness to noise, albeit with slightly slower inference times. However, the DLN_M variant suffers from the vanishing gradient problem, as the partial derivatives diminish when $\partial T_M(a, b)/\partial a < 1$, causing the backpropagated error signal to attenuate and hindering effective learning in deeper networks. To address this [19] proposed scaling the gradients by a factor $f > 1$ for DLNs with more than six layers. While this technique mitigates the vanishing gradient problem, it also introduces the risk of exploding gradients, potentially destabilizing training.

B. Zadeh Based Differentiable Logic Gate Networks

To mitigate gradient vanishing, [28] proposed replacing Menger's T-norm with Zadeh's T-norm [29], defined as $T_Z(a, b) = \min(a, b)$. While the forward computation is straightforward, care must be taken during backpropagation to ensure consistent gradient flow, especially when $a = b$, where the gradient could flow to either input.

To address this, the authors [28] chose to default to propagating the gradient through a in the case of equality (although choosing b would yield the same results). The partial derivatives are defined as:

$$\frac{\partial T_Z(a, b)}{\partial a} = \begin{cases} 1 & \text{if } a \leq b, \\ 0 & \text{otherwise,} \end{cases} \quad \frac{\partial T_Z(a, b)}{\partial b} = \begin{cases} 0 & \text{if } a \leq b, \\ 1 & \text{otherwise.} \end{cases}$$

This setup ensures a well-defined and consistent backward pass. One key advantage of Zadeh's T-norm is that the derivatives of all fuzzy operations are limited to the set $\{-1, 0, 1\}$, and it is never simultaneously zero for both a and b. This guarantees that the gradient is neither vanishing nor exploding, enabling more stable and effective training.

Consequently, Zadeh based DLNs (DLNs_Z) are good at recognizing discrete, singular features and outperform DLNs_M in inference speed, albeit with a slight trade off in accuracy on some benchmarks [28].

C. Binary Logic Gate Networks

Compared to traditional neural networks, the inference process in Differentiable Logic Networks can be significantly less efficient due to the need to compute all 16 fuzzy logic operations for every computational unit. In contrast, neural networks rely heavily on matrix multiplication as their computational backbone, a process that is now widely accelerated

by modern hardware, such as Nvidia's Tensor Cores, among others.

One potential optimization is to reduce the number of fuzzy logic operations computed by selecting only the logic gate with the highest probability for each unit. Since fuzzy operations generalize Boolean operations, when inputs are strictly binary (i.e., 0 or 1), the resulting outputs are the same as the binary connectives equivalent; converting from smooth fuzzy operators to crisp binary ones theoretically does not compromise accuracy, especially when the training data is already binary. This transition could also substantially reduce memory consumption and computational complexity due to the efficiency of binary operations.

To implement this optimization, DLNs can be converted into Binary Logic Gate Networks (LGNs) by choosing the fuzzy operation where the softmax function results in the highest probability; thus, this retains the same architectural structure, namely, the number of layers and computational units; however, each unit in LGNs only computes a single binary logic gate instead of a weighted sum of all 16 functions.

III. Łukasiewicz based Differentiable Logic Gate Networks

In this study, we investigate fuzzy logic operations based on the Łukasiewicz T-norm, defined as $T_Ł(a,b) = \max\{0, a+b-1\}$. Like the Zadeh T-norm used in [28], the Łukasiewicz T-norm relies on $\min$ and $\max$ operators. Consequently, it exhibits similar challenges during backpropagation, particularly regarding non smooth gradients and ambiguous gradient flow paths.

To address this, we adopt the same backward pass strategy proposed in [28], which resolves the gradient ambiguity by enforcing consistent choices. The partial derivatives with respect to the inputs are defined as:

$$\frac{\partial T_Ł(a,b)}{\partial a} = \begin{cases} 0 & \text{if } a+b-1 \leq 0, \\ 1 & \text{otherwise,} \end{cases}$$

$$\frac{\partial T_Ł(a,b)}{\partial b} = \begin{cases} 0 & \text{if } a+b-1 \leq 0, \\ 1 & \text{otherwise.} \end{cases}$$

This yields sharp gradients in the set $\{-1, 0, 1\}$ for the network, which can help limit exploding gradients. However, a significant portion of the domain of $T_Ł$ results in zero gradients, leading to the risk of vanishing gradients during training.

To mitigate this, we implement a residual initialization scheme inspired by [20]. Instead of initializing the learnable weights uniformly at random, the authors bias them toward logic gates that act as pass through operators. Owing to the symmetry between inputs, they treat gates corresponding to A and B (i.e., the 3rd and 5th gates in Table I) as functionally equivalent in this context. We experiment with two variants of this initialization: one where gate A is assigned a 90% initial probability (with all other gates initialized to 0.67%) as in [20], and in this work we introduce another variant where both gates A and B are initialized with 45% probability each.

For inference, we transform the computational units of Differentiable Logic Networks into hard logic gates and generate corresponding C code for the network. This C code is compiled using the `-O1` optimization flag for models with fewer than 50,000 computational units and `-O0` for larger models due to limitations in compilation time and memory overhead.

Modern 64-bit CPUs are ubiquitous across consumer devices, which implies that Boolean values must typically be extended via zero extension or sign extension to align with the ALU's expected input width. The implementation proposed by [19] uses a 64-bit encoding for Boolean inputs to match the native word size of contemporary CPUs. For consistency, our initial implementation of Zadeh-based Differentiable Logic Networks (DLNs) [28] followed the same encoding strategy. However, empirical evaluation with our available devices for experiments revealed that using an 8-bit encoding for Boolean values significantly reduces inference time. This change in using different numbers of bits to represent Boolean values does not affect the accuracy of the models. Therefore, in order to maximize performance under our setup, we proceed with using 8-bit encodings for all classical logic operations in the networks instead of 64-bit.

IV. Experiments and Results

To ensure a fair comparison with previous work [19], [28], we evaluate our proposed approach on both structured and unstructured data types. For structured datasets, categorical attributes are one-hot encoded, while continuous features such as age or tumor size are discretized and then followed by one-hot encoding. This preprocessing ensures compatibility with post training discretized models that require binary or discrete input representations. For unstructured data such as CIFAR-10 [12], we adopt a similar discretization strategy as [19], [28] for pixel intensities by applying a series of progressive thresholds, enabling a finer grained binary encoding of continuous values. This technique reduces information loss during binarization and supports more expressive representations in logic based models. We created two variants of the CIFAR-10 dataset for 3 thresholds and 31 thresholds, which we will discuss in more detail in CIFAR-10. In contrast, for the MNIST dataset, we retain the original grayscale pixel values and train directly on the real valued inputs to evaluate DLNs' performance on real values data.

During evaluation, all DLNs, regardless of implementation, are converted into Binary Logic Gate Networks where the computational unit calculates fixed Boolean functions with the highest probability. All the variants of $\text{DLN}_Ł$ are denoted as $\text{DLN}_Ł$, $\text{DLN}_{Ł^*}$, $\text{DLN}_{Ł^{**}}$ for without residual initialization, residual initialization at the feedforward gate A, and residual initialization at gates A and B. All models are implemented using PyTorch [18], Adam optimizer [9] and trained on an NVIDIA RTX 2000 Ada GPU. Inference time is measured using a single threaded setup on an AMD Ryzen 5 3550H CPU and averaged across 100 runs. To indicate statistically significant differences in inference time across models, we

TABLE I
LIST OF ALL ŁUKASIEWICZ T-NORM LOGIC GATES.

ID	Operator	Łukasiewicz T-norm	00	01	10	11
0	$False$	0	0	0	0	0
1	$a \wedge b$	$max\{0, a+b-1\}$	0	0	0	1
2	$\neg(a \Rightarrow b)$	$max\{0, a-b\}$	0	0	1	0
3	a	a	0	0	1	1
4	$\neg(a \Leftarrow b)$	$max\{0, -a+b\}$	0	1	0	0
5	b	b	0	1	0	1
6	$\neg(a \Leftrightarrow b)$	$min\{1, max\{0, a-b\} + max\{0, -a+b\}\}$	0	1	1	0
7	$a \vee b$	$min\{1, a+b\}$	0	1	1	1
8	$\neg(a \vee b)$	$max\{0, 1-a-b\}$	1	0	0	0
9	$a \Leftrightarrow b$	$max\{0, min\{1, 1-a+b\} + min\{1, 1+a-b\} - 1\}$	1	0	0	1
10	$\neg b$	$1-b$	1	0	1	0
11	$a \Leftarrow b$	$min\{1, 1+a-b\}$	1	0	1	1
12	$\neg a$	$1-a$	1	1	0	0
13	$a \Rightarrow b$	$min\{1, 1-a+b\}$	1	1	0	1
14	$\neg(a \wedge b)$	$min\{1, 2-a-b\}$	1	1	1	0
15	$True$	1	1	1	1	1

TABLE II
RESULTS ON THE MONK DATA SETS AVERAGED OVER 100 RUNS.

Method	MONK-1	MONK-2	MONK-3
ANN	100.00%	100.00%	93.5%
DLN_M	100.00%	78.24%	95.37%
DLN_Z	100.00%	89.32%	91.74%
$DLN_Ł$	100.00%	56.94%	93.98%
$DLN_{Ł^*}$	100.00%	61.57%	78.24%
$DLN_{Ł^{**}}$	100.00%	49.77%	69.44%
	# Parameters	Inf. Time	Memory
ANN	162	161.20 ± 26.50ns	648B
DLN_M	144 | 72 | 72	8.06 ± 0.17ns	72B | 36B | 36B
DLN_Z	144 | 72 | 72	7.77 ± 0.13ns	72B | 36B | 36B
$DLN_Ł$	144 | 72 | 72	7.20 ± 0.12ns	72B | 36B | 36B
$DLN_{Ł^*}$	144 | 72 | 72	6.83 ± 0.16ns	72B | 36B | 36B
$DLN_{Ł^{**}}$	144 | 72 | 72	6.78 ± 0.15ns	72B | 36B | 36B

denote: $p \in [0.001, 0.5)$ by *, $p \in [0.0001, 0.001)$ by **, and $p \in [0, 0.0001)$ by ***.

A. MONK

The MONK datasets [26] constitute a benchmark suite of binary classification tasks designed to evaluate the performance of machine learning models on symbolic and logical reasoning problems. Each dataset consists of discrete-valued features and a binary target label, with the classification rules defined using logical expressions.

MONK-1 defines a concept in disjunctive normal form (DNF), making it relatively simple and well suited for symbolic learners such as Differentiable Logic Gate Networks. **MONK-2** resembles a parity problem, where the target function is based on a combination of attributes that cannot be easily expressed in either DNF or conjunctive normal form (CNF). **MONK-3** is structurally similar to MONK-1 and is also defined using DNF. However, it introduces label noise into the training set, making it a useful benchmark for evaluating a model's robustness and generalization under imperfect data.

The ANN achieves the highest accuracy on the MONK-1 and MONK-2 datasets, but it requires significantly more memory than DLNs, regardless of the implementation. DLN_M outperforms all other models on the MONK-3 dataset. Interestingly, $DLN_Ł$ achieves a higher accuracy than the ANN (94.0% vs. 93.5%) in the last dataset (see table II), results of significant difference tests are presented in table III and basic statistical analysis in figure 1. Adding residual initializations to $DLN_Ł$ yields mixed results: accuracy improves on MONK-2 but decreases on the other datasets. In terms of inference time, the $DLN_{Ł^*}$ $DLN_{Ł^{**}}$ methods show similar improvement in performance.

TABLE III
MONK DATA SET SIGNIFICANT INFERENCE TIME DIFFERENCE TESTS. WE DENOTE: $p \in [0.001, 0.5)$ BY *, $p \in [0.0001, 0.001)$ BY **, AND $p \in [0, 0.0001)$ BY ***

	DLN_M	DLN_Z	$DLN_Ł$	$DLN_{Ł^*}$	$DLN_{Ł^{**}}$
DLN_M			***	***	***
DLN_Z			*	***	***
$DLN_Ł$	***	*			*
$DLN_{Ł^*}$	***	***			
$DLN_{Ł^{**}}$	***	***	*		

B. Adult dataset

The Adult dataset [1], also known as the "Census Income" dataset, is a widely used benchmark in the machine learning community. It originates from the UCI Machine Learning Repository and involves a binary classification task aimed at predicting whether an individual's annual income exceeds $50,000. The dataset comprises 48,842 samples and includes both continuous and categorical features, such as age, education, occupation, and work class.

TABLE IV
RESULTS FOR THE ADULT DATA SET AVERAGED OVER 100 RUNS.

Adult	Acc.	#Param.	Infer. Time	Memory
ANN	84.90%	3,810	287.95 ± 6.19ns	$15KB$
DLN_M	84.83%	1,280	59.47 ± 2.34ns	640B
DLN_Z	75.40%	1,280	71.12 ± 0.45ns	640B
$DLN_Ł$	82.62%	1,280	94.98 ± 0.88ns	640B
$DLN_{Ł^*}$	84.57%	1,280	59.99 ± 2.02ns	640B
$DLN_{Ł^{**}}$	84.59%	1,280	81.27 ± 1.91ns	640B

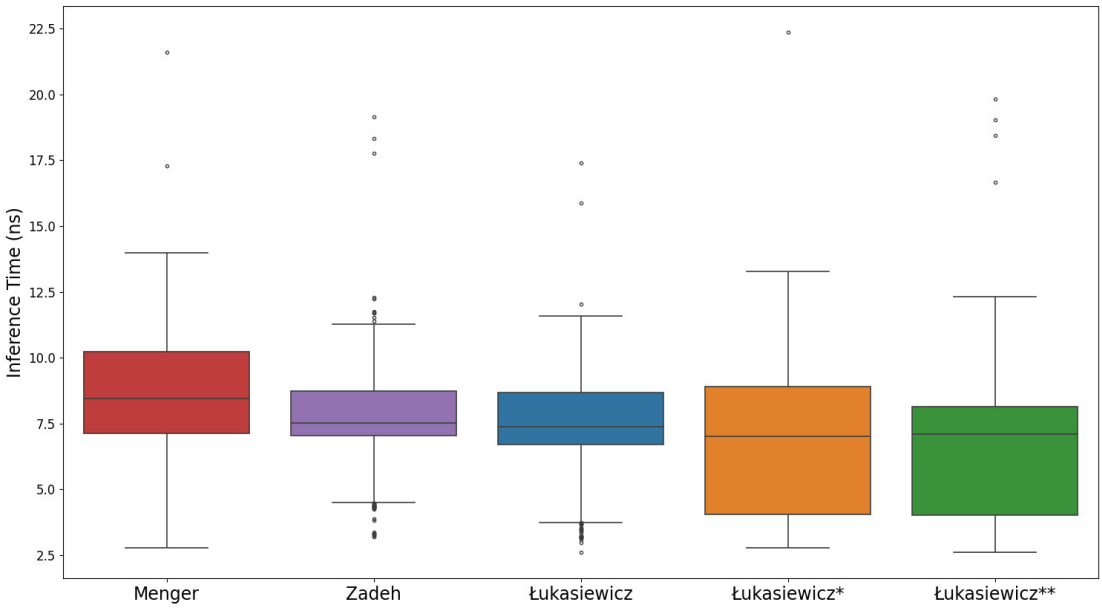


Fig. 1. Basic statistical analysis for MONK dataset inference time (ns) results (lower is better)

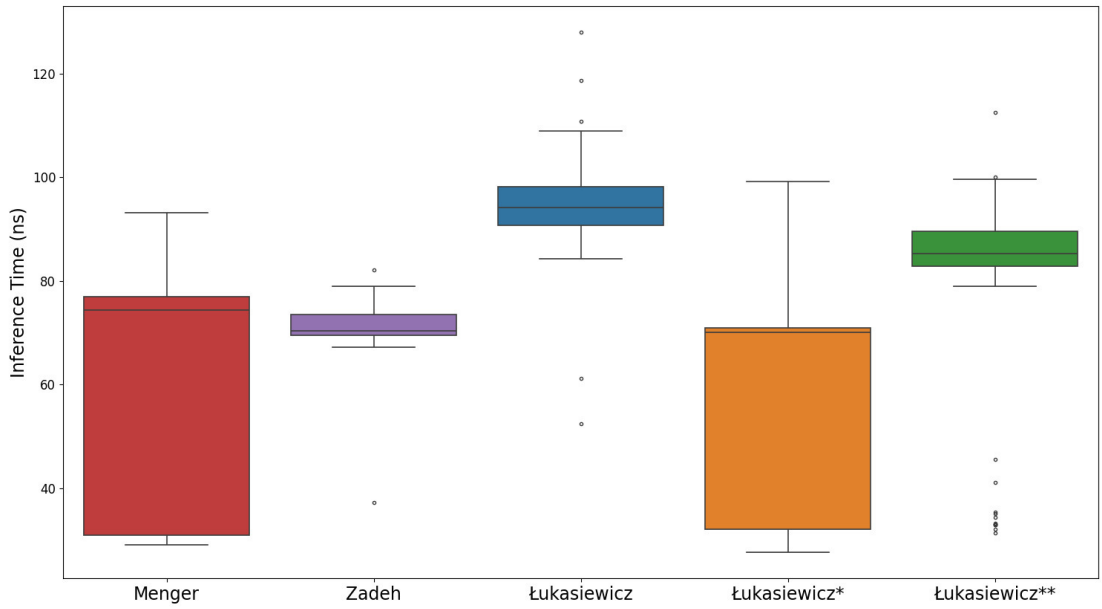


Fig. 2. Basic statistical analysis for Adult dataset inference time (ns) results (lower is better)

TABLE V
ADULT DATA SET SIGNIFICANT INFERENCE TIME DIFFERENCE TESTS. WE DENOTE: $p \in [0.001, 0.5)$ BY *, $p \in [0.0001, 0.001)$ BY **, AND $p \in [0, 0.0001)$

	DLN_M	DLN_Z	$DLN_Ł$	$DLN_{Ł^*}$	$DLN_{Ł^{**}}$
DLN_M		***	***		***
DLN_Z	***		***	***	***
$DLN_Ł$	***	***		***	***
$DLN_{Ł^*}$		***	***		***
$DLN_{Ł^{**}}$	***	***	***	***	

TABLE VII
BREAST CANCER DATA SET SIGNIFICANT INFERENCE TIME DIFFERENCE TESTS. WE DENOTE: $p \in [0.001, 0.5)$ BY *, $p \in [0.0001, 0.001)$ BY **, AND $p \in [0, 0.0001)$

	DLN_M	DLN_Z	$DLN_Ł$	$DLN_{Ł^*}$	$DLN_{Ł^{**}}$
DLN_M		***	***	***	*
DLN_Z	***		***	**	
$DLN_Ł$	***	***		*	***
$DLN_{Ł^*}$	***	**	*		*
$DLN_{Ł^{**}}$	*		***	*	

In this benchmark, DLN_M, $DLN_{Ł^*}$, and $DLN_{Ł^{**}}$ demonstrate comparable accuracy to the artificial neural network (ANN), but with much lower inference time and memory usage (table IV), inference time significant difference tests are in table V and see figure 2 for basic statistical analysis. Among all implementations, DLN_M is the fastest in terms of inference speed and similar accuracy compared to the baseline ANN. Interestingly, when we initialize the Łukasiewicz models with higher probabilities at gates A and B, the accuracy doesn't improve, but inference time goes down. This shows $DLN_{Ł^*}$ a good compromise between speed and performance.

C. Breast Cancer dataset

The Breast Cancer dataset [30] was collected from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. It is one of three domains provided by the Oncology Institute that have frequently appeared in the machine learning literature.

The dataset consists of 286 instances, with 201 belonging to one class and 85 to another. Each instance is described by nine attributes. The dataset is commonly used to assess classification performance on medical diagnostic tasks, particularly in scenarios involving imbalanced class distributions and heterogeneous feature types.

TABLE VI
RESULTS FOR THE BREAST CANCER DATA SET AVERAGED OVER 100 RUNS.

Breast Cancer	Acc.	#Param.	Infer. Time	Memory
ANN	75.31%	434	792.43 ± 100.23ns	1.4KB
DLN_M	71.43%	640	38.34 ± 0.27ns	320B
DLN_Z	68.56%	640	35.14 ± 0.22ns	320B
$DLN_Ł$	70.00%	640	26.72 ± 1.07ns	320B
$DLN_{Ł^*}$	74.29%	640	30.62 ± 1.25ns	320B
$DLN_{Ł^{**}}$	70.00%	640	35.58 ± 1.10ns	320B

In this experiment, table VI shows ANN reaches an accuracy of 75.3%, followed closely by $DLN_{Ł^*}$ at 74.3%, while using only a fraction of the inference time and memory. $DLN_Ł$ stands out as the fastest model, although it has very low accuracy. The results of significant difference tests for inference speed are presented in table VII, and essential statistical analysis is presented in figure 3. Notably, using residual initialization brings a major boost for $DLN_Ł$, raising its accuracy from 70.0% to 74.3%. However, when both gates A and B are initialized together, the model's accuracy actually drops. Once again, $DLN_{Ł^*}$ has a good balance between accuracy and run time complexity.

D. MNIST Dataset

To ensure comparability with prior studies [19], [28], we evaluate our proposed approach on the MNIST dataset [13]. MNIST is a standard benchmark in computer vision and machine learning, consisting of grayscale images of handwritten digits ranging from 0 to 9. The dataset comprises 70,000 samples, partitioned into 60,000 training images and 10,000 test images, each originally sized at 28×28 pixels.

To examine the models' adaptability to reduced input dimensionality, we additionally resize the images to 20×20 pixels by removing the borders, thereby constructing a secondary, lower-resolution version of the dataset. Each image is associated with a label corresponding to the digit it represents, making this a ten class classification problem. The dataset serves as a robust testbed for evaluating both accuracy and generalization across a range of model architectures.

In both the small and normal configurations, the results in table VIII show ANNs outperform DLNs in terms of accuracy across all implementations (ANN *(small)*: 96.3%, ANN: 98.40%), though this comes at the cost of higher inference time and memory usage. Among the smaller DLN models, DLN_M *(small)* achieves the highest accuracy, but it is the slowest ($1.127 \pm 0.03\mu$s) see table IX and figure 4 for inference speed significant difference tests and statistical analysis. On the other end, $DLN_Ł$ *(small)* offers the fastest inference but with the lowest accuracy. $DLN_{Ł^{**}}$ *(small)* yields higher accuracy than $DLN_{Ł^*}$ *(small)* and on average, it also has a faster inference time; however, this is not a statistically significant increase. Both initialization strategies show clear improvements in accuracy compared to the uninitialized base model.

It's also worth pointing out that DLN_M and $DLN_{Ł^*}$ have comparable accuracy (97.6% vs. 97.3%), but $DLN_{Ł^*}$ has the advantage in terms of faster inference.

E. CIFAR-10 Dataset

CIFAR-10 [12] is a well established benchmark dataset composed of 60,000 color images, each with a resolution of 32×32 pixels. The images span 10 distinct object categories, including airplanes, automobiles, birds, cats, and more. The dataset is partitioned into 50,000 training samples and 10,000 test samples. Its compact image size makes it particularly suitable for experimentation under limited computational resources, facilitating the development and evaluation of novel algorithms.

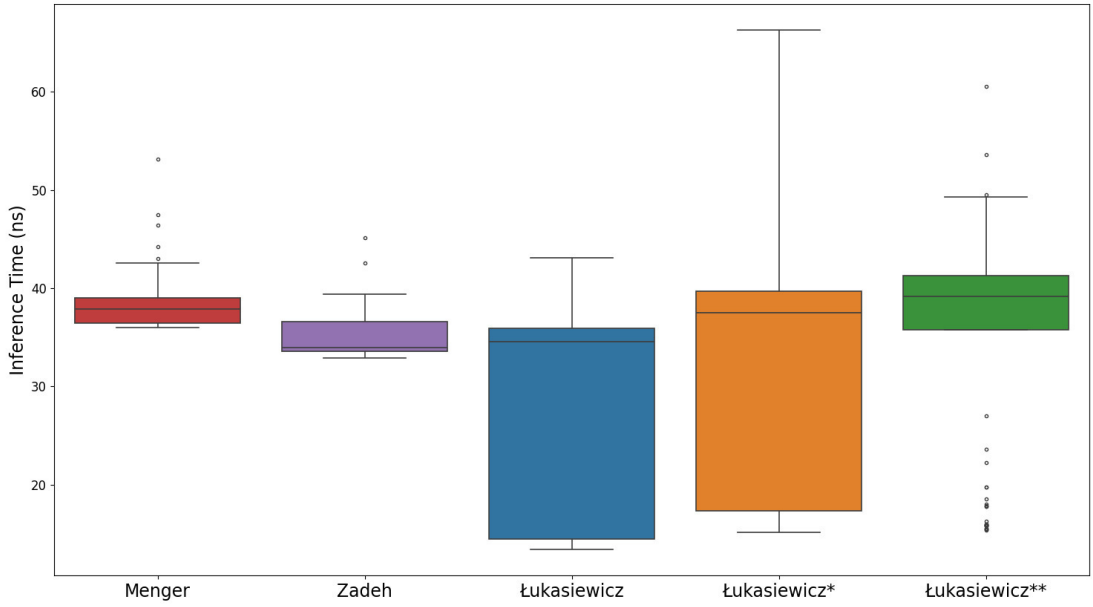


Fig. 3. Basic statistical analysis for Breast Cancer dataset inference time (ns) results (lower is better)

TABLE VIII
RESULTS FOR THE MNIST DATA SET AVERAGED OVER 100 RUNS, WHERE OPS STAND FOR NUMBER OF BINARY OPERATIONS NEEDED TO PERFORM COMPUTATIONS.

MNIST	Acc.	#Param.	Memory	Infer. Time	OPs
ANN (*small*)	97.92%	118282	462KB	$3.66 \pm 0.15\mu$s	236M
DLN_M (*small*)	96.34%	48000	23KB	$1.13 \pm 0.03\mu$s	48K
DLN_Z (*small*)	94.52%	48000	23KB	$1.10 \pm 0.03\mu$s	48K
$\text{DLN}_Ł$ (*small*)	93.93%	48000	23KB	$1.25 \pm 0.30\mu$s	48K
$\text{DLN}_{Ł^*}$ (*small*)	95.55%	48000	23KB	$1.09 \pm 0.03\mu$s	48K
$\text{DLN}_{Ł^{**}}$ (*small*)	96.31%	48000	23KB	$1.05 \pm 0.03\mu$s	48K
ANN	98.40%	22609930	86MB	$1009.17 \pm 9.23\mu$s	45G
DLN_M	97.61%	384000	188KB	$83.16 \pm 0.09\mu$s	384K
DLN_Z	93.33%	384000	188KB	$79.12 \pm 0.06\mu$s	384K
$\text{DLN}_Ł$	95.05%	384000	188KB	$79.63 \pm 0.11\mu$s	384K
$\text{DLN}_{Ł^*}$	97.36%	384000	188KB	$78.58 \pm 0.13\mu$s	384K
$\text{DLN}_{Ł^{**}}$	97.40%	384000	188KB	$78.49 \pm 0.10\mu$s	384K

TABLE IX
MNIST DATA SET SIGNIFICANT INFERENCE TIME DIFFERENCE TESTS. WE DENOTE: $p \in [0.001, 0.5)$ BY *, $p \in [0.0001, 0.001)$ BY **, AND $p \in [0, 0.0001)$

	DLN_M (*small*)	DLN_Z (*small*)	$\text{DLN}_Ł$ (*small*)	$\text{DLN}_{Ł^*}$ (*small*)	$\text{DLN}_{Ł^{**}}$ (*small*)
DLN_M (*small*)			*		
DLN_Z (*small*)			*		
$\text{DLN}_Ł$ (*small*)	*	*		**	***
$\text{DLN}_{Ł^*}$ (*small*)			**		
$\text{DLN}_{Ł^{**}}$ (*small*)			***		
	DLN_M	DLN_Z	$\text{DLN}_Ł$	$\text{DLN}_{Ł^*}$	$\text{DLN}_{Ł^{**}}$
DLN_M		***	***	***	***
DLN_Z	***		***	**	***
$\text{DLN}_Ł$	***	***		***	***
$\text{DLN}_{Ł^*}$	***	**	***		
$\text{DLN}_{Ł^{**}}$	***	***	***		

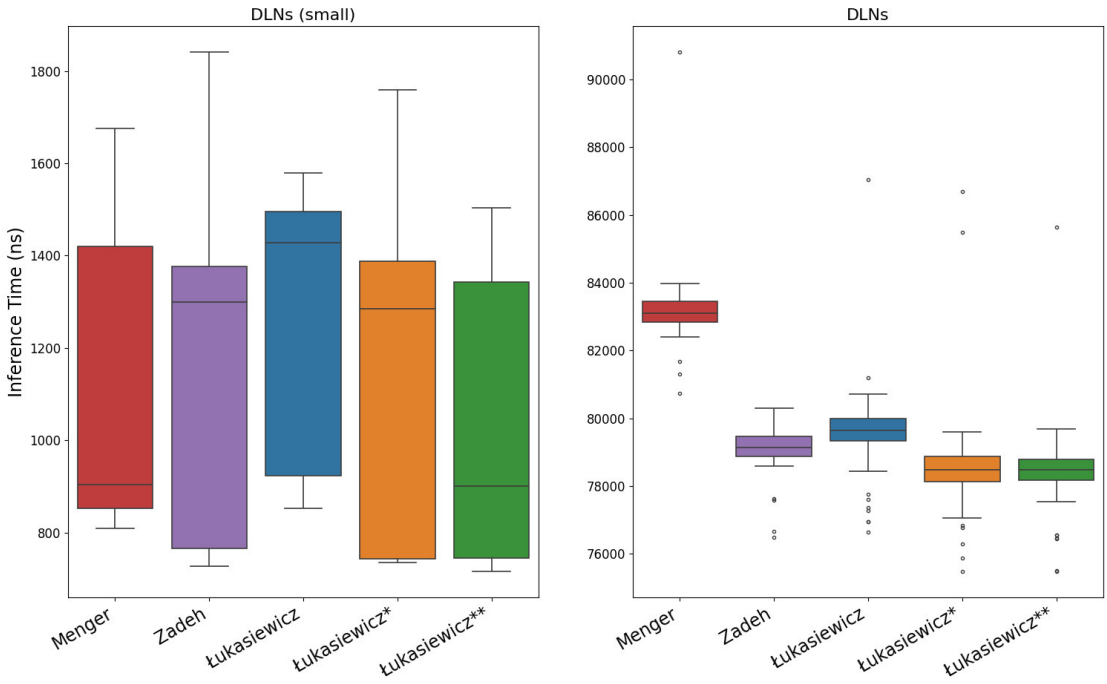


Fig. 4. Basic statistical analysis for MNIST dataset inference time (ns) results (lower is better)

Following the methodology proposed in [19], we reduce the color channel resolution to discrete threshold levels to accommodate binary input constraints of logic based models. Specifically, we apply three value thresholds for the *small* and *medium* models and 31 thresholds for the *large* models. Instead of a simple binarization using a 0.5 cutoff, which often results in significant information loss, we use a set of progressive thresholds (e.g., 0.25, 0.5, 0.75) to discretize pixel intensities more effectively. For instance, a pixel with intensity $p_i = 0.6$ can be represented as $p_i = [1, 1, 1, 0]$, where i denotes the pixel's position in the flattened image. This multi threshold encoding mitigates the discretization loss when converting fuzzy inputs into binary representations, thus improving the suitability of the data for binary logic gate models.

Although dropout and data augmentation are widely used to improve generalization in vision tasks, we intentionally refrain from employing these techniques to maintain consistency with our experimental framework. For all DLN implementations, we use the notation S, M, and L to represent small, medium, and large models, respectively, based on the number of parameters.

In the smaller model configurations, DLN implementations achieve lower accuracy compared to ANNs (see table X). Among the different DLN implementations, DLN_M consistently achieves the highest accuracy across all model sizes (S, M, L). The initialization methods show consistent improvement in accuracy over the base DLN_L model for all model sizes without significant impact on the inference time (see table XI and figure 5).

V. Discussion

The experimental results show consistently low accuracy for Łukasiewicz T-norm based models, which aligns with our initial hypothesis. We attribute this to a substantial portion of the domain in Łukasiewicz based fuzzy logic yielding zero gradients, which prevents error signals from backpropagating effectively to higher layers. Moreover, across all experiments, larger networks benefit significantly from both proposed initialization schemes, showing notable improvements over the baseline DLNs_L. This supports the idea that biasing the model toward feedforward gates while avoiding gates that contribute to gradient vanishing, particularly during the early stages of training when error rates are highest, can substantially enhance error propagation. Interestingly, these initialization methods also increase the models inference speed compared to the baseline models. While increasing the number of feedforward-biased gates does lead to higher accuracy, the improvement is modest and not consistently observed across all experiments. In smaller models, such as those trained on MONK-2, MONK-3, and the Breast Cancer dataset, biasing toward feedforward gates results in a decrease in accuracy performance. We hypothesize that when the number of model parameters is limited, the biasing restricts the model's capacity to flexibly select appropriate logic gates, thereby impairing its learning ability. However, in larger models, the inflexibility introduced by the initialization schemes results in measurable improvements in both accuracy and inference time. One

TABLE X
RESULTS FOR THE CIFAR-10 DATA SET AVERAGED OVER 100 RUNS, WHERE OPS STAND FOR NUMBER OF BINARY OPERATIONS NEEDED TO PERFORM COMPUTATIONS, S,M,L STAND FOR SMALL, MEDIUM AND LARGE NUMBER OF PARAMETERS IN THE MODEL RESPECTIVELY PRESENTED IN THE THIRD COLUMN.

CIFAR-10	Acc.	#Param.	Memory	Infer. Time	OPs
ANN	57.08%	12.6M	48MB	$519.38 \pm 6.70\mu$s	25G
DLN_M (S)	45.75%	48K	24KB	$1.75 \pm 0.00\mu$s	48K
DLN_Z (S)	45.45%	48K	24KB	$2.00 \pm 0.00\mu$s	48K
$\text{DLN}_Ł$ (S)	36.92%	48K	24KB	$1.63 \pm 0.00\mu$s	48K
$\text{DLN}_{Ł^*}$ (S)	41.14%	48K	24KB	$1.66 \pm 0.00\mu$s	48K
$\text{DLN}_{Ł^{**}}$ (S)	42.44%	48K	24KB	$1.88 \pm 0.00\mu$s	48K
DLN_M (M)	57.34%	512K	250KB	$167.19 \pm 0.14\mu$s	512K
DLN_Z (M)	55.55%	512K	250KB	$163.54 \pm 0.11\mu$s	512K
$\text{DLN}_Ł$ (M)	52.23%	512K	250KB	$158.81 \pm 0.25\mu$s	512K
$\text{DLN}_{Ł^*}$ (M)	54.88%	512K	250KB	$159.22 \pm 0.11\mu$s	512K
$\text{DLN}_{Ł^{**}}$ (M)	55.30%	512K	250KB	$159.59 \pm 0.18\mu$s	512K
DLN_M (L)	60.32%	1.28M	625KB	$399.08 \pm 0.34\mu$s	1.28M
DLN_Z (L)	57.28%	1.28M	625KB	$382.81 \pm 0.16\mu$s	1.28M
$\text{DLN}_Ł$ (L)	53.48%	1.28M	625KB	$377.30 \pm 0.21\mu$s	1.28M
$\text{DLN}_{Ł^*}$ (L)	57.78%	1.28M	625KB	$379.28 \pm 0.26\mu$s	1.28M
$\text{DLN}_{Ł^{**}}$ (L)	58.09%	1.28M	625KB	$377.20 \pm 0.23\mu$s	1.28M

TABLE XI
CIFAR-10 DATA SET SIGNIFICANT INFERENCE TIME DIFFERENCE TESTS. WE DENOTE: $p \in [0.001, 0.5)$ BY *, $p \in [0.0001, 0.001)$ BY **, AND $p \in [0, 0.0001)$

	DLN_M (S)	DLN_Z (S)	$\text{DLN}_Ł$ (S)	$\text{DLN}_{Ł^*}$ (S)	$\text{DLN}_{Ł^{**}}$ (S)
DLN_M (S)		***	***	***	***
DLN_Z (S)	***		***	***	***
$\text{DLN}_Ł$ (S)	***	***		***	***
$\text{DLN}_{Ł^*}$ (S)	***	***	***		***
$\text{DLN}_{Ł^{**}}$ (S)	***	***	***	***	
	DLN_M (M)	DLN_Z (M)	$\text{DLN}_Ł$ (M)	$\text{DLN}_{Ł^*}$ (M)	$\text{DLN}_{Ł^{**}}$ (M)
DLN_M (M)		***	***	***	***
DLN_Z (M)	***		***	***	***
$\text{DLN}_Ł$ (M)	***	***			*
$\text{DLN}_{Ł^*}$ (M)	***	***	*		
	DLN_M (L)	DLN_Z (L)	$\text{DLN}_Ł$ (L)	$\text{DLN}_{Ł^*}$ (L)	$\text{DLN}_{Ł^{**}}$ (L)
DLN_M (L)		***	***	***	***
DLN_Z (L)	***		***	***	***
$\text{DLN}_Ł$ (L)	***	***		***	
$\text{DLN}_{Ł^*}$ (L)	***	***	***		***
$\text{DLN}_{Ł^{**}}$ (L)	***	***		***	

possible explanation for the improved runtime efficiency is that a majority of the operations in these models are identity functions. At lower compiler optimization levels (e.g., `-O0`), identity operations are implemented as simple memory transfers between registers. However, under aggressive optimization settings (e.g., `-O3`), compilers often eliminate such operations entirely by reusing registers, resulting in minimal computational overhead. This behavior significantly reduces the runtime burden of inference in models where identity gates dominate. Figure 6 illustrates the gate selection distributions for $\text{DLNs}_Ł$ (Baseline), $\text{DLNs}_{Ł^*}$ (A), and $\text{DLNs}_{Ł^{**}}$ (A - B) on the MNIST dataset. A clear concentration of feedforward gates a for $\text{DLNs}_{Ł^*}$, and feedforward gates a and b for $\text{DLNs}_{Ł^{**}}$, which correspond to identity mappings, are evident in the latter two configurations. This suggests that the initialization schemes effectively steer the models toward simpler computational pathways. Interestingly, these identity heavy models not only benefit from faster inference but also demonstrate improved gradient stability, particularly during the early stages of training. The stable signal propagation enabled by identity operations likely contributes to the observed accuracy gains. However, the dominance of feedforward gates in $\text{DLNs}_{Ł^*}$ and $\text{DLNs}_{Ł^{**}}$ after training also implies underutilization of the network's expressive capacity. While this may be beneficial for efficiency, it raises questions about the models' ability to learn more complex representations when required.

Furthermore, we observe that increasing model size generally leads to an improvement in accuracy. This trend is consistent with the two findings from [19], [28] show that larger DLN models possess greater representational capacity. In current DLNs implementations, input pairs are connected randomly to computational units. This random connectivity hinders memory coalescing during data loading, which in turn leads to data starvation as computational kernels are forced to wait for data to be fetched. It is due to transferring memory from global memory (VRAM) to local memory (Registers) cost hundreds times more compute cycles comparing to the actually computations itself [17]. Replacing a convex combination of gates with a single, discrete gate can lead to significant performance degradation. For instance, if a computational unit

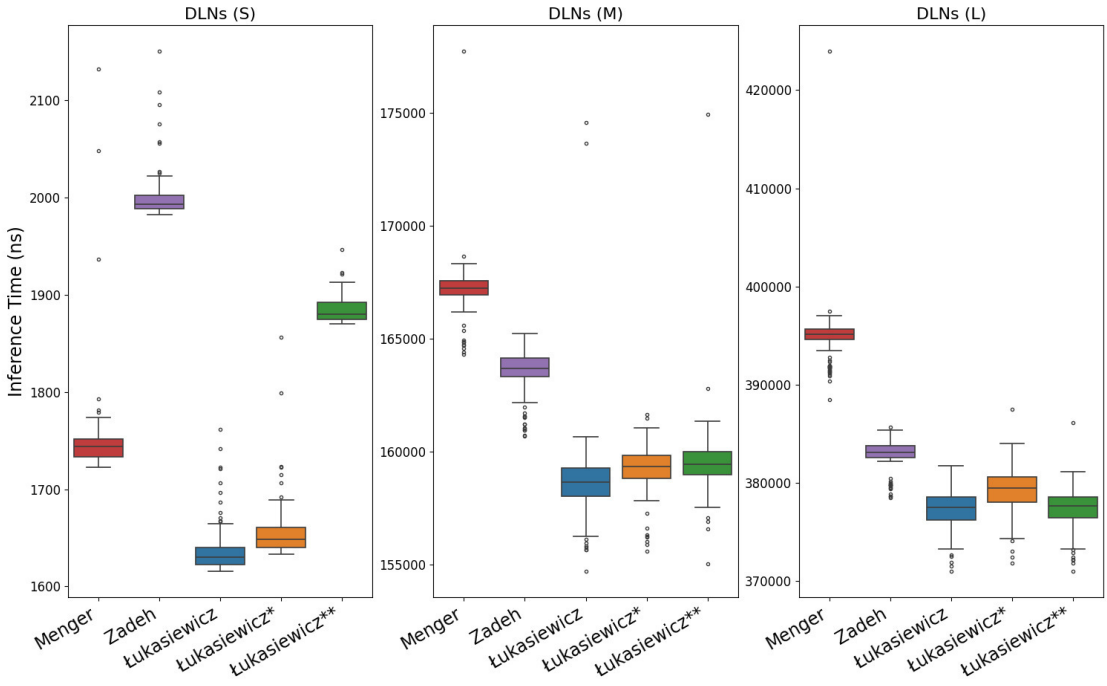


Fig. 5. Basic statistical analysis for CIFAR-10 dataset inference time (ns) results (lower is better)

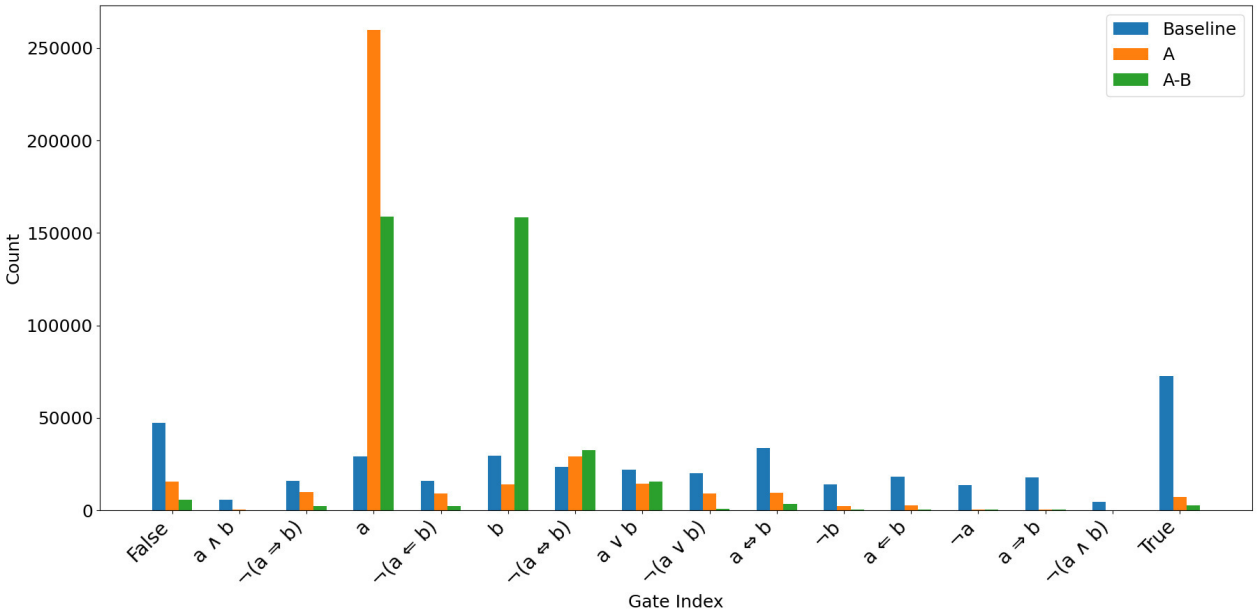


Fig. 6. DLNs$_{Ł}$, DLNs$_{Ł^*}$, DLNs$_{Ł^{**}}$ gate distributions in MNIST dataset

has probabilities $p_1 = 0.60$ and $p_4 = 0.40$ for two logic gates, deterministically selecting a gate at p_1 fails to capture the nuanced behavior expressed in the soft combination. This simplification may overlook critical interactions among the logic operators.

VI. Conclusions

We proposed a differentiable logic gate network based on a relaxation of classical logic gates using the Łukasiewicz T-norm. Our empirical findings support the hypothesis that the Łukasiewicz T-norm is not well-suited for gradient based training due to its gradient vanishing properties. This is consistent with observations reported in [10]. The experiments further demonstrate that encouraging better error propagation by biasing the weights of the computational units toward feedforward logic gates and avoiding the gates that can cause zeros in gradient significantly improves model accuracy and mitigates the gradient vanishing issue.

DLNs offer a distinct advantage over traditional neural networks in terms of interpretability. Due to the structural clarity, this facilitates direct inspection and tracing of information flow, making DLNs particularly suitable for applications in explainable artificial intelligence (XAI). Unlike standard neural networks, where interpretability often relies on post hoc analysis or surrogate explanations, DLNs provide intrinsic transparency through their architecture. This transparency reveals a recurrent pattern in trained DLN models: a disproportionately high number of units are configured as feedforward (identity) gates. While this contributes to computational efficiency and stable training dynamics, it also suggests a significant underutilization of the available computational capacity within the network. Many logic gate units effectively act as passthrough mechanisms, bypassing transformation or decision making roles.

Regardless of implementation details, differentiable logic gate networks generally require less memory and exhibit lower inference time compared to standard neural networks, highlighting their computational efficiency despite variable performance in accuracy. These findings underscore key differences across implementations of DLNs. While logic networks trained using Menger's T-norm may achieve higher accuracy, the Łukasiewicz based models exhibit faster inference times. This emphasizes the importance of careful design and evaluation when developing differentiable logic gate networks. In some applications, reducing inference time with a marginal decrease in accuracy may be preferable to higher accuracy but significantly slower performance.

Further performance gains in DLNs may be realized by optimizing the flow of information through structured and efficient connectivity patterns. One notable limitation in existing architectures is the underutilization of computational units, which can be mitigated by reusing logic elements across input features in a convolution like fashion [20] By enabling each compute unit to process multiple bit pairs in parallel rather than in a one-by-one manner, the network can exploit the capabilities of modern 64-bit arithmetic logic units (ALUs) on CPUs. This approach not only enhances the utilization of computational units in DLNs but also contributes to reducing computational overhead during inference. Introducing regularization strategies to constrain the weight vectors of computational units may help reduce discretization loss when translating from fuzzy to hard logic (Boolean). These offer promising directions for improving both the accuracy and efficiency of DLNs in practical applications.

Finally, while most medical machine learning studies focus on developing high-accuracy and computationally intensive models, DLNs offer a practical potential in low-latency, resource-constrained use cases. Applications such as oxygen blood saturation measuring or non-invasive glucose monitoring suggest alternative research directions for logic gate networks.

Acknowledgment

This paper presents results which will be presented as a part of master thesis of the first author prepare under the supervision of the second author. Authors would like to thank anonymous Reviewers for valuable comments and remarks which enabled to improve the paper.

References

[1] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996. Available: https://doi.org/10.24432/C5XW20

[2] S. Bosse, "IoT and Edge Computing using virtualized low-resource integer Machine Learning with support for CNN, ANN, and Decision Trees," *Proc. 18th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 35, pp. 367–376, 2023, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., IEEE. Available: http://dx.doi.org/10.15439/2023F7745

[3] J. Choi, Z. Wang, S. Venkataramani, P. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized Clipping Activation for Quantized Neural Networks," 2018. Available: https://arxiv.org/abs/1805.06085

[4] L. da Cruz, C. Sierra-Franco, G. Silva-Calpa, and A. Raposo, "Enabling Autonomous Medical Image Data Annotation: A human-in-the-loop Reinforcement Learning Approach," in *Proc. 16th Conf. on Computer Science and Intelligence Systems (FedCSIS)*, vol. 25, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., IEEE, 2021, pp. 271–279. Available: http://dx.doi.org/10.15439/2021F86

[5] L. Dey, S. Jana, T. Dasgupta, and T. Gupta, "Deciphering Clinical Narratives – Augmented Intelligence for Decision Making in Healthcare Sector," in *Proc. 18th Conf. on Computer Science and Intelligence Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35, *Annals of Computer Science and Information Systems*, IEEE, 2023, pp. 11–24. Available: http://dx.doi.org/10.15439/2023F3385

[6] D. Długosz, A. Królak, T. Eftestøl, S. Ørn, T. Wiktorski, K. R. J. Oskal, and M. Nygård, "ECG Signal Analysis for Troponin Level Assessment and Coronary Artery Disease Detection: the NEEDED Study 2014," in *Proc. 2018 Federated Conf. Comput. Sci. Inf. Syst.*, vol. 15, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds. IEEE, 2018, pp. 1065–1068. Available: http://dx.doi.org/10.15439/2018F247

[7] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," *in Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, 2015, pp. 1737–1746.

[8] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks," *J. Mach. Learn. Res.*, vol. 22, no. 1, art. no. 241, Jan. 2021, 124 pp.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014. Available: https://arxiv.org/abs/1412.6980

[10] E. van Krieken, E. Acar, and F. van Harmelen, "Analyzing Differentiable Fuzzy Logic Operators," *Artificial Intelligence*, vol. 302, 2022, pp. 103602. Available: https://www.sciencedirect.com/science/article/pii/S0004370221001533

[11] J. Kosiński, K. Szklanny, A. Wieczorkowska, and M. Wichrowski, “An Analysis of Game-Related Emotions Using EMOTIV EPOC,” *Proc. 2018 Federated Conf. on Computer Science and Information Systems (FedCSIS)*, vol. 15, Annals of Computer Science and Information Systems, pp. 913–917, 2018, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds. IEEE. Available: http://dx.doi.org/10.15439/2018F296

[12] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images," Univ. of Toronto, 2012.

[13] Y. LeCun, C. Cortes, and C. J. C. Burges, “The MNIST database of handwritten digits," 1998. Available: http://yann.lecun.com/exdb/mnist/

[14] M. Marcinkiewicz and G. Mrukwa, “Quantitative Impact of Label Noise on the Quality of Segmentation of Brain Tumors on MRI scans,” *Proc. 2019 Federated Conf. on Computer Science and Information Systems (FedCSIS)*, vol. 18, pp. 61–65, 2019. Edited by M. Ganzha, L. Maciaszek, and M. Paprzycki. IEEE. Available: http://dx.doi.org/10.15439/2019F273

[15] K. Menger, “Statistical Metrics," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 28, no. 12, Dec. 1942, pp. 535–537. Available: https://doi.org/10.1073/pnas.28.12.535

[16] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta, “Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science," *Nature Communications*, vol. 9, no. 1, Jun. 2018, art. 2383. Available: https://doi.org/10.1038/s41467-018-04316-3

[17] A. Morar, F. Moldoveanu, A. Moldoveanu, O. Balan, and V. Asavei, “GPU Accelerated 2D and 3D Image Processing,” in *Proc. 2017 Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, vol. 11, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., IEEE, 2017, pp. 653–656. Available: http://dx.doi.org/10.15439/2017F265

[18] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019. Available: https://arxiv.org/abs/1912.01703

[19] F. Petersen, C. Borgelt, H. Kuehne, and O. Deussen, “Deep Differentiable Logic Gate Networks," *in Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 2006–2018. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/0d3496dd0cec77a999c98d35003203ca-Paper-Conference.pdf

[20] F. Petersen, H. Kuehne, C. Borgelt, J. Welzel, and S. Ermon, “Convolutional Differentiable Logic Gate Networks," *in Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 121185–121203. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/db988b089d8d97d0f159c15ed0be6a71-Paper-Conference.pdf

[21] M. Pudo, M. Wosik, and A. Janicki, “Open Vocabulary Keyword Spotting with Small-Footprint ASR-based Architecture and Language Models,” *Proc. 18th Conf. Computer Science and Intelligence Systems (FedCSIS)*, Annals of Computer Science and Information Systems, vol. 35, pp. 657–666, 2023, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., IEEE. Available: http://dx.doi.org/10.15439/2023F8594

[22] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, “Binary neural networks: A survey," *Pattern Recognition*, vol. 105, 2020, art. 107281. Available: https://www.sciencedirect.com/science/article/pii/S0031320320300856

[23] M. Rapoport and T. Tamir, “Best Response Dynamics for VLSI Physical Design Placement,” in *Proc. 2019 Federated Conf. on Computer Science and Information Systems (FedCSIS)*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., IEEE, vol. 18, 2019, pp. 147–156. Available: http://dx.doi.org/10.15439/2019F91

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, Oct. 1986, pp. 533–536. Available: https://doi.org/10.1038/323533a0

[25] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, “Evolutionary Machine Learning: A Survey," *ACM Comput. Surv.*, vol. 54, no. 8, art. 161, Oct. 2021. Available: https://doi.org/10.1145/3467477

[26] S. Thrun et al.,“The Monk’s Problem’s: A Performance Comparison of Different Learning Methods," 1991. Available: https://api.semanticscholar.org/CorpusID:59810521

[27] L. Uberg and S. Kadry, “Analysis of Brain Tumor Using MRI Images,” in *Proc. 17th Conf. Computer Science and Intelligence Systems (FedCSIS)*, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30, *Annals of Computer Science and Information Systems*, IEEE, 2022, pp. 201–204. Available: http://dx.doi.org/10.15439/2022F69

[28] P. Wasilewski, and Ch. D. Nguy, “Deep Differentiable Logic Gate Networks Based on Fuzzy Zadeh’s T-norm,” Proceedings of 6th Polish Conference on Artificial Intelligence (PP-RAI 2025), Katowice, Poland, Apr. 7-9, 2025. to appear in Lecture Notes in Networks and Systems, Springer.

[29] L. A. Zadeh, “Fuzzy sets," *Information and Control*, vol. 8, no. 3, 1965, pp. 338–353. Available: https://www.sciencedirect.com/science/article/pii/S001999586590241X

[30] M. Zwitter and M. Soklic, “Breast Cancer," UCI Machine Learning Repository, 1988. Available: https://doi.org/10.24432/C51P4M

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 231–241
DOI: 10.15439/2025F4755
ISSN 2300-5963 ACSIS, Vol. 43

Segmentation and Process Assignment of Semi-Structured Event Logs

Piotr Przymus, Krzysztof Rykaczewski, Janusz Zieliński, Łukasz Mikulski
0000-0001-9548-2388, 0000-0002-4772-6330, 0000-0002-1127-9874, 0000-0002-6711-557X
Nicolaus Copernicus University in Toruń
ul. Gagarina 11, 87-100 Toruń, Poland
Email: {piotr.przymus, krzysztof.rykaczewski, janusz.zielinski, lukasz.mikulski}@mat.umk.pl

***Abstract*—Process mining provides valuable insights by discovering process models from execution logs. However, its effectiveness depends heavily on high-quality, well-structured logs. Many real-world systems produce low-level, semi-structured logs lacking clear process identifiers, causing misalignment with their intended process models.**

This paper introduces a method for structuring raw event logs by segmenting event streams and mapping them to known processes. Using process traces from experienced users, we develop a model that infers process assignments in unstructured logs. Our approach is motivated by a modular enterprise system without predefined workflows, where dynamic processes generate low-level logs requiring interpretation.

We validate our method on a semi-synthetic business dataset and a fully synthetic dataset from PLG2. Our results demonstrate that trace segmentation improves process discovery, aligns logs with meaningful structures, and significantly enhances process mining in unstructured environments.

I. Introduction

PROCESS mining is a powerful methodology that uses event logs to gain insights into business processes, enabling organizations to analyze, monitor, and optimize their workflows. Traditionally, process mining relies on structured event logs, where each event is explicitly linked to a process instance identifier. However, in many real-world contexts, event logs are only semi-structured, lacking process identifiers yet containing timestamps, event types, and user identifiers. This limitation complicates the reconstruction of process execution and hinders effective process analysis.

This research addresses that challenge by presenting an approach to structure semi-structured event logs. Specifically, we assign process identifiers to raw event data by leveraging a set of ideal process traces performed by experienced users. These labeled traces, which include known process identifiers and types, serve as training data for a machine learning-based system. The trained model is then applied to raw event logs to infer process assignments and identify user activities linked to specific processes.

We explored a practical application of this approach in a modular enterprise management system provided to external entities. These entities use the system to execute diverse business processes spanning multiple modules. Because of the variability of clients, the heterogeneity of business processes, and the system's flexibility, incorporating process identifiers into the logs is not feasible from a business perspective. Therefore, our methodology relies exclusively on semi-structured data.

By implementing this approach, we provide a solution that enhances process mining capabilities in environments where structured event logs are unavailable. This research contributes to process mining by introducing a method for structuring semi-structured event logs, enabling more effective business process analysis, anomaly detection, and performance monitoring.

a) Replication Package: To facilitate reproducibility and further research, we provide a complete replication package containing code, data, and experimental scripts. It is publicly available at:

`https://github.com/ncusi/Segmentation_and_Process_Assignment_of_Semi-StructuredEvent_Logs`

The remainder of this paper is organized as follows. Section II discusses related work. Sections III and IV reviews necessary preliminaries on event logs, process mining, and similarity measures. Section V states the problem, and Section VI describes our methodology for structuring semi-structured event logs. In Section VII, we discuss our experimental design, and Section VIII presents the results. In Section IX, we assess threats to validity. Finally, we conclude and propose future directions in Section X.

A. Running Example (Motivation)

Consider a customer-support system where each event is logged as `[time, user, activity, ...]`. A typical log snippet might look like:

```
2025-05-10T09:12:03Z, alice, OpenComplaint, details
2025-05-10T09:14:21Z, alice, AssignAgent, agentId=42
2025-05-10T09:15:07Z, alice, ResolveComplaint, status=closed
2025-05-10T09:20:00Z, alice, OpenComplaint, details
...
```

Without a case identifier, it is unclear which `OpenComplaint` corresponds to which conversation. This can cause supervisors to miss abandoned requests or misattribute resolution times. To address this, we propose to reconstruct each case by matching segments of the event stream to known process templates.

This work was supported by the Regional Operational Programme of the Kuyavian-Pomeranian Voivodeship for 2014–2020 under the grant titled "Budowa zaplecza badawczo-rozwojowego w MGA Sp. z o.o."

II. Related Work

Workflow and process mining have been extensively studied. Classical surveys such as [1], [2], [3] provide a comprehensive overview. A critical challenge in this field involves deriving high-quality, structured event logs from raw, often semi-structured data sources. Our work sits at the intersection of several active research areas, including event-case correlation, trace segmentation, similarity-based activity matching, and process discovery under uncertainty.

A fundamental problem in process mining is event-case correlation inferring case identifiers when they are absent or ambiguous. This challenge has been addressed through various heuristics and model-aware approaches. For example, Helal and Awad [4] propose a runtime event correlation method that uses process models and task durations to group events into likely cases. Brzychczy et al. [5] introduce a rule-based approach for detecting case boundaries in time series data, leveraging domain-specific structural regularities.

The extraction of event logs from relational or unstructured sources has also been widely explored. Andrews et al. [6] present RDB2Log, a quality-aware, semi-automated log extraction framework. Hernandez-Resendiz et al. [7] extend this work by introducing techniques to convert relational data into XES-compliant event logs, facilitating downstream analysis.

The problem of reconstructing traces from flat event streams or system logs has been tackled through temporal and contextual grouping methods. In the domain of unstructured customer service data, Kecht et al. [8] apply natural language inference to extract structured traces from customer service conversations. Korzeniowski and Goczyca [9] propose an automated log-template generation method (SLT) and an interaction-extraction pipeline that discovers application-to-application dependencies in enterprise logs, demonstrating its utility on a large banking system. One of the earliest treatments of the segmentation problem appears in [10], where the authors propose an event correlation function to partition unstructured logs into coherent subsequences.

One of the earliest treatments of the segmentation problem appears in [10], where the authors propose an event correlation function to partition unstructured logs into coherent subsequences. Our approach similarly focuses on identifying process instances from flat user streams, using data-driven similarity and conformance models.

Sequence similarity measures have been applied in activity and behavior mining. Levenshtein similarity, in particular, has proved effective by Tax et al. in [11] for clustering and aligning activity sequences. To the best of our knowledge, although the Tversky similarity measure is widely used in set and pattern matching, it has not previously been applied to trace similarity in process mining. Our work leverages both Levenshtein and Tversky measures to enable efficient and flexible segmentation.

The challenge of multi-case event streams where multiple processes or objects interact has been identified by Martin et al. in [12] as a key limitation in traditional process mining. Recent methods under the umbrella of object-centric process mining [13], [14], [15] provide more expressive models, but they require richer data and frequently involve more complex analysis pipelines. Our work addresses a simpler setting where each user executes one process at a time, but case boundaries remain implicit and must be inferred.

Event log imperfections such as missing case identifiers, incomplete traces, or overlapping activities are well-documented threats to process analysis. Jans et al. [16] discuss the implications of log quality in auditing scenarios. Our method specifically targets one imperfection: the lack of case identifiers in user-specific event streams.

Streaming scenarios add complexity, requiring online segmentation and model updating. Burattin et al. [17] propose methods for control-flow discovery from live event streams. While their focus is on evolving process models, their incremental techniques are relevant to our goal of continuously analyzing user activity streams.

In contrast to prior work, we assume access to a small set of example traces per process and center our efforts on segmenting flat, user-specific event streams with no case identifiers. We combine prefix-based segmentation with syntactic and conformance-based similarity scoring, which allows us to reconstruct approximate process instances even under noisy or truncated conditions. Our results show that lightweight similarity functions outperform full model conformance for segmentation tasks, offering a practical balance between accuracy and efficiency.

a) Theoretical and Practical Implications:

- *Theory:* We introduce a novel combination of syntactic (Levenshtein, Tversky) and semantic (log skeleton) similarity within a unified segmentation framework. Our proofs of injectivity in activity encoding and the boundary properties of our similarity measures contribute formal insights for future process-mining research.
- *Practice:* The method integrates seamlessly with existing BPM tools (e.g., PM4Py) without modifying ERP system logs. Its lightweight shingle computations and XTructure-based regex checks enable real-time segmentation on production event streams.

III. Preliminaries

We follow standard process mining notation, as introduced in [18]. The goal of process mining is to derive meaningful insights into the execution of business processes from recorded event data. The key input for process mining is the *event log*, which captures sequences of events corresponding to the execution of individual process instances, referred to as *cases*.

A *trace* is a finite sequence of events that represents the observed execution of a single case. By comparing and analyzing these traces, we can uncover patterns, identify deviations, and evaluate performance characteristics of the underlying process.

A *process* P represents the system or workflow whose behavior we aim to analyze. In process mining, the focus is on gaining insights into the execution of P by studying event logs,

which contain traces that correspond to specific executions or instances of process P.

A. Event Logs and Process Mining

We now introduce the formal elements used to represent event data, traces, and structured event logs.

a) Events: Let AN be a finite set of *attribute names*, and let Val be the set of all possible attribute values. We introduce a special symbol $\perp$ (bottom), with $\perp \notin \mathsf{Val}$, to denote "undefined."

The universe of events is a (possibly infinite) set E. Each event $e \in E$ is described by a *partial* function

$$\pi_e \colon \mathsf{AN} \rightharpoonup \mathsf{Val}, \tag{1}$$

where the arrow $\rightharpoonup$ indicates that for some attribute names a the value $\pi_e(a)$ may be undefined (i.e. $\pi_e(a) = \perp$). Concretely, for every $e \in E$ and $a \in \mathsf{AN}$, we have $\pi_e(a) \in \mathsf{Val} \cup \{\perp\}$.

b) Conditions on Events: We impose the following two conditions on every event $e \in E$:

(A1) $\pi_e(time) \neq \perp$, meaning that every event has a timestamp.
(A2) There exists an attribute $a \in \mathsf{AN} \setminus \{time\}$ such that $\pi_e(a) \neq \perp$, meaning every event has at least one non-time attribute with a value not equal to $\perp$.

Definition III.1 (Event Stream)**.** *An* event stream *(or event table) ET is a finite sequence of events:*

$$ET = \langle e_1, e_2, \ldots, e_n \rangle, \quad e_i \in E, \tag{2}$$

such that there exists an attribute $a \in \mathsf{AN} \setminus \{time\}$ with $\pi_a(e_i) \neq \perp$ for all $i = 1, \ldots, n$.

c) Finite Sequences: In what follows, we adopt the standard definition of finite sequences using Kleene star notation. This notation, common in formal language theory, is used throughout this work to denote sequences (e.g., traces of events).

Notation III.2 (Kleene star)**.** *Given a set X, we define X^* to be the set of all finite sequences over X. Formally, each element of X^* is of the form $\langle x_1, x_2, \ldots, x_n \rangle$ for some $n \geq 0$ and $x_i \in X$. The empty sequence is allowed if $n = 0$.*

Let $\Sigma \subseteq \mathsf{Val}$ be a finite set of *activity names*, known as the *activity alphabet*. A *trace* is a finite sequence $\sigma \in \Sigma^* = \{\text{all finite sequences over } \Sigma\}$ representing the control-flow perspective of a case.

Definition III.3 (Case Identifier and Cases)**.** *Let $id \in \mathsf{AN}$ be a selected* case identifier *attribute. The set of cases in ET is defined as:*

$$\mathsf{Cases}(ET, id) = \{\pi_{id}(e) \mid e \in ET,\ \pi_{id}(e) \neq \perp\}. \tag{3}$$

Remark III.4. *Each element $c \in \mathsf{Cases}(ET, id)$ is interpreted as a distinct* case *of the process within ET. Concretely, "case" refers to the actual value c of the attribute id. In this way, $\mathsf{Cases}(ET, id)$ collects exactly those case-identifier values that appear in the event stream ET, ensuring each value corresponds to a unique case.*

Definition III.5 (Trace of a Case)**.** *Let $c \in \mathsf{Cases}(ET, id)$. Then the* trace of case c *is the unique sequence of events*

$$\pi_{trace}(c) = \langle e_1, \ldots, e_k \rangle \in E^* \tag{4}$$

satisfying:

i) For all i, $\pi_{id}(e_i) = c$. In other words, each e_i belongs to the case identified by c.
ii) The sequence $\langle e_1, \ldots, e_k \rangle$ is sorted by non-decreasing timestamp, i.e.

$$\pi_{e_1}(time) \leq \pi_{e_2}(time) \leq \cdots \leq \pi_{e_k}(time), \tag{5}$$

recalling that $e.a := \pi_e(a)$.
iii) $\{e_1, \ldots, e_k\} = \{e \in ET \mid \pi_{id}(e) = c\}$, so we include all *events of that case (no more, no fewer).*

Remark III.6. *For each $c \in \mathsf{Cases}(ET, id)$, the sequence $\pi_{trace}(c)$ is defined as the ordered list of all events $e \in ET$ for which $\pi_{id}(e) = c$, arranged in non-decreasing order of $time$.*

IV. Process Models, Event Logs, and Conformance Checking

A. Process Models

Definition IV.1 (Process Models)**.** *Let Σ be an* activity alphabet *(i.e., a finite set of symbols). Denote by $\mathcal{M}_P(\Sigma)$ the set (or universe) of all process models that describe the behavior of a specific process P and are defined over the activity alphabet Σ. These models belong to a given class $\mathcal{P}$ (e.g., Petri nets, BPMN models, etc.) relevant to this work. An element $M \in \mathcal{M}_P(\Sigma)$ is called a* process model *if it prescribes which traces over Σ are possible in the corresponding process.*

Remark IV.2. *Let P denote a particular business process (e.g. "Order Handling"), and let Σ be its activity alphabet. We write*

$$\mathcal{M}_P(\Sigma) \subseteq \mathcal{P} \tag{6}$$

to mean "the set of all models in the model class $\mathcal{P}$ that describe process P over Σ."

Here:

- *P is the specific process whose behavior we wish to capture.*
- *$\mathcal{P}$ is a* class of modeling formalisms *(for example, Petri nets, BPMN diagrams, log skeletons, etc.).*
- *$\mathcal{M}_P(\Sigma)$ is the subset of $\mathcal{P}$ consisting of those individual models that generate or accept exactly the traces of P over alphabet Σ.*

Definition IV.3 (Induced Language of a Model)**.** *For each process model $M \in \mathcal{M}_P(\Sigma)$, associate a* language $\mathcal{L}_P(M) \subseteq \Sigma^*$*, where Σ^* denotes the set of all finite sequences over Σ. Formally,*

$$\mathcal{L}_P(M) = \{w \in \Sigma^* \mid w \text{ is a possible execution of } M\}. \tag{7}$$

Thus, each $w \in \mathcal{L}_P(M)$ is a finite sequence of activities representing one possible execution of M.

B. Finite Multisets of Traces

Definition IV.4 (Finite Multisets of Traces)**.** *Let X be a set. A* finite multiset *over X is a function $N\colon X \to \mathbb{N}$ such that $\{x \in X \mid N(x) \neq 0\}$ is finite. We call $N(x)$ the* multiplicity *of x in the multiset N. We write $B(X)$ for the set of all finite multisets of elements of X.*

Notation IV.5 (Multisets of Traces)**.** *Since Σ^* is the set of all finite sequences over Σ, the set of all finite multisets of traces is denoted by $B(\Sigma^*)$. An element $L \in B(\Sigma^*)$ is a multiset of traces, where $L(\sigma) \in \mathbb{N}$ indicates the number of occurrences of the trace σ in L.*

C. Structured Event Log

Assume that events are drawn from a universe E, each event $e \in E$ being described by a partial function $\pi_e\colon \mathsf{AN} \rightharpoonup \mathsf{Val}$.

Definition IV.6 (Structured Event Log)**.** *Let $id \in \mathsf{AN}$ be a designated case identifier attribute. A* structured event log *is a finite set of case-trace pairs:*

$$L = \{(c_i, \pi_{trace}(c_i)) \mid c_i \in \mathsf{Cases}(ET, id)\} \subseteq \mathsf{Val} \times E^*, \quad (8)$$

such that for each pair $(c, \pi_{trace}(c)) \in L$, the corresponding trace

$$\pi_{trace}(c) = \langle e_1, \ldots, e_k \rangle \in E^* \quad (9)$$

satisfies:

(i) $\forall e_j \in \pi_{trace}(c)$ *we have* $\pi_{id}(e_j) = c$*;*
(ii) $\forall e_j \in \pi_{trace}(c)$ *we have* $\pi_{time}(e_j) \neq \perp$*;*
(iii) $\forall e_j \in \pi_{trace}(c)\ \exists a \in \mathsf{AN} \setminus \{time, id\}$ *such that* $\pi_a(e_j) \neq \perp$.

Remark IV.7. *Each element $(c, \pi_{trace}(c)) \in L$ represents a single case c together with its corresponding trace. Thus, a structured event log explicitly associates each case identifier with the complete, temporally ordered sequence of events belonging to that case. Specifically, the log has the following hierarchical structure:*

- *A finite set of case-trace pairs L;*
- *Each trace $\pi_{trace}(c) \in E^*$ is a sequence of events related to a single case c;*
- *Every event $e \in \pi_{trace}(c)$ has a timestamp, the associated case identifier c, and at least one additional meaningful attribute (e.g., activity, resource).*

Notice that each case identifier c_i appears exactly once in the set L.

From this structured event log, one can perform further abstraction steps, such as applying *event classifiers*, to derive a *simple event log* $L' \in B(\Sigma^*)$ suitable for control-flow analysis.

D. Process Models and Conformance Checking

When process models are available, conformance checking can be used to compare a candidate trace to a known process structure. This enables semantic similarity beyond syntactic matching, especially useful in process mining when control-flow behavior must be validated.

Let $\mathcal{L}_P \subseteq \Sigma^*$ be a set of example traces for a process P. From this set, we can derive a process model $\mathcal{M}_P(\Sigma)$ using one of several discovery algorithms:

- **Alpha Miner** [19]: Constructs a Petri net by detecting causal and concurrent relations from the event log.
- **Heuristics Miner** [20]: Enhances the Alpha Miner with frequency-based dependency filtering to handle noise.
- **Log Skeleton** [21]: Builds a constraint-based model capturing behavioral relations such as equivalence, always-before, and always-after.

Let $\sigma \in \Sigma^*$ be a trace and $\mathcal{M}_P(\Sigma)$ the process model discovered from $\mathcal{L}_P$. We evaluate how well σ conforms to $\mathcal{M}_P(\Sigma)$ using one of the following conformance techniques:

- **Token-Based Replay Fitness (TBR)** [22]: Evaluates trace fitness by simulating token movement in a Petri net. We use the PM4Py function `pm4py.conformance.fitness_token_based_replay`. This is applied to models discovered via Alpha Miner and Heuristics Miner.
- **Log Skeleton Conformance**: Computes trace fitness by comparing the trace to a log skeleton model using structural constraint diagnostics. We use: `pm4py.conformance_log_skeleton`.

While these model-based approaches capture process semantics, we also employ syntactic techniques that operate on string representations of event sequences. These methods allow for efficient trace comparison when no explicit model is available or required.

Process Discovery Configuration: All discovery was done with PM4Py v2.9.3:

- **Heuristics Miner:** `dependency_threshold=0.99, and_threshold=0.65, or_threshold=0.65, min_support=0.10`.
- **Alpha Miner:** `dependency_threshold=0.50` (PM4Py default).

These were chosen because Heuristics is noise-robust and frequency-based, whereas Alpha provides a simple, well-understood baseline.

E. String Representation and Similarity Measures

To support efficient and scalable process matching, we transform sequences of events into compact string representations that preserve their sequential structure. This enables the application of string-based similarity techniques, including shingling and character-edit distance.

a) Activity Encoding: To enable string-based similarity comparison, each activity in a trace is deterministically mapped to a fixed-length character code. This allows the transformation of activity sequences into strings over a finite alphabet.

Definition IV.8 (Activity Encoding)**.** *Let Σ denote the finite set of activity labels. Define an injective encoding function*

$$C\colon \Sigma \to \Gamma^n, \quad (10)$$

where Γ is a finite character alphabet, and $n \in \mathbb{N}$ is a fixed code length. The set Γ^n represents all possible strings of length n formed from alphabet Γ. This encoding assigns each activity a unique code of length n, but can be extended to $\widetilde{C}\colon \Sigma^* \to (\Gamma^n)^*$ by allowing any trace $\sigma = \langle a_1, \ldots, a_k \rangle \in \Sigma^*$ to be represented as a concatenated string $\widetilde{C}(\sigma) = C(a_1) \cdot C(a_2) \cdot \cdots \cdot C(a_k) \in (\Gamma^n)^*$, where $\widetilde{C}(\langle\rangle)$ is the empty string and $\cdot$ (dot) denotes concatenation.

Remark IV.9 (Practical Consideration:)**.** *In our implementation, we use $n = 3$ and define*

$$\Gamma = \{A, B, \ldots, Z, a, b, \ldots, z, 0, \ldots, 9\}, \tag{11}$$

with $|\Gamma| = 62$. This provides a sufficiently large encoding space for typical process alphabets, ensuring compactness and facilitating downstream operations like shingling.

We now formalize the injectivity of the activity encoding in the following proposition:

Proposition IV.10 (Uniqueness Preservation)**.** *Let $C\colon \Sigma \to \Gamma^n$ be an injective encoding function as defined. Then, for any two distinct traces $\sigma_1, \sigma_2 \in \Sigma^*$, their encoded representations $\widetilde{C}(\sigma_1)$ and $\widetilde{C}(\sigma_2)$ are distinct.*

Proof: We prove the claim by induction on the length k of the sequences.

Base Case: For $k = 0$ (the empty sequence), $\widetilde{C}(\langle\rangle)$ is the empty string, and the claim holds trivially.

Inductive Step: Assume the statement holds for all sequences of length k. Let

$$\sigma_1 = (a_1, a_2, \ldots, a_k, a_{k+1}) \quad \text{and} \quad \sigma_2 = (b_1, b_2, \ldots, b_k, b_{k+1})$$

be sequences in Σ^{k+1} such that

$$\widetilde{C}(\sigma_1) = \widetilde{C}(\sigma_2). \tag{12}$$

Writing

$$\widetilde{C}(\sigma_1) = C(a_1) \cdot C(a_2) \cdots C(a_k) \cdot C(a_{k+1}), \tag{13}$$

and

$$\widetilde{C}(\sigma_2) = C(b_1) \cdot C(b_2) \cdots C(b_k) \cdot C(b_{k+1}), \tag{14}$$

since each $C(a_i)$ is a fixed-length string (of length n), the concatenation is unambiguous. Comparing the first n characters of both sides gives

$$C(a_1) = C(b_1). \tag{15}$$

Because C is injective, it follows that

$$a_1 = b_1. \tag{16}$$

Removing the first code from both concatenations, we then have

$$\widetilde{C}(a_2, \ldots, a_{k+1}) = \widetilde{C}(b_2, \ldots, b_{k+1}). \tag{17}$$

By the induction hypothesis,

$$(a_2, \ldots, a_{k+1}) = (b_2, \ldots, b_{k+1}). \tag{18}$$

Thus,

$$(a_1, a_2, \ldots, a_{k+1}) = (b_1, b_2, \ldots, b_{k+1}). \tag{19}$$

This completes the inductive step, and therefore $\widetilde{C}$ is injective. ∎

b) Shingling: To capture local sequential patterns within the encoded string representation of traces, we apply a fixed-length sliding window operation known as *shingling*.

Definition IV.11 (Shingle Set)**.** *Let $x \in (\Gamma^n)^*$ be an encoded string representation of a trace, and let $k \in \mathbb{N}$ be a fixed window size. The* shingle set *of x is defined as:*

$$\text{Shingle}(x, k) = \{x[i : i+k-1] \mid 1 \le i \le |x| - k + 1\}, \tag{20}$$

where $|x|$ is the length of the string x, and $x[i : i + k - 1]$ denotes the substring of length k starting at position i.

c) Similarity Measures: We employ two types of similarity functions to compare string-encoded traces or their shingle sets.

Definition IV.12 (Similarity Measures)**.**

(1) ***Tversky Similarity.***
Let $A, B \subseteq \mathcal{U}$ be finite sets (e.g., sets of shingles). Define the similarity function:

$$\tau(A, B) := \frac{|A \cap B|}{|A \cap B| + \alpha|A \setminus B| + \beta|B \setminus A|}, \tag{21}$$

where $\alpha, \beta \ge 0$ are tunable parameters (commonly $\alpha = 0$, $\beta = 1$), and $|X|$ denotes the cardinality of set X. This asymmetric variant prioritizes recall over precision, treating containment as more significant than a symmetric match. Consequently, if one set is entirely contained within the other, the similarity measure reaches its maximum value, reflecting a strong inclusion relationship.

(2) ***Levenshtein Similarity.***
Let $a, b \in (\Gamma^n)^*$ be two encoded strings, and let $d(a, b) \in \mathbb{N}_0$ be the Levenshtein distance between them (i.e., the minimum number of single-character insertions, deletions, or substitutions needed to transform a into b). A common normalization for Levenshtein similarity is:

$$\ell(a, b) := 1 - \frac{d(a, b)}{\max\{|a|, |b|\}}. \tag{22}$$

Here, $\ell(a, b)$ yields 1 if the strings are identical and 0 if $d(a, b)$ is as large as the maximum length of the two strings (e.g., $d(a, b) = |b|$ if $|a| = 0$). This approach is widely used due to its intuitive interpretation.

Remark IV.13 (Alternative Normalization)**.** *In some applications, one might prefer a custom normalization to give special treatment to length differences or partial overlaps (e.g., subtracting $\big||a| - |b|\big|$ from $d(a, b)$). If that approach is desired, it must be clearly justified. For example, one might wish to penalize purely length-based mismatches less than mismatches in the overlapping region of the two strings. In such cases, the*

similarity function could be:

$$\ell'(a,b) := \begin{cases} 1, & \text{if } d(a,b) - \big||a| - |b|\big| = 0, \\ \frac{1}{d(a,b) - \big||a| - |b|\big|}, & \text{otherwise.} \end{cases}$$

However, this version is less common and can be confusing without a thorough explanation of the motivation behind it.

This is a well-known property of the above-mentioned similarity measures:

Theorem IV.14 (Boundary Conditions of Similarity Functions)**.** *For any two encoded strings $a, b \in (\Gamma^n)^*$:*

- *The Levenshtein similarity $\ell(a,b)$ satisfies $\ell(a,b) = 1$ if and only if $a = b$, and $\ell(a,b) = 0$ when the number of edits equals the length of the longer string (i.e., maximally different).*
- *For the Tversky index $\tau(A,B)$ (with a typical setting of $\alpha = 0$ and $\beta = 1$), it holds that $\tau(A,B) = 1$ if and only if $B \subseteq A$.*

Remark IV.15. *The normalization in $\ell(a,b)$ subtracts the minimum number of edits required to align the lengths of a and b, isolating the structural mismatches. This ensures that string pairs differing only in length are treated as more similar than those with substantive internal edits.*

V. Problem Statement

In many enterprise systems, users interact with the platform through various interfaces, executing tasks that correspond to different business processes (e.g., onboarding a customer, handling a complaint, reviewing a case). These interactions are recorded as user-specific event streams. However, in most operational settings, this data is captured in a semi-structured format: each event includes the time it occurred, the action taken (activity name), and the user who performed it, but does not explicitly identify which process or case it belongs to.

Let $ET_u = \langle e_1, e_2, \ldots, e_n \rangle$ denote the event stream of a user $u \in \mathsf{Val}$, where each event $e_i \in E$ satisfies:

$$\pi_{user}(e_i) = u, \quad \pi_{time}(e_i) \neq \bot, \quad \pi_{event}(e_i) \in \Sigma. \tag{23}$$

This stream captures a chronologically ordered sequence of user interactions, i.e.

$$\pi_{e_1}(time) \leq \pi_{e_2}(time) \leq \cdots \leq \pi_{e_n}(time), \tag{24}$$

but without any case identifier. As such, it is considered a *semi-structured event stream*.

To support process-aware analysis and monitoring, the business needs to understand which processes a user was involved in, when they started and ended, and what behavior was exhibited within each instance. For example, a supervisor may wish to assess whether an employee followed the correct handling procedure for different types of customer requests.

To enable this, the organization provides, for each process type P_i, $i = 1, \ldots, k$, a set of example traces:

$$\mathcal{L}_{P_i} \subseteq \Sigma^*, \tag{25}$$

which have been collected in a controlled environment. In this setting, expert users executed process cases one at a time, with clear boundaries between cases. These labeled traces serve as templates for recognizing process behavior in live, unstructured event streams.

Definition of the Process-Specific Alphabet. We define

$$\Sigma_{P_i} = \bigcup_{\sigma \in \mathcal{L}_{P_i}} \{\text{activities in } \sigma\} \subseteq \Sigma, \tag{26}$$

i.e., Σ_{P_i} is the set of activity labels that appear in the example traces for process P_i, $i = 1, \ldots, k$.

Several challenges arise in this scenario:

- **Overlapping Activities:** Activities (event names) may appear in multiple processes, i.e. $\Sigma_{P_i} \cap \Sigma_{P_j} \neq \emptyset$ for some $i \neq j$, $i, j = 1, \ldots, k$.
- **Multiple Instances:** The same user u may execute multiple process instances across different process types.
- **Sequential, Possibly Abandoned Executions:** The user performs processes sequentially (never in true parallel) but may abandon one and start another.

a) Objective: Given a semi-structured event stream ET_u and example trace sets $\mathcal{L}_{P_1}, \ldots, \mathcal{L}_{P_k}$ for known processes, the goal is to approximate the user's activity history in terms of which processes were executed and when.

We aim to segment the user stream into trace-like subsequences that are likely to correspond to process instances. Each segment is matched to one of the known processes based on similarity. At this stage, our goal is not to produce a perfect segmentation, but rather a **good estimation** of the user's behavior from a process-centric perspective.

VI. ProcessMatcher Overview

This section introduces **ProcessMatcher**, a two-phase framework for real-time detection of known business processes in semi-structured event streams. In the **offline modeling** phase, curated example traces are filtered for noise and standardized; events are encoded into fixed-length tokens and passed through XTructure to learn hierarchical prefix regexes, alongside a k-shingle extractor that produces compact signatures. Optionally, a light-weight conformance analysis validates pattern quality. In the **online segmentation** phase, incoming streams are partitioned into candidate segments and matched against the learned regexes and shingle signatures using fast approximate similarity measures, yielding polynomial-time detection.

A. Methodology

The ProcessMatcher pipeline comprises three stages:

(a) **Training Data Preparation.** Noisy or incomplete traces are removed; activity labels are normalized; events are encoded as fixed-length character codes (Definition IV.8).
(b) **Pattern Learning.** Normalized sequences feed XTructure for prefix generalization, producing a regex $\mathcal{R}_i$ for each process, and into the k-shingle extractor to compute each process's signature.

(c) **Parameter Tuning.** A grid search over shingle length k, similarity threshold θ, and prefix depth p is conducted on a held-out validation set to jointly optimize classification accuracy and runtime.

This clear separation of phases ensures reproducibility, modularity, and straightforward extensibility of individual components.

B. *XTructure* Primer

`XTructure` [23] is a lightweight, polynomial-time structure for learning and representing the syntax of semi-structured strings (e.g., activity-encoded traces) as a DFA-equivalent model. Unlike full regular-expression inference—which is NP-hard—`XTructure` builds a concise pattern model in a single pass over the data. Its design comprises three hierarchical layers:

1) **Symbol Layer:**
   - Records empirical counts of characters at each position.
   - Groups symbols into character classes (e.g. digit, letter) via a χ^2 homogeneity test when possible, or else enumerates the top-k symbols to cover a fixed fraction of observations.
2) **Token Layer:**
   - Splits input strings on delimiters into tokens.
   - Represents each token as a sequence of symbol-layer distributions, capturing both fixed and variable-length substrings within that token.
3) **Branch Layer:**
   - Accommodates multiple syntactic variants (e.g. date formats) by maintaining up to a bounded number of token-sequence "branches."
   - Creates a new branch when an incoming example's *fit score* exceeds a threshold for all existing branches, then performs branch-and-merge to enforce a global branch limit.

Once learned in a single (streaming) pass, an `XTructure` instance supports:

- **Serialization** to a compact, human-readable regular expression (an OR of its branches).
- **Sampling** of synthetic examples by drawing from its learned symbol distributions.
- **Comparison** to other patterns (regular expressions or other `XTructures`) via Monte Carlo sampling and CLT-based fitness estimation.

This combination of expressivity and efficiency makes `XTructure` orders of magnitude faster than general regex-inference techniques, yet sufficiently powerful to capture all finite-language structure observed in real-world enterprise event data.

C. ProcessMatcher Architecture

To identify instances of a process within semi-structured event streams, we introduce the `ProcessMatcher`, a modular component designed to match candidate trace segments to a known process P using a combination of syntactic and conformance-based techniques.

For each known process P_i, $i = 1, \ldots, k$, a separate matcher instance M_i is constructed using its set of example traces $\mathcal{L}_{P_i} \subseteq \Sigma^*$. The matcher learns a lightweight representation of the process, which is then used to detect and score segments within user-specific event streams.

The approach is divided into two phases: an offline *process modeling phase*, and an online *stream segmentation and matching phase*.

D. Parameter Specification

Default values and allowable ranges for the key parameters of `ProcessMatcher`:

- **Shingle length k:** default $k = 6$, range $4 \leq k \leq 10$.
- **Similarity threshold θ:** default $\theta = 0.8$, range $0.5 \leq \theta \leq 0.95$.
- **Prefix length for `XTructure`:** default $p = 18$ characters (i.e., 6 activity codes $\times$ 3 chars each), range $9 \leq p \leq 30$.
- **Tversky parameters (α, β):** default $(\alpha, \beta) = (0, 1)$, both non-negative.

The method `contains_process`(S) works as follows:

1) Compute the set of shinglets (character k-grams) from the encoded string S:
$$A = \mathrm{Sh}(S). \tag{27}$$
2) For each reference shinglet set $\mathrm{Sh}(\sigma_j)$ in $\mathcal{L}_{P_i}$, $i = 1, \ldots, k$, set
$$B = \mathrm{Sh}(\sigma_j), \tag{28}$$
then compute the Tversky index $\tau(A, B)$ given by Equation (21).
3) If any $\tau(A, B) \geq \theta$, return `true`; otherwise return `false`.

E. *XTructure* Explanation

We construct a prefix regular expression $\mathcal{R}_i$, $i = 1, \ldots, k$, from observed prefixes of length p using `XTructure`:

1) Initialize: `xpref = XTructure(max_branches=2)`.
2) For each trace $\sigma \in \mathcal{L}_{P_i}$, $i = 1, \ldots, k$, extract `prefix =` σ`[:p]` and call `xpref.learn_new_word(prefix)`.
3) After all prefixes are learned, `str(xpref)` yields a regex, e.g. `(ABC|XYZ)`, capturing the generalized prefix patterns.

F. Phase 1: Offline Process Modeling

Given example traces $\mathcal{L}_{P_i}$ of process P_i, $i = 1, \ldots, k$, the `ProcessMatcher` performs the following:

1) **Process Signature Construction:** A process signature is built using *shinglets*, which are sets of overlapping substrings (character n-grams) extracted from traces in $\mathcal{L}_{P_i}$, $i = 1, \ldots, k$. This signature enables fast approximate matching using the Tversky index during online presence checking.
2) **Prefix Pattern Learning:** The initial prefixes of example traces are used to learn a regular expression $\mathcal{R}_i$ that characterizes common process starts. This is done using the

`XTructure` algorithm [23], which generalizes prefixes into a compact, discriminative regex by merging observed patterns. This allows rapid detection of potential case beginnings within user streams.

3) **Optional Process Model Discovery:** If desired, a process model $\mathcal{M}_{P_i}(\Sigma)$ (e.g., Petri net, log skeleton) is discovered from $\mathcal{L}_{P_i}$ using PM4Py, $i = 1, \ldots, k$. This enables conformance-based fitness computation as an alternative to syntactic similarity.

G. Phase 2: Online Stream Segmentation and Matching

Given a user event stream $ET_u = \langle e_1, e_2, \ldots, e_n \rangle$ with $\pi_{user}(e_i) = u$ and no case identifier, we segment and label the stream with likely process instances using the following steps, executed per process matcher M_i, $i = 1, \ldots, k$:

1) **Process Presence Check:** Using the shinglet-based Tversky similarity, we determine whether the user stream contains any segment sufficiently similar to $\mathcal{L}_{P_i}$, $i = 1, \ldots, k$. This acts as a fast filter to skip irrelevant streams.
2) **Prefix Detection:** We apply the learned prefix regular expression $\mathcal{R}_i$ over the stream to detect potential case start positions. Each match defines a candidate window boundary.
3) **Similarity Scoring:** For each candidate segment $\sigma \in \Sigma^*$ (bounded by prefix matches), we compute its similarity to the process using one of several methods:
   - **Tversky Index (default)**: uses shinglets and the Tversky index.
   - **Levenshtein Distance**: normalized edit distance.
   - **Conformance Fitness**: if a process model $\mathcal{M}_{P_i}(\Sigma)$, $i = 1, \ldots, k$, is available, token-based replay or log skeleton diagnostics can be used (e.g. Conformance checking based on Log Skeleton Conformance, Heuristics Miner, and Alpha Miner).
4) **Instance Counting:** The number of likely process instances within the stream is estimated by summing the similarity scores of detected segments. Only segments whose score exceeds a predefined threshold are counted.

This approach offers a flexible, efficient method for process instance identification from flat user event streams, enabling reconstruction of activity histories even in the absence of explicit case identifiers.

Remark VI.1. *Set $\mathcal{M} = \{M_1, \ldots, M_k\}$ denotes a set of process matchers, each of which corresponds to a* different *known procedure (process) $P_i = M_i.\texttt{process}$, $i = 1, \ldots, k$. In other words, for each process P_i we have a separate matcher M_i assigned.*

Remark VI.2. *In Algoithm 1 we use function `EncodeStreamAsActivityString` which is a procedure that transforms a user's event stream ET_u into a string where each character (or substring) represents a specific activity. In practice, for each event $e_i \in ET_u$, this function:*

- *Retrieves the value of the attribute corresponding to the activity (for example, $\pi_{event}(e_i)$).*
- *Maps this value to a fixed character code according to a predefined activity encoding function $C \colon \Sigma \to \Gamma^n$. As defined earlier in Definition IV.8, C is an injective function that assigns each activity $a \in \Sigma$ a unique fixed-length code $C(a) \in \Gamma^n$, where Γ is a finite character alphabet and n is a fixed code length.*
- *Concatenates these codes to construct a string S, which represents the ordered sequence of activities in the stream.*

This string representation enables further operations, such as prefix detection, stream segmentation, and the calculation of similarity measures between segments of events, which are crucial in the process matching algorithm.

Algorithm 1 Segmenting a User Event Stream Using Process Matchers

Require: User event stream $ET_u = \langle e_1, e_2, \ldots, e_n \rangle$
Require: Process matchers $\mathcal{M} = \{M_1, \ldots, M_k\}$
Ensure: Set of labeled trace segments $\mathcal{T}$

$\mathcal{T} \leftarrow \emptyset$
$S \leftarrow$ EncodeStreamAsActivityString(ET_u)
for all $M_i \in \mathcal{M}$ $(i = 1, \ldots, k)$ **do**
  if M_i.contains_process(S) **then**
    $B \leftarrow M_i$.prefix_regex.find_all_matches(S)
    for $j = 1$ **to** $|B| - 1$ **do**
      $\sigma \leftarrow S[B[j] : B[j+1]]$
      $s \leftarrow M_i$.process_similarity(σ)
      if $s \geq \theta$ **then**
        $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\sigma, M_i.\texttt{process})\}$
      end if
    end for
    $\sigma \leftarrow S[B[-1] :]$
    $s \leftarrow M_i$.process_similarity(σ)
    if $s \geq \theta$ **then**
      $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\sigma, M_i.\texttt{process})\}$
    end if
  end if
end for

return $\mathcal{T}$

Proposition VI.3. *Assume that a user event stream ET_u is formed by the concatenation of process traces*

$$ET_u = \sigma_1 \cdot \sigma_2 \cdots \sigma_k, \tag{29}$$

where each $\sigma_i \in \mathcal{L}_P$ (the set of example traces for process P). Suppose further that the learned prefix regular expression $\mathcal{R}$ recognizes exactly the starting positions of any trace in $\mathcal{L}_P$. Then, applying the `ProcessMatcher` segmentation algorithm to ET_u will recover the segments $\sigma_1, \sigma_2, \ldots, \sigma_k$ exactly.

Proof: Let S be the string representation of ET_u obtained by the encoding function $\widetilde{C}$. By assumption, each process instance σ_i corresponds to an encoded substring $s_i = \widetilde{C}(\sigma_i)$

and the overall stream is

$$S = s_1 \cdot s_2 \cdots s_k. \tag{30}$$

By the correctness of the learned prefix regular expression $\mathcal{R}$, for every i we have:

$$\mathcal{R}(S[i]) = \text{true} \quad \Longleftrightarrow \quad i \text{ is the starting position of some } s_j.$$

Let $B = \{i_1, i_2, \ldots, i_k\}$ be the set of positions where $\mathcal{R}$ matches in S, with $i_1 = 1$. The segmentation algorithm then forms segments by taking

$$\sigma_j = S[i_j : i_{j+1} - 1] \quad \text{for } j = 1, \ldots, k-1, \tag{31}$$

and $\sigma_k = S[i_k : |S|]$. Since the stream S is exactly $s_1 s_2 \cdots s_k$, it follows that for each j,

$$\sigma_j = s_j = \widetilde{C}(\sigma_j). \tag{32}$$

Thus, the segmentation algorithm recovers the exact boundaries of each process instance ■

Corollary VI.4. *The segmentation algorithm will always find all process instances if the assumptions about prefixes are satisfied.*

VII. Experiment Design

To evaluate the proposed segmentation method, we conduct two experiments: one semi-synthetic (based on real process traces) and one fully synthetic (generated via process simulation). Both experiments follow a similar structure: we assume that for each process P_i a set of example traces $\mathcal{L}_{P_i}$, $i = 1, \ldots, k$, is available, and we generate user event streams by concatenating process executions. We evaluate the model's ability to identify the underlying process structure from these unstructured streams.

A. Process Setup and Trace Sources

- **Experiment 1 Semi-Synthetic (Real Traces):** We use two real-world processes from our business use case. For each process, we obtain at least 5000 example traces (of various events long). These are used both as reference traces for training the `ProcessMatcher` and as source traces for simulating user activity.
- **Experiment 2 Fully Synthetic (Simulated Processes):** We generate five synthetic process models using the PLG2 framework [24]. For each model, we generate:
  - 20 ideal (noise-free) traces for $\mathcal{L}_{P_i}$, $i = 1, \ldots, k$.
  - 100 perturbed traces using PLG2's default noise settings (e.g., skipped steps, inserted noise, reordering).

B. User Event Stream Synthesis

In both experiments, we generate synthetic user event streams by randomly sampling process traces (ideal or perturbed) and concatenating them. For each user event stream ET_u^{syn}, we apply the following:

- Each user stream contains a random number of process executions (between $[10, 40]$).
- For each selected execution:
  - In the semi-synthetic setting, traces are sampled from the set of ideal real-world traces.
  - In the fully synthetic setting, traces are sampled from the pool of perturbed traces generated via PLG2.
- In the semi-synthetic setting, sampled traces may be further perturbed by truncating their suffix, simulating premature termination of process instances.
- For each synthetic user event stream, we retain the ground truth segmentation: the start and end positions of each process instance, and the corresponding process identity.

This setup yields realistic, unstructured user streams while preserving the ability to evaluate segmentation and classification performance precisely.

C. Iteration and Distribution Variation

To evaluate robustness across different usage distributions, we repeat each experiment over multiple iterations. In each iteration, the probability of selecting each process is varied, producing both balanced and skewed distributions of process executions within streams. This simulates scenarios ranging from uniformly active users to users focused on one or two dominant processes.

D. Evaluation Metrics

Since the true structure of each synthetic user event stream is known, we evaluate performance using the following metrics: **Process Classification Accuracy:** Proportion of predicted segments assigned to the correct process. Together, these experiments allow us to evaluate both the accuracy and robustness of the method under realistic variability and controlled noise.

VIII. Results and discussion

We evaluate our method in two benchmark scenarios, described in Section VIII-A: a semi-synthetic setting based on real process traces, and a fully synthetic setting using processes generated by PLG2 [24]. In both scenarios, we assess the performance of different similarity scoring functions within the segmentation pipeline described earlier. Table I summarizes the accuracy and average processing time for each similarity scoring function, evaluated across both experimental settings.

Each experiment reports both classification accuracy how accurately the detected segments match the correct process and processing time per user stream (in seconds). We compare five similarity or conformance scoring approaches:

- **Levenshtein Similarity**: Edit distance over string-encoded traces.
- **Tversky Similarity**: Shingle-based set similarity.
- **Log Skeleton Conformance** [21]: Constraint-based model alignment.
- **Heuristic Miner + Token Replay** [20]: Frequency-based process model with token-based fitness scoring.
- **Alpha Miner + Token Replay** [19]: Basic Petri net discovery and replay fitness.

TABLE I
RESULTS: ACCURACY AND TIME ON ALL FUNCTIONS ON BOTH EXPERIMENTAL SETUPS

Scoring Function	**Fully Synthetic**		**Semi-Synthetic**	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
Levenshtein Similarity	**99.56**	**81.03**	**100.0**	**7.12**
Tversky Similarity	**91.14**	**94.55**	**99.43**	**14.10**
Log Skeleton Conformance	**79.55**	**11242**	**88.79**	**812**
Heuristic Miner + TBR	**60.8**	**13170**	**53.32**	**1214**
Alpha Miner + TBR	**65.5**	**5572**	**14.78**	**1273**

A. Discussion

As shown in Table I, Levenshtein similarity consistently achieves the highest segmentation and classification accuracy across both experimental settings. It also outperforms the other methods in terms of runtime efficiency, making it a strong default choice.

Tversky similarity, while slightly less accurate, remains competitive and exhibits favorable performance in both accuracy and runtime. Both syntactic similarity methods clearly outperform conformance-based approaches in this segmentation task.

Among the token-replay-based methods, the Log Skeleton model achieves the best accuracy but still falls short of the syntactic techniques. Heuristic Miner and Alpha Miner produce lower classification accuracy and efficiency, suggesting that these models are not well suited to trace-level segmentation in the absence of case structure.

a) Takeaway: For user stream segmentation tasks without explicit case identifiers, syntactic similarity-based scoring functions (Levenshtein and Tversky) provide the best balance between accuracy and computational cost.

IX. THREATS TO VALIDITY

Although the proposed approach shows promising results, several limitations must be acknowledged. We discuss potential threats to validity across four standard dimensions: internal, external, construct, and conclusion validity.

a) Internal Validity: The segmentation algorithm relies on synthetic or semi-synthetic user streams, where process boundaries are derived from known trace sets and artificially injected noise. While this setup allows for precise ground-truth evaluation, it may not capture all forms of real-world user behavior, such as interleaved or overlapping process executions. In addition, prefix learning and similarity scores are sensitive to parameter choices (e.g., shingle length, Tversky threshold), which may influence the outcomes.

b) External Validity: Our experiments are conducted on two real processes and five simulated ones. Although these span diverse scenarios, they may not fully reflect the complexity or variability of processes in other domains (e.g., healthcare, manufacturing). Moreover, user behavior in production systems can involve more irregular or non-sequential activity patterns, which are not entirely accounted for by our current assumptions.

c) Construct Validity: We measure segmentation accuracy using a ground truth built from ideal and truncated traces. However, the assumption that a truncated trace still represents a meaningful process instance may not always hold. Furthermore, our evaluation treats the best-matching process as correct, even though some traces could be ambiguous due to shared activities across multiple processes.

d) Conclusion Validity: Our evaluation metrics focus on classification accuracy and runtime per stream, providing clear comparative insights but excluding statistical significance testing. In future work, we plan to incorporate cross-validation, error analysis, and statistical robustness measures to reinforce confidence in the observed performance differences.

X. CONCLUSION AND FUTURE WORK

In this paper, we propose a method for segmenting semi-structured user event streams into process instances using a combination of syntactic similarity and model-based conformance checking. Our approach leverages known example traces for each process to construct lightweight matcher components that identify candidate segments based on learned prefixes and similarity scoring. This design enables segmentation in scenarios where case identifiers are absent and activities may overlap across processes.

We evaluated the method in two settings: (1) a semi-synthetic setup based on real-world processes, and (2) a fully synthetic benchmark generated with PLG2. The results show that syntactic similarity measures particularly Levenshtein and Tversky offer strong accuracy while remaining computationally efficient. Although semantically richer, model-based conformance approaches performed less effectively in this segmentation context, likely because of their sensitivity to partial traces and noise.

a) Future Work: Several extensions of this work are planned. First, we intend to address the issue of overlapping or conflicting segment assignments when multiple processes match the same trace window. Incorporating overlap resolution strategies such as greedy selection based on similarity scores or probabilistic modeling could yield cleaner segmentations. Second, we aim to evaluate the method on real, unlabeled user data in production environments, with expert annotations serving as ground truth. Finally, integrating sequence models (e.g., RNNs or transformers) or embedding-based representations may further enhance trace similarity beyond symbolic encoding.

Our findings highlight the potential of combining simple syntactic techniques with process mining insights to extract structured information from unstructured logs, thereby enabling more interpretable and process-aware user behavior analysis.

b) Prospects for Sequence Models: In future work, we plan to explore sequence-model approaches (RNNs, LSTMs, or Transformers). This would allow us to:

- Learn continuous embeddings of event sequences instead of manual character codes.
- Capture long-range temporal dependencies and non-linear patterns inherent in complex processes.
- Support online adaptation via incremental fine-tuning on newly observed streams without retraining shingle sets or XTructure models.

These extensions should improve robustness to noise and adaptivity to evolving business logics.

REFERENCES

[1] W. M. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. Weijters, "Workflow mining: A survey of issues and approaches," *Data & knowledge engineering*, vol. 47, no. 2, pp. 237–267, 2003.

[2] J. Claes and G. Poels, "Process mining and the prom framework: an exploratory survey," in *Business Process Management Workshops: BPM 2012 International Workshops, Tallinn, Estonia, September 3, 2012. Revised Papers 10*. Springer, 2013, pp. 187–198.

[3] S. J. van Zelst, F. Mannhardt, M. de Leoni, and A. Koschmider, "Event abstraction in process mining: literature review and taxonomy," *Granular Computing*, vol. 6, pp. 719–736, 2021.

[4] I. Helal and A. Awad, "Correlating unlabeled events at runtime," *arXiv preprint arXiv:2004.09971*, 2020.

[5] E. Brzychczy, T. Pełech-Pilichowski, and Z. Dworakowski, "Case id detection based on time series data – the mining use case," 2024. [Online]. Available: https://arxiv.org/abs/2410.23846

[6] R. Andrews, C. G. van Dun, M. T. Wynn, W. Kratsch, M. K. Röglinger, and A. H. ter Hofstede, "Quality-informed semi-automated event log generation for process mining," *Decision Support Systems*, vol. 132, p. 113265, 2020.

[7] D. Sanchez-Charles, J. Carmona, and R. Raventos, "Building event logs from relational databases in the XES format," *Applied Sciences*, vol. 12, no. 21, p. 10832, 2022.

[8] C. Kecht, A. Egger, W. Kratsch, and M. Röglinger, "Event log construction from customer service conversations using natural language inference," in *3rd International Conference on Process Mining (ICPM)*, 2021.

[9] Ł. Korzeniowski and K. Goczyła, "Discovering interactions between applications with log analysis," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2022, pp. 861–869.

[10] C. W. Günther, A. Rozinat, and W. M. van der Aalst, "Activity mining by global trace segmentation," in *Business Process Management Workshops: BPM 2009 International Workshops, Ulm, Germany, September 7, 2009. Revised Papers 7*. Springer, 2010, pp. 128–139.

[11] N. Tax, N. Sidorova, R. Haakma, and W. M. van der Aalst, "Event abstraction for process mining using supervised learning techniques," in *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 1*. Springer, 2018, pp. 251–269.

[12] N. Martin, B. Depaire, and A. Caris, "The use of process mining in a business process simulation context: Overview and challenges," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 2014, pp. 381–388.

[13] W. M. van der Aalst, "Object-centric process mining: dealing with divergence and convergence in event data," in *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*. Springer, 2019, pp. 3–25.

[14] W. M. van der Aalst and A. Berti, "Discovering object-centric Petri nets," *Fundamenta informaticae*, vol. 175, no. 1-4, pp. 1–40, 2020.

[15] A. F. Ghahfarokhi, G. Park, A. Berti, and W. M. van der Aalst, "Ocel: a standard for object-centric event logs," in *European Conference on Advances in Databases and Information Systems*. Springer, 2021, pp. 169–175.

[16] M. J. Jans, M. Alles, and M. A. Vasarhelyi, "Process mining of event logs in auditing: Opportunities and challenges," *International Journal of Accounting Information Systems*, vol. 12, no. 1, pp. 1–20, 2011.

[17] A. Burattin, A. Sperduti, and W. M. van der Aalst, "Control-flow discovery from event streams," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, 2014, pp. 2420–2427.

[18] D. Fahland, "Extracting and pre-processing event logs," *CoRR*, vol. abs/2211.04338, 2022.

[19] W. M. van der Aalst, A. J. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[20] A. J. Weijters, W. M. van der Aalst, and A. de Medeiros, "Process mining with the heuristics miner algorithm," in *Technical Report WP*, vol. 166, 2003, pp. 1–34.

[21] S. J. van Zelst and W. M. van der Aalst, "Log skeletons: A classification approach to process discovery," in *International Conference on Advanced Information Systems Engineering (CAiSE)*. Springer, 2018, pp. 309–324.

[22] A. Rozinat and W. M. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Information Systems*, vol. 33, no. 1, pp. 64–95, 2008.

[23] A. Ilyas, J. M. da Trindade, R. Castro Fernandez, and S. Madden, "Extracting syntactical patterns from databases," in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*. ACM, 2018, pp. 1773–1788.

[24] A. Burattin and W. M. van der Aalst, "PLG2: Multiperspective process randomization with online and offline simulations," in *International Conference on Business Process Management*. Springer, 2015, pp. 214–219.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 243–253
DOI: 10.15439/2025F3612
ISSN 2300-5963 ACSIS, Vol. 43

Addressing Skills Shortages through Low-Code/No-Code: Skill Profiles and Implementation Challenges

Paul Rozbitski
0009-0008-5413-0165
Technische Hochschule Mittelhessen (THM) – University of Applied Sciences
THM Business School
Giessen, Germany
Email: paul.rozbitski@w.thm.de

Christian Leyh
0000-0003-0535-0336
Technische Hochschule Mittelhessen (THM) – University of Applied Sciences
THM Business School
Giessen, Germany
Email: christian.leyh@w.thm.de

Ching-Ho Ip
Technische Hochschule Mittelhessen (THM) – University of Applied Sciences
THM Business School
Giessen, Germany
Email: ching-ho.ip@w.thm.de

***Abstract*—Low-code/no-code (LCNC) platforms promise low-barrier access to software development, increased efficiency, and programming with little to no code. This paper explores whether it is really possible to create software without writing code and with little programming knowledge. Using a systematic literature analysis, we lay a foundation for future research on the skills needed by citizen developers using LCNC tools in corporate and public-sector contexts. Our analysis of 58 articles identifies three key skill categories: hard skills, such as programming and working with APIs; soft skills, such as communication and collaboration; and domain-specific knowledge. However, challenges exist in the upskilling process. Additionally, generative AI (GenAI) shows potential to expand LCNC capabilities. Nevertheless, the intersection of LCNC and public-sector transformation remains underexplored, highlighting the need for deeper studies focused on sector-specific challenges, use cases, and long-term adoption.**

***Index Terms*—Low-Code, No-Code, Citizen Developer, Skills, Innovations, Digital Transformation.**

I. Introduction

INFORMATION and communication technology (ICT) has undergone rapid changes over the decades—particularly in the last ten to fifteen years—and has had a far-reaching impact on almost every aspect of daily life. These changes and the associated digitalization have caused not only companies but also public institutions, organizations, and private individuals to increasingly shift their activities to the digital world. Industrial and commercial companies are increasingly utilizing the potential of e-commerce to tap into new market opportunities; public authorities are turning to e-government to reduce the administrative workload; and social interactions are increasingly shifting to the digital world of social networks. Even a large proportion of global financial transactions would be almost inconceivable without the use of ICT [1] As a result, digitalization and digital transformation are far more than just buzzwords: they have become decisive strategic competitive factors. Digitalization is fundamentally changing the economy and society and shaping the way we live and work in unprecedented ways [1], [2].

For companies, digitalization opens up the opportunity to develop and establish new business models. In order to actualize these new business models as well as to remain competitive in global markets, the majority of companies must deal with the challenges of digitalization and digital transformation. Relatedly—and also revealed by the global crises of recent years—a critical characteristic of successful companies has been their ability to adapt quickly to external circumstances, respond to external factors, and become a learning organization that quickly adopts innovations, with a focus not only on efficiency but also on resilience. Digitalization can be a key enabler of innovations and competitiveness; organizations that do not push ahead with digitalization risk falling behind, because they may struggle to adapt to changing market conditions and external factors [3], [4].

Digitalization has been a central topic in the private sector for many years and is now often considered an integral part of business operations. At the same time, digitalization is also of great importance for public administrations, because administrations are often in contact with citizens, work with many people, and link infrastructure; therefore, the efficiency of administration has an impact on cities and municipalities [5]. Public administrations also face digitization challenges and pursue the goal of improving their own processes. Specifically, there is a fundamental need for e-government services among the German population, as the range of services on offer is too limited: according to the Bitkom survey examining people's preferences regarding administrative services, 79% of respondents consider Germany to be lagging behind regarding digitalization, and 78% would like to see significantly more investment in the digitalization of public administration. The use of artificial intelligence (AI) is particularly in demand: 61% are in favor of public authorities using AI [6]. Accordingly, given these aspects, the focus of research on digitalization should not only be on companies but also on public institutions with numerous stakeholders.

In addition to the organizational perspective on the benefits of digitalization, there are also macroeconomic and labor-market implications. Automation and the implementation of software systems can reshape the labor market, leading to job

losses and the creation of new job profiles. The difference between the demand and the supply of qualifications leads to a skill mismatch; it remains uncertain whether there will be enough new jobs for those who have lost their jobs, and whether they will be able to fill these positions. Those with medium qualifications are particularly at risk [7]. One thing is certain: there will be "losers of automation"—i.e., people who are unable to find a job in the long term or who are forced to take underpaid jobs in other regions [8].

The effects of skill mismatch are already evident in Germany, where small and medium-sized enterprises (SMEs) and public authorities are struggling to recruit qualified professionals despite the fact that the number of registered unemployed people far exceeds the number of vacancies [9], [10]. In 2023, German SMEs reported significant difficulties in filling information technology (IT) positions, with 149,000 roles remaining unfilled—almost twice the number of vacancies recorded five years earlier. According to a Bitkom study examining the shortage of IT professionals across all sectors, Germany could face a shortage of around 663,000 IT professionals by 2040 [11]. Therefore, organizations must take action to avoid the problems associated with a shortage of skilled employees. Introducing personnel development measures could help bridge the gap between the qualifications on offer and those in demand [7].

Innovative approaches such as citizen development (CD) and low-code/no-code (LCNC) are well suited to this purpose. People without in-depth IT knowledge could use LCNC platforms to automate processes and develop applications. Through CD, companies can reduce the workload of IT departments by enabling employees in other departments to develop their own solutions, thereby reducing their dependence on IT professionals [12], [13]. CD and LCNC can also help drive and promote digitalization by making use of existing employees.

Given the current developments and the opportunities that LCNC can provide for companies and governmental institutions, this paper focuses on two research questions (RQs) that address the tension between the usage of LCNC and the necessary employee skills and competencies:

RQ1: What competencies are discussed in the literature in the context of LCNC platforms?

RQ2: What is the state of research on LCNC platforms in the public sector, and what skill profiles and implementation challenges are discussed in this context?

To answer these research questions, our paper proceeds as follows. After the introduction, a short theoretical background related to LCNC platforms is provided; this includes information about the typical users of these platforms, as well as their advantages and disadvantages. Thereafter, in Section 3, we describe our methodology (namely systematic literature review) to make our analysis reproducible. Section 4 presents the results of the literature review, highlighting the most discussed topics. In Section 5, we discuss the results of our analysis from different perspectives. Finally, the paper concludes with a summary and an outlook regarding potential further research steps.

II. Theoretical Background

A. Citizen Developers

Employees in specialist departments have in-depth domain knowledge and are familiar with the workflows in their respective areas. However, due to a lack of formal programming education and skills, they are often unable to develop digital solutions on their own. At the same time, IT departments that could take on such tasks are often overloaded [13], which leads to a backlog of innovation-related activities due to a focus on operational tasks [14].

In this context, the term citizen developers is employed to denote individuals who lack extensive expertise in IT yet employ platforms to automate processes and develop applications. These people are guided by platforms, which are toolboxes containing the necessary tools for creating customized applications.

B. Low-Code/No-Code Platforms

LCNC platforms are visual development environments that enable users to build software applications and automate workflows with little to no coding experience. The platforms provide users with a graphical user interface (GUI), where they can choose from available elements—such as buttons or application programming interface (API) calls—which can be visually combined through drag-and-drop into ready-to-use software [12].

While NC platforms require no programming knowledge, LC platforms require a little programming knowledge to make applications more flexible. With LC, existing modules can be extended with the user's own code [15].

Another important feature of LCNC platforms is that they not only provide support for visual programming and handle the entire application lifecycle, but they also typically provide integrated tools for the deployment process. They also can enable collaboration between software developers in terms of distributed work [16].

Transferring selected development work to citizen developers reduces the need for expensive IT expertise [17]. CD relieves the strain on scarce IT capacity, as implementations can be distributed across more resources. In particular, this reduces technical dependencies on IT departments; as a result, professional developers can focus on more strategic and value-adding tasks [18].

LCNC platforms combine declarative programming and model-driven development with visual interfaces, making software development more user-friendly and efficient [18]. These advantages have led to their high adoption among non-technical users looking to leverage their strengths. However, professional developers are skeptical about using LCNC platforms, which they consider to be inferior to traditional programming tools [19], [20].

C. Advantages and Challenges of LCNC Platforms

The advantages of LCNC platforms are well documented in the literature. One particularly interesting feature of such platforms is that they have a very low barrier to entry for non-programmers. This has led to the democratization of software development, whereby more people can develop, which in turn enables companies to benefit from new, fast prototypes and innovations [21], [19]. Furthermore, according to expert assessments, LCNC application development can in some cases be at least twice as fast as traditional code-based development [17]. Response times to customer feature requests are also reduced due to faster implementation and shorter delivery times and cycles. Such platforms also emphasize the high agility of development processes. Overall, using LCNC platforms can lead to increased efficiency and productivity, thus saving costs [17], [18], [12]. Additionally, technical training and CD initiatives form a key part of LCNC platforms: targeted training and education programs introduce employees to basic and advanced digital skills, thereby helping companies to boost the digital literacy of their entire workforce [18].

On the other hand, the literature also highlights some challenges associated with LCNC platforms. These include difficult-to-estimate licensing costs, which nevertheless can be justified by faster development processes and potential resource savings related to expensive IT specialists [17], [22]. A bigger challenge is vendor lock-in, which has been highlighted by several authors. Users are tied to the platform manufacturers of their chosen platform, making it difficult to transfer solutions to another platform and thus reducing flexibility [17], [18], [22]. Therefore, it is important to choose the right LCNC platform; otherwise, the investment costs will be too high. Companies experience the difficulty of choosing the right LCNC platform from dozens of providers. To this end, a systematic approach can be adopted to address this challenge [22].

The last aspect that is often discussed is the quality dimension. To start, there are always concerns about data protection and IT security. Although certain security standards can be met or even enhanced by LCNC platforms, a significant number of citizen developers lack the necessary knowledge to ensure secure development. Additionally, these platforms can be hard to debug, because the underlying code is typically generated by the platform's developers and is not directly accessible to users, making it difficult to identify and maintain issues [17]. Opinions on data protection vary: some see advantages, while others warn of risks, particularly when citizen developers create applications [20].

D. Institutionalization and the Rise of LCNC

According to reports by Gartner and Forrester, the market for LCNC is growing year on year. These market research companies predict that this market will continue to grow as a result of the ongoing digital transformation, which is accelerating the demand for software while at the same time leading to skill shortages in the IT sector. As outlined above, as new digital products and services emerge and new markets are developed, the demand for digital skills rises. In an effort to reduce costs and increase efficiency while meeting the growing demand for software, organizations are increasingly institutionalizing LC solutions within their IT departments [23], [24]. This means that LC solutions are used by IT companies to develop and deliver professional software solutions for their customers. This results in the acceleration of the distribution of LCNC. Therefore, the LCNC market will continue to grow and reach a volume of $50 billion by 2028. The fact that larger providers of LCNC platforms are being bought by corporations such as Siemens also indicates the relevance of this topic [25].

Given the growing importance of LCNC platforms, especially regarding organizational usage, it is important to understand which providers are currently available on the market.

E. Overview of LCNC Vendors

To classify technology providers, especially in markets that are rapidly changing and growing, Gartner has developed a methodology for categorizing IT providers into four distinct categories. The measurement of various factors—such as geographical strategy, marketing strategy, business model, and innovative strength—contribute to the assessment of "completeness of vision." Other criteria—such as customer experience, product or service quality, responsiveness to market changes, and operational performance—are used to measure the "ability to execute" [26]. By combining these two dimensions, Gartner categorizes vendors into four groups: "Leaders," "Challengers," "Visionaries," and "Niche Players."

- Leaders are highly competitive, have a strong market presence, possess a clear vision, demonstrate strong ability to execute, and are well positioned for the future. According to the Gartner report on LCNC from 2024, the following vendors are classified as Leaders: Mendix, OutSystems, Microsoft, Appian, ServiceNow, and Pegasystems.
- Challengers can also execute well today and may even dominate large market segments, but they do not have a clear vision of where the market is heading or they fail to respond to emerging trends. Challenger LCNC vendors include Oracle, Salesforce, and Zoho.
- Visionaries are companies with a clear vision of the future and an understanding of market trends, but they are not yet executing well or their products' features are not stable enough. The following vendors are classified as Visionaries: SAP and Retool.
- Niche players, such as Creatio and Newgen, focus successfully on a small part of the market or a specific geographical region. They may be a suitable choice for particular use cases or industries, but they do not yet match the breadth nor innovation of the other categories.

III. Methodology – Systematic Literature Review

Our literature review follows the approach of Webster and Watson [27]. The details of our methodology and the individual steps can be found in Fig. 1. The review consists of three main steps: literature search, topic identification, and analysis. In the first step, keywords and operators are used to identify the relevant literature. The second step involves identifying relevant topics. The final step consists of a detailed analysis of the literature and central topics. We selected relevant databases, defined the inclusion and exclusion criteria, documented the search process, and ultimately stored the identified literature in the Zotero literature management program

A. Phase 1 – Literature Search

We considered various databases and selected the following as sources: IEEE Xplore, ACM DL, and Springer Link. For titles, we used the search terms "low-code" and "no-code" along with additional keywords, such as "digital transformation" and "citizen developer," to achieve more precise results. Since the search functions in the individual databases are structured differently, we adapted the search fields depending on the database: in IEEE, we searched across all metadata; in ACM, within all abstracts; and in Springer Link, using the "with all the words" function.

The first phase of the literature search, which entailed a broad search, was carried out in November 2024. The time frame was set for 2019–2024 to ensure that the research was up to date. Sources with a high degree of relevance to the topic requirements were selected.

During the screening process, titles were first analyzed for their relevance. This was followed by in-depth reviews and summaries. Only German- and English-language sources were considered. Papers such as workshop documents, short articles, and articles without full-text access were excluded from the evaluation. After Zotero removed the duplicates, we were left with 151 articles, which provide a solid foundation for further research in the field of LCNC. However, in order to gain a better overview of certain topics, we narrowed the sample even further for our analysis.

The second step of the literature search involved conducting a targeted search guided by the research questions, particularly with regard to competencies (RQ1) and challenges in the public sector (RQ2). This step occurred in March 2025. The remaining articles were screened once more; in addition, further relevant articles were searched using backward and forward searches in accordance with the method described by [27]. We analyzed the papers until no new concepts emerged. In total, the steps in phase 1 led to the identification of 58 relevant papers that were analyzed in detail in the subsequent phases of our literature review. A complete list of these 58 papers is available upon request from the authors.

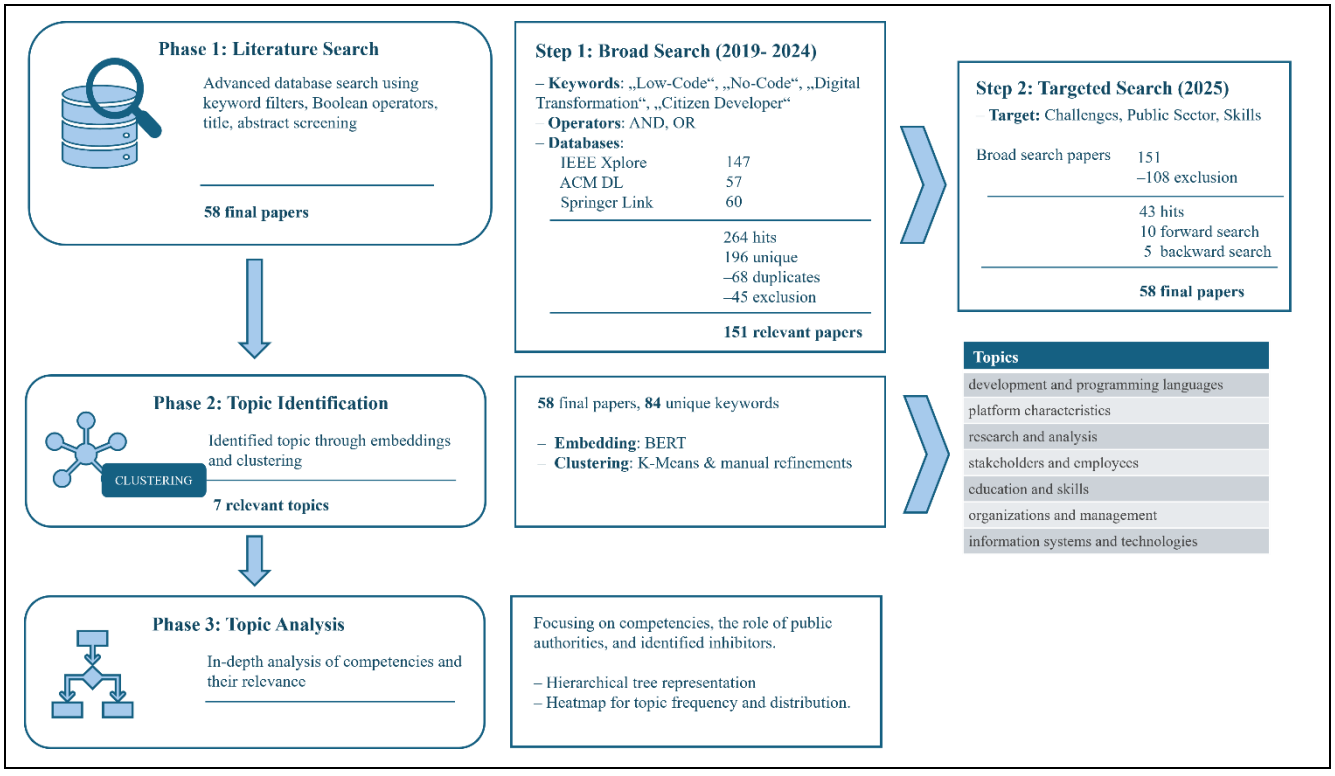


Fig. 1. Overview of the Three Steps of the Methodological Approach

B. Phase 2 – Topic Identification

During the screening of titles and abstracts, suitable keywords from the databases, as well as keywords identified manually, were entered for each article to enable subsequent classification. Recurring keywords were identified and organized by topic. However, these keywords had different levels of abstraction, which made further analysis difficult; therefore, an attempt was made to make the levels of abstraction more similar by combining some keywords into generic terms. As there were 84 keywords, an automatic clustering method was initially employed, which was then supplemented by our own review.

Python and the KNIME Analytics Platform were utilized for the data analysis, especially to analyze keywords, clusters, and frequencies, as well as to create tables. Additionally, we incorporated D3.js into our workflows for the purpose of displaying the tree structure of keywords. We also used the Seaborn Python data visualization library to create a heatmap. Token-level embeddings were created using the Bidirectional Encoder Representations from Transformers (BERT) bert-base-uncased model to evaluate the semantic similarity of keywords. BERT is a state-of-the-art language representation model that captures the contextual meaning of terms based on how they are used in natural language; although the model is frequently used to analyze entire sentences, it can also be used to understand specific keywords. As BERT does not require labeled data, it allows us to apply unsupervised clustering without the need for manual annotations. The classical k-means clustering algorithm was then applied, with varying numbers of clusters, to identify coherent term groups [28]. The analysis produced the best results with seven clusters, which were then manually checked and optimized.

C. Phase 3 – Topic Analysis

In order to gain more detailed insights into thematic trends, an analysis was conducted to determine which articles covered which topics. This made it possible to identify both gaps in content and potential trends. A heat map (see Fig. 2) was used for this purpose, with individual articles shown on the x-axis and the thematic clusters shown on the y-axis. The heat map depicts which topics dominate the literature and which are underrepresented. Each cell contains a number indicating how many aspects of a particular cluster are covered in the respective article. This facilitates the identification of thematic foci and research gaps at a glance.

Another tool used to represent the clusters is a hierarchical tree structure. This is particularly well suited for visualizing subordinate concepts and highlighting thematic connections, dependencies, and analysis paths. The tree displays the main topics at the top level, the subcategories at the second level, and the individual keywords (including the number of times they are mentioned in the analyzed papers) at the bottom level.

Finally, a concept-centered method based on Webster and Watson [27] was used to systematically synthesize the authors' findings and positions on specific, relevant concepts.

IV. Results

This section presents the results of the keyword and thematic analyses, which were based on a concept matrix comprising 58 articles on LCNC platforms. First, each article was assigned to relevant keywords, and then the keywords were grouped into overarching thematic clusters. In total, 84 unique keywords were extracted, with 518 mentions across all papers. Several keywords with the same meaning but different spellings were standardized.

The results of the literature review are presented in three visual formats: a heatmap, two tables, and a hierarchical tree diagram. The heatmap (see Fig. 2) illustrates the seven main thematic clusters and their distributions across the selected articles; the x-axis shows the unique article numbers in ascending order (i.e., from article 1 to article 58) and the y-axis shows the clusters, or subject areas, sorted by how often they were mentioned in articles. The frequency with which each cluster was mentioned is indicated by a number in the cell at the intersection of the cluster and paper number. Table I contains these clusters, the numbers of articles in which these topics were mentioned, the corresponding keywords, and the number of times each keyword appears in the papers. Table II shows the most and least frequently used keywords. The hierarchical tree diagram (see Fig. 3) groups keywords into thematic clusters based on their contextual relationships; each level has its own color to aid in distinguishing between them. In Fig. 3, we have chosen to omit certain individual clusters—i.e., platform characteristics, research analysis, and development and programming languages—so that the illustration fits into our article; however, the full illustration is available upon request.

A. Topic Coverage across Articles

As shown in Fig. 2 and in Table I, the seven thematic clusters are:

- #1 information systems and technologies,
- #2 development and programming languages,
- #3 education and skills,
- #4 platform characteristics,
- #5 organizations and management,
- #6 stakeholders and employees, and
- #7 research and analysis.

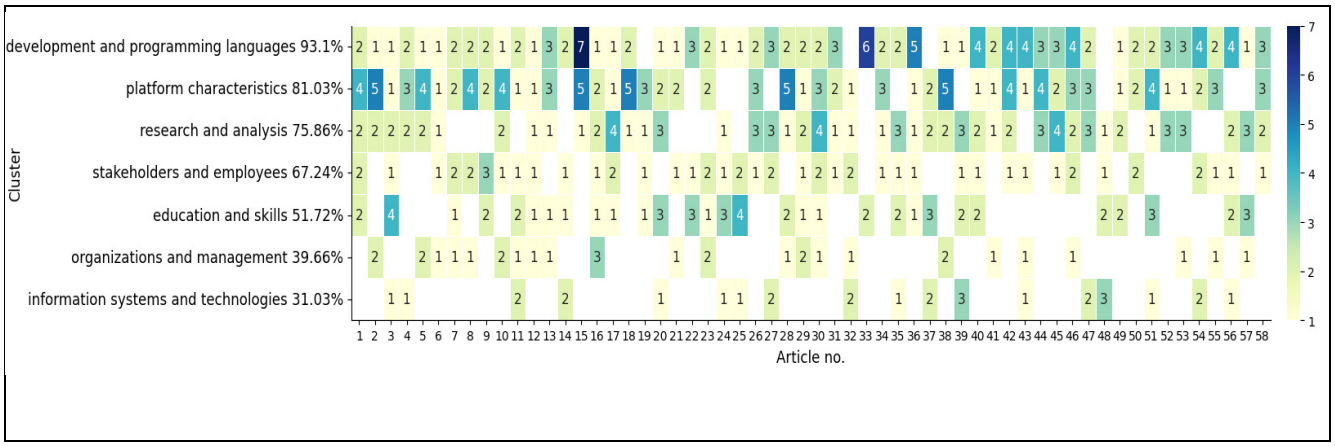


Fig. 2. Heatmap of Topic Coverage Across Articles

TABLE I.
TOPICS, ASSOCIATED KEYWORDS, AND FREQUENCY OF OCCURRENCE

No.	Cluster	Article Count	Keywords with Occurrences Count
#1	information systems and technologies	18	information systems (9), artificial intelligence (7), integration (5), data handling (4), IoT (2), connectors (1), machine learning (1)
#2	development and programming languages	54	low-code (51), software development (20), no-code (16), rapid application development (6), model-driven engineering (6), traditional development (5), programming languages (4), prototyping (3), domain-specific language (3), testing (2), devops (2), generators (1), mobile applications (1), digital process automation (1), paradigm (1), visual programming (1), automation (1), design to code (1)
#3	education and skills	29	skills (21), education (12), skills shortage (12), re- and upskilling (5), learning (5), experiential learning (1), authentic learning (1)
#4	platform characteristics	47	challenges (30), platform (21), drivers (19), fundamentals (9), forecast (8), adoption (7), security (5), recommendations (4), quality (4), inhibitors (3), classification (3), taxonomy (2), features (2), architecture (1), technical roles (1)
#5	organizations and management	24	companies (10), organization (4), governance (3), management (2), job crafting (2), work system (2), resource demand theory (1), business unit development (1), project management (1), agile methodology (1), socio-technical theory (1), decision making (1), manufacturing (1), public administrations (1), paradox theory (1), product line (1)
#6	stakeholders and employees	39	democratization (21), citizen developer (18), innovations (7), employees (3), employee-driven innovation (2), stakeholders (1), job roles (1)
#7	research and analysis	44	conceptional (36), use case (22), survey (8), experiment (6), interviews (4), analysis (2), research (2), group discussion (2), metamodeling (2), task analysis (1), empirical study (1), expert interviews (1), hackathon (1), model (1)

A notable finding is that *development and programming languages* (cluster #2) dominates, with a frequency of 93.1%, followed by *platform characteristics* (cluster #4) with 81.03%. Cluster #7, *research and analysis,* shows which articles explicitly mention the type of method or research approach used (75.86%). The clusters *education and skills* (#3, 51.72%), *organizations and management* (#5, 39.66%), and *information systems and technologies* (#1, 31.3%) are significantly less represented.

Table I illustrates the distribution of the 58 articles across the seven theme clusters; for each cluster, the table shows the number of articles in which each topic is addressed. The corresponding keywords are listed in descending order of frequency (provided in parentheses). The number of keywords per topic vary: across all topics, *development and programming languages* (#2) contain the highest number of unique keywords (18), while *education and skills* (#3), *information systems and technologies* (#1), and *stakeholders and employees* (#6) include only 7 keywords.

As shown in Table I, *development and programming languages* (#2) is the most frequently addressed cluster, with 54 mentions across the articles. The most frequently mentioned keywords in this cluster are “low-code” (51), “software development” (20), and “no-code” (16). This cluster and its keywords are followed by *platform characteristics* (#4, 47 articles); the most frequent keywords here are “challenges” (30), “platform” (21), and “drivers” (19). The third most represented cluster is *research and analysis* (#7), with 44 mentions across the articles. The most frequent keywords in this cluster are “conceptual” (36), “use case” (22), and “survey” (8). The least addressed clusters are *organizations and management* (#5, 24 articles) and *information systems and technologies* (#1, 18 articles).

Fig. 3 provides an overview of our classification structure through a hierarchical tree representation of topics and keywords. The seven thematic clusters form the basic components for either keywords or subcategories, as certain keywords are connected to more specific subtopics. For example, *education and skills* (#3) has been divided into the "skill demand" subcategory—with the keywords "skills", "skill shortage", and "re- and upskilling"—and the subcategory "learning approaches"—with the keywords "experimental learning", "learning" and "authentic learning"—as well as the keyword "education". The cluster *platform characteristics* (#4) relates to fundamental platform properties, such as "architecture" and "features", and comprises the subcategory "categorization" with the keywords "classification" and "taxonomy". Other clusters have a similar structure of branches, subcategories, and keywords.

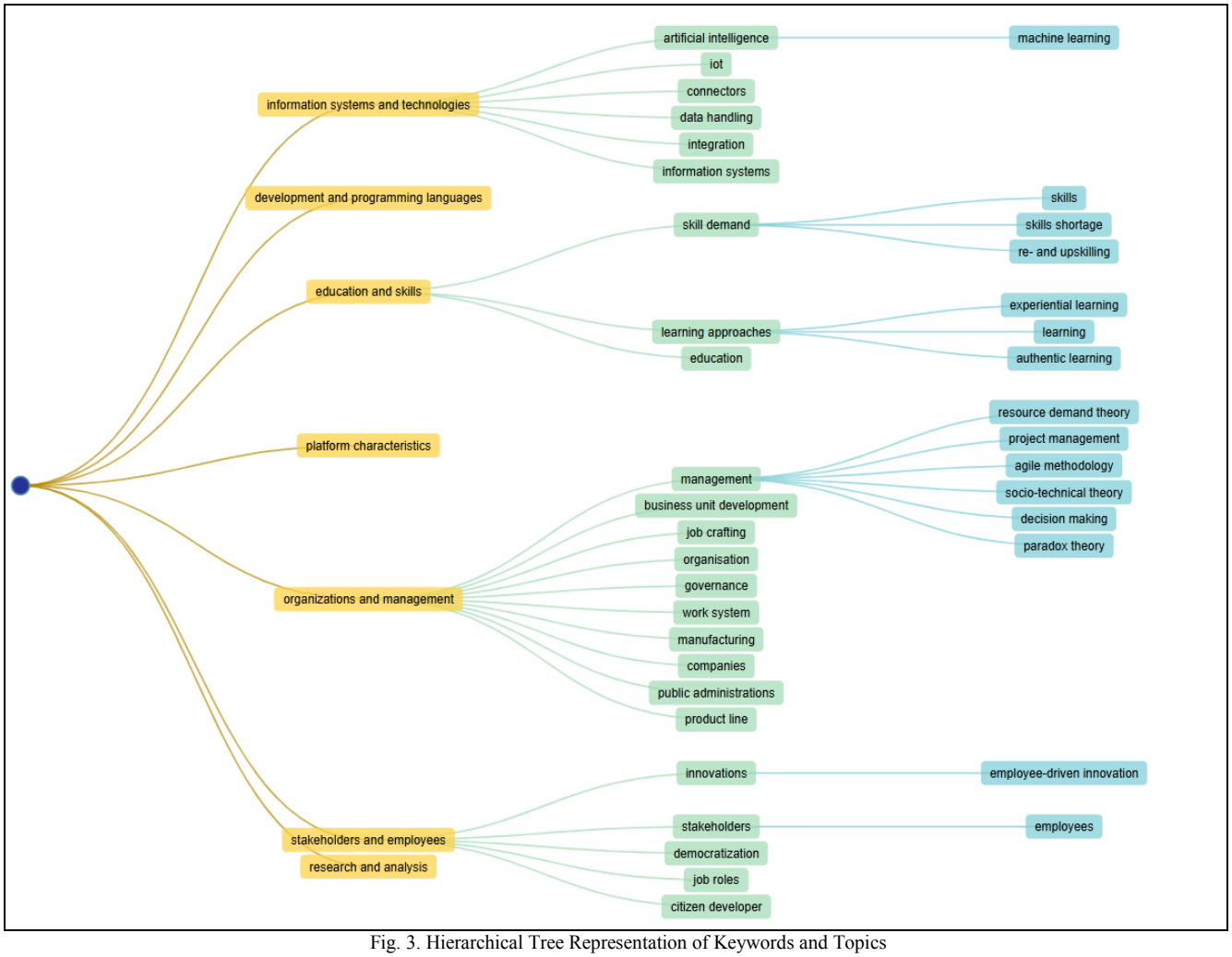


Fig. 3. Hierarchical Tree Representation of Keywords and Topics

B. Keywords' Frequency

A keyword analysis was conducted to better elucidate the thematic focus of the selected literature. An overview of the 15 most- and least-frequently occurring keywords found in the selected publications is shown in Table II. Notably, approximately 35.7% of all keywords were mentioned only once.

Interestingly, the keyword "skills" first appeared in 2021, after which it has continuously increased (2021=2, 2022=5, 2023=6, 2024=10, 2025 (January–May)=3). Notably, the related term "re- and upskilling" did not appear until 2022. "Low-code" was mentioned significantly more often than "no-code", with 51 mentions compared to 16. Frequently discussed topics included challenges related to LCNC, platform characteristics, software development drivers, and information systems—keywords with strong technical focus. The keywords "citizen development" (18), "democratization (21)", and "skills" (21) also appeared frequently, as did "education" and "skills shortage" (both with 21 mentions).

In contrast, non-technical aspects were rarely mentioned. For example, terms like "agile methodology" and "project management" appeared only once; notably, "public administration" and "manufacturing" were also mentioned only once each.

TABLE II.
MOST AND LEAST FREQUENTLY USED KEYWORDS IN THE ANALYZED LITERATURE

Most frequently occurring keywords		Least frequently occurring keywords	
Keyword	**Frequency**	**Keyword**	**Frequency**
low-code	51	project management	1
conceptual	36	agile methodology	1
challenges	30	socio-technical theory	1
use case	22	decision-making	1
skills	21	manufacturing	1
platform	21	public administration	1
democratization	21	paradox theory	1
software development	20	product plan	1
drivers	19	stakeholders	1
citizen development	18	job roles	1
no-code	16	task analysis	1
education	12	empirical study	1
skills shortage	12	expert interviews	1
companies	10	hackathon	1
information systems	9	model	1

V. DISCUSSION

A. Overview of General Findings

An important observation concerning the results was that the two most frequent clusters were technical. The strong concentration of keywords in these clusters indicates that most of the literature focuses on the technical foundations of LCNC platforms. This finding aligns with previous literature reviews in this context, which also found that the majority of papers concentrate on the technical aspects of LCNC platforms [20]. Less prevalent terms—particularly regarding skills, organizational context, and emerging technologies—highlight areas that may be underexplored, suggesting potential directions for further research.

The cluster research and analysis show that conceptual approaches and use cases are particularly common. This indicates that current research is primarily focused on developing fundamental ideas and describing specific applications. More empirical investigations would be useful for obtaining a more comprehensive picture: while many works present practical examples (i.e., use cases), systematic studies that quantify these experiences are much rarer.

Most articles include keywords from multiple clusters, reflecting the interdisciplinary nature of LCNC research. On average, each article contributes to 4.4 clusters, indicating that LCNC platforms are examined from multiple perspectives simultaneously.

B. Competencies, Skills, Education (RQ1)

The skills topic has gained noticeable momentum in recent years, reflecting the growing relevance of competency-related aspects in the context of LCNC platforms.

In general, the skills and competencies of citizen developers can be classified into two areas: hard skills and soft skills [29]. Elshan et al. [30] add a third dimension of business or domain expertise (which we combine with soft skills in the article).

Many training programs are aimed at professional developers. In contrast, citizen developers would benefit more from "knowledge nuggets" [29], which address specific knowledge deficiencies without overwhelming citizen developers. Knowledge nuggets are small, focused learning units that fill specific topic or skill gaps in a concise and accessible format. Citizen developers certainly require technical competencies, but only those relevant to their specific fields of application, and these can be conveyed in the form of courses, presentations, and short videos. It is important to ensure that these materials are tailored to citizen developers. Currently, most offerings still focus on professional software developers, and they unfortunately rarely meet the needs of citizen developers, who need concrete, relevant, bite-sized pieces of knowledge for their development [22]. Accordingly, proper didactic concepts should be taken into account [29].

Although LCNC minimizes entry barriers to software development, employees still need some technical skills to work with LCNC efficiently. The most commonly identified hard skill is basic programming knowledge [30], [29]: authors emphasize proficiency in LCNC, data management and analytics, familiarity with integration technologies and APIs, general IT knowledge, an understanding of technical requirements, knowledge of business processes (domain), and design knowledge.

As mentioned by Bernsteiner et al. [29], individuals with strong technical skills often still fail in LCNC projects due to the lack of soft skills. One of the key responsibilities of citizen developers is coordinating and communicating with the various stakeholders involved in the process. Therefore, interpersonal skills and the ability to communicate between multiple stakeholders are crucial for citizen developers. To this end, reskilling and upskilling initiatives should not focus only on technical aspects, but also on soft skills. Another four important soft skills were also emphasized: problem-solving, communication and collaboration, adaptability and flexibility, and initiative and self-motivation. In contrast to professional developers, citizen developers benefit especially from skills like team understanding, user comprehension, and creativity, as well as business skills including domain expertise, an understanding of business processes, and project management [30].

In response to RQ1:

The literature identifies three main areas of competence for citizen developers: *hard skills*, such as IT knowledge, programming (e.g., basic programming skills), and working with integrations and API; *soft skills*, such as communication and collaboration; and *domain-specific business knowledge*. Technical, communication, and domain-specific skills are equally important for implementing CD. Difficulties arise with the learning process: it is not particularly limited, but citizen developers should be taught differently than IT professionals. To this end, learning processes based on targeted knowledge nuggets can help close gaps relevant to their work context.

C. Governmental Authorities (RQ2)

The keyword distribution shows that the dominant discourse is centered around technical foundations, skills, and the broad application of LCNC platforms in the private sector, even though the search was not limited to this domain.

Of the articles on LCNC platforms, only one addresses the public sector directly: Gialitakis et al. [31]. This study highlights the potential for the broader use of LCNC tools in the digitalization of bodies of government authority, and it suggests wider application to demonstrate its feasibility. In that article, the researchers analyzed the characteristics of several low-code development platforms available on the market, and they pointed to well-known technical challenges that are not specific to the public sector. They limited their focus to vendor lock-in, lack of interoperability with other systems, lack of flexibility in custom logic or integrations, and security concerns. Subsequently, a system that supports business process model and notation (BPMN) was selected, and a five-step approach was conceptualized. The proposed methodology—spanning BPMN modeling to database design, user interface creation, optional scripting in procedural language/structured query language (PL/SQL), and the integration of process models into the application—was devised for use in both business and governmental contexts, with the objective of facilitating the implementation of applications for workflow-based processes. To demonstrate the practical application of the concept, the authors implemented a case study: following consultation with the Greek government, the researchers selected a public service process (specifically, the reimbursement of travel expenses) for digitization, which they subsequently undertook. The researchers demonstrated that the implementation of such an application was not only feasible with LCDP but also that the digital process now took one to three days, in comparison to twenty days for the in-person process.

The objective of the present study was to demonstrate the feasibility of implementing public service processes in low-code development platforms, with the goal of highlighting the potential for developing process-driven applications in this domain. This approach has been demonstrated to reduce the need for extensive coding methods and is accessible to individuals without formal computer science degrees. Consequently, this approach has the potential to facilitate digital transformation in public administrations by streamlining developer cycles, minimizing costs, and leveraging citizen-centric design to obtain feedback from users. This, in turn, could result in more effective public services.

Nevertheless, despite the clear advantages of this approach, it is important to acknowledge its limitations. It is clear that many real-world processes involve a high level of complexity, largely due to the participation of numerous organizations. This complexity creates challenges with process modeling. A further issue that must be addressed is the synchronization of models and tables in databases. The process of database synchronization is a prerequisite for implementing model changes; however, this procedure is not without its complications.

Overall, governmental authorities remain less represented in the literature. Further studies are necessary to identify sector-specific barriers, skill requirements, and implementation strategies in public administrations.

In response to RQ2:

Research on governmental administrations is limited; only one article was identified. The main challenges were not explicitly mentioned—only the standard limitations of LCNC platforms, such as issues with interoperability, flexibility, and vendor lock-in. Competency profiles were also not identified. Further research is needed in this area.

D. Future of LCNC Platforms

The low coverage of the cluster information systems and technologies (#1) may be due to the inclusion of newer emergent topics in the LCNC context, such as AI, machine learning (ML), and the internet of things (IoT). AI has recently gained attention in the context of LCNC platforms and tools, as the integration of AI into LCNC could potentially expand the latter's usage and help overcome issues such as lack of flexibility, inadequate documentation, and technical constraints, thereby leading to synergies [32], [33].

Generative AI (GAI) can support the work of citizen developers in LCNC platforms, and automating certain tasks may compensate for missing technical skills [34]. This is seen as a way to address the technical limitations and skills gaps that LCNC alone cannot overcome.

It is also notable that "low-code" is mentioned far more often than "no-code." This may be because the use of NC platforms is more limited compared to that of LC platforms [30]. However, recent conceptual work suggests that this distinction may blur in the future: for example, [35] suggests a conceptual model which integrates large language models (LLMs) into LC tools, bringing them closer to true NC solutions.

LCNC is also occasionally discussed in the educational sector, particularly in connection with modern learning methods. It is interesting to note that within the skills cluster, LCNC is occasionally discussed in the context of modern learning methods: students can learn how AI works and how to deal with data in practice, guided through the ML workflow [36].

Implication:

Emerging technologies such as AI, ML, and IoT were underrepresented in the context of LCNC and CD research. Studies are currently being conducted in the area of merging GAI with LCNC to support and simplify CD. There are also practical applications in education, where AI supports both teaching and student learning in IT. These are promising areas for future research.

VI. Conclusion and Limitations

This study presents a systematic literature review of the current research on LCNC platforms, with a particular focus on skills and governmental authorities. The analysis reveals that the examined literature primarily addresses technical aspects, while organizational and educational dimensions are less frequently examined, although the topic of skills has gained greater visibility in recent years. This study indicates that the competencies required for citizen developers are generally categorized, but challenges remain in adapting training methods to the specific needs of citizen developers.

The use of LCNC tools in the context of governmental administration is seldom discussed, and sector-specific barriers are insufficiently explored. The integration of GAI into LCNC platforms may help reduce technical barriers and support the empowerment of citizen developers. Nonetheless, further (empirical) research is necessary to validate these concepts. Future research could focus on the use of LCNC platforms in government settings and the development of upskilling strategies for employees.

We are aware that this study has several limitations, which should be considered when interpreting the results. For example, we did not identify all relevant papers published in journals and conference proceedings, because we made our selection from specific databases. Other journals and conference proceedings not included in the databases we examined might also contain relevant articles, and expanding the analysis to include additional databases could have produced different results and a different number of hits. A literature review cannot provide a comprehensive overview of a topic, particularly when the focus is limited to three databases, and so this may not have provided sufficient coverage of the topic. Therefore, a broader search using more databases, and particularly interdisciplinary ones, could highlight further limitations of our research and findings. Additionally, the focus on LCNC in the context of digital transformation may have led to some bias. Another limiting factor is that although a technically-driven clustering approach was used, assigning topics to clusters is still somewhat subjective, as some aspects may overlap between categories depending on the context. Additional clustering based on abstracts or keywords could improve the analysis further and reduce bias. Last but not least, we are aware that the study has temporal limitations, as the relevance and accuracy of the findings may change over time due to the quick speed of development in this field.

VII. References

[1] R. Hentschel, and C. Leyh, "Cloud Computing: Status quo, aktuelle Entwicklungen und Herausforderungen," in *Cloud Computing*, S. Reinheimer, Ed. Wiesbaden, Germany: Springer, 2018, pp. 3–20. https://doi.org/10.1007/978-3-658-20967-4_1..

[2] C. Leyh, and T. Schäffer, "Digital Competencies as a Necessary Foundation for Digital Transformation," *HMD Prax. Wirtsch.*, vol. 61, no. 1, pp. 12–26, Feb. 2024, https://doi.org/10.1365/s40702-024-01044-9

[3] J. A. R. Awad, and R. Martín-Rojas, "Digital transformation influence on organisational resilience through organisational learning and innovation," *J. Innov. Entrep.*, vol. 13, article no. 69, Oct. 2024, https://doi.org/10.1186/s13731-024-00405-4.

[4] C. Leyh, T. Schäffer, K. Bley, and S. Forstenhäusler, "SIMMI 4.0 – A Maturity Model for Classifying the Enterprise-wide IT and Software Landscape Focusing on Industry 4.0," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS 2016)*, 2016, http://dx.doi.org/10.15439/2016F478.

[5] A. Lorenz, P. Rozbitski, C. Leyh, and N. Madeja "From traffic dynamics to smart city solutions: Identifying key elements for data-driven traffic management," in *Proceedings of the 2025 Americas Conference on Information Systems (AMCIS 2025)*, 2025.

[6] Bitkom e.V, "Digitale Verwaltung: Menschen wollen Anliegen online erledigen [Digital administration: Most people want to handle their concerns online, in German]," Bitkom e.V., 2024. Accessed: May 24, 2025. [Online]. Available: https://www.bitkom.org/Presse/Presseinformation/Digitale-Verwaltung-Menschen-wollen-Anliegen-online-erledigen.

[7] H. Troger, "Auswirkungen auf den Arbeitsmarkt," in *7 Erfolgsfaktoren für wirksames Personalmanagement*, Wiesbaden, Germany: Springer, 2019, pp. 19–53. https://doi.org/10.1007/978-3-658-24437-8_2.

[8] A. Spermann, "Beendet die Digitalisierung das deutsche Jobwunder?," in *Megatrends aus Sicht der Volkswirtschaftslehre*, L. Rebeggiani, C. B. Wilke, and M. Wohlmann, Eds. Wiesbaden, Germany: Springer, 2020, pp. 215–230. https://doi.org/10.1007/978-3-658-30129-3_11.

[9] Statista, "Annual average unemployment rate in Germany from 2005 to 2025," Statista, 2025. Accessed: May 24, 2025. [Online]. Available: https://www.statista.com/statistics/227005/unemployment-rate-in-germany/.

[10] Statista, "Bestand von allen der BA gemeldeten offenen Arbeitsstellen in Deutschland im Jahresdurchschnitt von 2013 bis 2025," Statista, 2025. Accessed: May 24, 2025. [Online]. Available: https://de.statista.com/statistik/daten/studie/2903/umfrage/jahresdurchschnittswerte-des-bestands-an-offenen-arbeitsstellen/.

[11] Bitkom e.V, "Mangel an IT-Fachkräften droht sich dramatisch zu verschärfen," Bitkom e.V., 2024. Accessed: May 24, 2025. [Online]. Available: https://www.bitkom.org/Presse/Presseinformation/Mangel-an-IT-Fachkraeften-droht-sich-zu-verschaerfen.

[12] C. Chuanjian, G. Shuze, and W. Hua, "Research on Software Development Based on Low-Code Technology," in *Proceedings of the 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, 2023, https://doi.org/10.1109/AIARS59518.2023.00049.

[13] H. ALSAADI, D. Radain, M. ALZAHRANI, W. Alshammari, D. ALAHMADI, and B. Fakieh, "Factors that affect the utilization of low-code development platforms: survey study," *Rev. Rom. Informatică Și Autom.*, vol. 31, no. 3 pp. 123–140, Sep. 2021, http://dx.doi.org/10.33436/v31i3y202110.

[14] D. Hoogsteen and H. Borgman, "Empower the Workforce, Empower the Company? Citizen Development Adoption," in *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS 2022)*, 2022.

[15] Y. Luo, P. Liang, C. Wang, M. Shahin, and J. Zhan, "Characteristics and Challenges of Low-Code Development: The Practitioners' Perspective," in *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '21)*, 2021, https://doi.org/10.1145/3475716.3475782.

[16] D. Di Ruscio, D. Kolovos, J. de Lara, A. Pierantonio, M. Tisi, and M. Wimmer, "Low-code development and model-driven engineering: Two sides of the same coin?," *Softw. Syst. Model.*, vol. 21, no. 2, pp. 437–446, Apr. 2022, https://doi.org/10.1007/s10270-021-00970-2.

[17] C. Baumgarten, R. Endl, and S. Stich, "Professionelle Softwareentwicklung mit Low Code optimieren – eine Fallstudie," *HMD Prax.*

Wirtsch., vol. 61, no. 5, pp. 1213–1234, Oct. 2024, https://doi.org/10.1365/s40702-024-01095-y.

[18] E. Elshan, and B. Binzer, "Mehr als ein Trend?: Wie Low-Code die digitale Transformation unterstützt," *HMD Prax. Wirtsch.*, vol. 61, no. 5, pp. 1070–1087, Oct. 2024, https://doi.org/10.1365/s40702-024-01107-x.

[19] A. Viljoen, E. N. Altın, A. Hein, and H. Krcmar, "Beyond Citizen Development: Exploring Low-Code Platform Adoption by Professional Software Developers," *Proceedings of the 2024 Americas Conference on Information Systems (AMCIS 2024)*, 2024.

[20] S. Käss, S. Strahringer, and M. Westner, "Drivers and Inhibitors of Low Code Development Platform Adoption," in *Proceedings of the 2022 IEEE 24th Conference on Business Informatics (CBI 2022)*, 2022, https://doi.org/10.1109/CBI54897.2022.00028.

[21] F. Temiz, "Der Einsatz von No-Code-/Low-Code-Plattformen für die Beschleunigung und Unterstützung der Digitalen Transformation," in *Praxishandbuch Kompetenzen in der Digitalen Transformation der Arbeit*, P. Bruns and W. Bruns, Eds. Wiesbaden, Germany: Springer, 2024, pp. 121–131. https://doi.org/10.1007/978-3-658-44084-8_8.

[22] A. Nikolenko, K. L. Becker, U. Wohlhage, B. Adrian, and S. Hinrichsen, "Auswahl und Einführung einer Low-Code-Plattform," in *Prozesse in Industriebetrieben mittels Low-Code-Software digitalisieren : Ein Praxisleitfaden*, S. Hinrichsen, S. Sauer, and K. Schröder, Eds. Berlin, Heidelberg, Germany: Springer, 2023, pp. 31–51. https://doi.org/10.1007/978-3-662-67950-0_3.

[23] J. Bratincevic, R. Taylor, and Z. Stone, "The Low-Code Market Could Approach $50 Billion By 2028," Forrester, 2024. Accessed: May 23, 2025. [Online]. Available: https://www.forrester.com/blogs/the-low-code-market-could-approach-50-billion-by-2028/

[24] Gartner, "Gartner Forecasts Worldwide Low-Code Development Technologies Market to Grow 20% in 2023," Gartner, 2022. Accessed: May 24, 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2022-12-13-gartner-forecasts-worldwide-low-code-development-technologies-market-to-grow-20-percent-in-2023

[25] J. Rymer, "Siemens Snaps Up Mendix; Low-Code Platforms Enter New Phase," Forrester, 2018. Accessed: Apr. 29, 2025. [Online]. Available: https://www.forrester.com/blogs/siemens-snaps-up-mendix-low-code-platforms-enter-new-phase/

[26] O. Matvitskyy, K. Davis, and A. Jain, *Magic Quadrant for Enterprise Low-Code Application Platforms*. Gartner, Inc., Jul. 12, 2025. [Online]. Available: https://www.gartner.com/doc/re-prints?id=1-2J40TC35&ct=241017&st=sb

[27] J. Webster, and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Q.*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[28] Y. S. Soekamto, I. Maryati, C. Christian, and E. Kurniawan, "Improving Publishing: Extracting Keywords and Clustering Topics," *J. Appl. Data Sci.*, vol. 5, no. 2, article no. 2, May 2024, https://doi.org/10.47738/jads.v5i2.199.

[29] R. Bernsteiner, S. Schlögl, C. Ploder, T. Dilger, and F. Brecher, "Citizen vs. Professional Developers: Differences and Similarities of Skills and Training Requirements for Low Code Development Platforms," in *Proceedings of the 15th annual International Conference of Education, Research and Innovation (ICERI 2022)*, 2022, https://doi.org/10.21125/iceri.2022.1036.

[30] E. Elshan, B. Binzer, and T. J. Winkler, "From Software Users to Software Creators: An Exploration of the Core Characteristics of the Citizen Developer Role and the Related Re- and Upskilling Programs," *Bus. Inf. Syst. Eng.*, vol. 67, no. 1, pp. 31–53, Feb. 2025, https://doi.org/10.1007/s12599-024-00915-1.

[31] M.-K. Gialitakis, G. Tsakalidis, N. Nousias, and K. Vergidis, "Rapid prototyping of process-driven applications using low-code development platforms: A case study from the Greek public sector," in *Proceedings of the 2024 International Conference Automatics and Informatics (ICAI 2024)*, 2024, https://doi.org/10.1109/ICAI63388.2024.10851543.

[32] G. Paliwal, A. Donvir, P. Gujar, and S. Panyam, "Low-Code/No-Code Meets GenAI: A New Era in Product Development," in *Proceedings of the 2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM)*, 2024, https://doi.org/10.1109/ETCM63562.2024.10746160.

[33] J. Martins, F. Branco, and H. Mamede, "Combining low-code development with ChatGPT to novel no-code approaches: A focus-group study," *Intell. Syst. Appl.*, vol. 20, article no. 200289, Nov. 2023, https://doi.org/10.1016/j.iswa.2023.200289.

[34] S. Leible, D. Simic, G.-L. Gücük, and C. von Brackel-Schmidt, "Citizen Developer als Katalysatoren für die Entwicklung mitarbeitergetriebener Innovationen," *HMD Prax. Wirtsch.*, vol. 61, no. 5, pp. 1088–1114, Oct. 2024, https://doi.org/10.1365/s40702-024-01102-2.

[35] N. Hagel, N. Hili, and D. Schwab, "Turning Low-Code Development Platforms into True No-Code with LLMs," in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, 2024, https://doi.org/10.1145/3652620.3688334.

[36] L. Sundberg, and J. Holmström, "Teaching Tip: Using No-Code AI to Teach Machine Learning in Higher Education," *J. Inf. Syst. Educ.*, vol. 35, no. 1, pp. 56–66, Mar. 2024, https://doi.org/10.62273/CYPL2902.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 255–263
DOI: 10.15439/2025F8791
ISSN 2300-5963 ACSIS, Vol. 43

Keypoint-based metric for evaluating image super-resolution quality

Jakub Sadel*†, Tomasz Tarasiewicz*†, Pawel Kowaleczko†‡, Maciej Ziaja*†,
Daniel Kostrzewa*†, Pawel Benecki*, Michal Kawulok*†
0009-0006-9227-7470, 0000-0002-7706-1317, 0000-0003-4689-6343, 0009-0009-8285-5881,
0000-0003-2781-3709, 0000-0003-4674-5393, 0000-0002-3669-5110
*Silesian University of Technology, Gliwice, Poland
Email: {jakub.sadel | tomasz.tarasiewicz | maciej.ziaja | daniel.kostrzewa | pawel.benecki | michal.kawulok}@polsl.pl
†KP Labs, Gliwice, Poland
‡Warsaw University of Technology, Warsaw, Poland
Email: pawel.kowaleczko@pw.edu.pl

***Abstract*—Recent advances in single- and multi-image super-resolution have revealed the limitations of classical image similarity metrics (like peak signal-to-noise ratio), as they often fail to align with human perception when evaluating the visual quality of super-resolved outputs. In this paper, we explore how to exploit keypoint-based metrics to evaluate super-resolution image quality. Specifically, we explore two correlated metrics: (*i*) a *multi-scale index proposal* measure capturing salience of keypoints, and (*ii*) a *repeatability* metric quantifying how consistently the corresponding keypoints are identified in super-resolved and ground-truth images. Experiments on several simulated and real-world datasets show that the repeatability correlates with subjective judgments, and multi-scale index proposal can be helpful for difficult datasets when other metrics are insufficient.**

I. Introduction

LOW-RESOLUTION (LR) images pose a significant challenge in diverse imaging fields, from satellite surveillance to medical diagnostics. Physical constraints or adverse capture conditions often lead to low-quality data with insufficient detail. To address this, image *super-resolution* (SR) techniques reconstruct *high-resolution* (HR) images from LR inputs, effectively enhancing visual clarity and texture.

In recent years, deep learning-based SR methods have achieved impressive results, producing images with sharper edges and more natural details than classic interpolation schemes. However, these advances introduce a pressing need for robust *image quality assessment* (IQA) metrics tailored to super-resolved images. Traditional *full-reference* (FR) metrics such as *peak signal-to-noise ratio* (PSNR) and *structural similarity index measure* (SSIM) compare the SR output against a ground-truth reference at the pixel or local structural level. While useful for measuring basic fidelity, these methods often do not capture the subjective quality that humans perceive—especially in SR outputs that introduce plausible, but non-identical, high-frequency details.

This work was supported by the National Science Centre, Poland, under Research Grant 2022/47/B/ST6/03009. JS was supported by the SUT Statutory Research funds under Grant 02/080/BKM25/0058. MK was supported by the SUT funds through the Rector's Research and Development Grant 02/080/RGJ25/0053.

Ultimately, human viewers prioritize perceptual realism, focusing on sharp textures, coherent structures, and absence of distracting artifacts. An SR image that looks over-smoothed or contains noticeable distortions in critical regions may be subjectively worse, even if it has a high value of PSNR or SSIM metric. Conversely, an image that retains the essential scene layout while adding visually consistent fine details can be perceived as high quality despite larger pixel-wise differences from the ground truth. This observation drives the development of alternative IQA methods that better reflect human visual preferences and the complex trade-offs between fidelity and realism in SR.

The primary objective of this study is to explore keypoint-based metrics as a means of evaluating the quality of super-resolved images. Unlike traditional IQA metrics that focus on pixel-wise fidelity or perceptual similarity, our approach investigates the geometric and structural consistency of SR output through the lens of local feature detection. Keypoint-based metrics were also exploited for that purpose in [1]; however, that approach relies on conventional techniques that are non-differentiable, which limits its applicability for training deep neural networks. In this work, on the other hand, we aim at adapting methods based on deep learning, in order to allow for applying them to task-driven training in the future. To achieve this, we leverage the capabilities of Key.Net [2], a neural network designed for keypoint detection tasks that combines hand-crafted and learned convolutional filters within a shallow multiscale architecture. In particular, we examine the utility of the repeatability, a widely adopted metric in local feature evaluation [3], which measures the consistency of keypoint detection under various transformations, and the *multi-scale index proposal* (MSIP), introduced as a loss function for training Key.Net [2]. We extend this idea to the SR domain, under the assumption that a high-quality super-resolved image should exhibit similar keypoint number and distributions to the ground-truth reference image.

The paper is organized as follows. Section II reviews related work on SR methods and quality assessment metrics. Section III describes the datasets used for the evaluation.

Section IV details the proposed keypoint-based approach and metric configurations. Section V presents the experimental setup and results. Section VI summarizes the paper and provides directions for future research.

II. Related Work

A. SR Techniques: Major Categories and State of the Art

SR is highly relevant in domains where capturing higher-resolution imagery directly is impractical due to physical or cost-related constraints [4]. Applications span security surveillance, medical imaging, and remote sensing [5]. For instance, in satellite imaging, SR algorithms may be the only way to obtain finer details when sensor resolution is limited or hiqh-quality image acquisition is too expensive [4]. Over the past two decades, a vast array of SR methods have been proposed, generally categorized into *single-image SR* (SISR) approaches (relying on a single LR input) and *multi-image SR* (MISR) approaches (fusing multiple observations of the same scene).

Early SISR techniques were based on interpolation and example-based methods, long before the deep learning era. Interpolation-based methods such as bilinear or bicubic interpolation provide simple baselines by smoothing and enlarging pixel grids, but tend to produce blurry results due to their limited modeling of image structure. More sophisticated classical methods leveraged natural image priors. For example, patch-based dictionary learning and sparse coding became influential around 2010. Yang et al. [6] pioneered learning pairs of small patches from LR–HR images, known as dictionary atoms, so that a sparse linear combination of LR atoms can reconstruct high-frequency details via the corresponding HR atoms. This sparse representation approach achieved sharper results than interpolation by reusing high-frequency patterns from training data.

Convolutional neural network (CNN) models brought a breakthrough in SISR by learning the LR-to-HR mapping from large datasets. The seminal SRCNN [7] of Dong et al. was the first CNN for SR, with a three-layer network directly learning to improve the quality of interpolated LR inputs. Despite its simplicity, SRCNN significantly outperformed sparse-coding methods, demonstrating the power of data-driven feature extraction. Dong et al. later proposed FSRCNN [8], an accelerated variant that operates in the LR space for efficiency. Kim et al. introduced a *very deep SR* network (VDSR) [9] with ca. 20 layers and residual learning to ease training. Subsequently, the EDSR model removed batch normalization to allow for even deeper residual networks, achieving state-of-the-art performance in the 2017 NTIRE challenge [10].

By 2017, *generative adversarial networks* (GANs) emerged to tackle a core limitation of purely *mean squared error* (MSE) optimized CNNs: overly smooth outputs lacking high-frequency texture. Ledig et al. proposed SRGAN [11], introducing adversarial training to favor perceptually sharp results. SRGAN's generator (a deep ResNet architecture, also known as SRResNet, due to its use of residual learning blocks) is trained with a discriminator to produce outputs that are difficult to distinguish from real HR images. This GAN-based approach produced much crisper and more detailed images, at the cost of lower PSNR. Wang et al. later improved this with ESRGAN [12], which used residual-in-residual dense blocks and a more perceptually aligned loss, resulting in more natural textures. Beyond perceptual losses, task-driven objectives have emerged to tailor super-resolution to downstream applications. For instance, Zyrek and Kawulok [13] proposed a method that optimizes SISR for improved text recognition from scanned documents by combining image similarity with text detection objectives.

While not directly focused on SR, Bairi et al. [14] addressed the related problem of image reconstruction from compressive sensing data. Their dual-path framework integrates Vision Transformers and perceptual optimization, combining global contextual modeling with perceptual loss to enhance reconstruction quality beyond traditional fidelity-driven methods.

Most recently, diffusion probabilistic models have been applied to SISR, marking a new frontier in fidelity and diversity. Saharia et al. introduced SR3 [15], which adapts denoising diffusion models for SR. Starting from pure noise, SR3 iteratively denoises the input while being conditioned on the LR image, generating highly photo-realistic outputs.

Whereas SISR relies on learned priors to hallucinate missing details, MISR methods exploit actual complementary information from multiple observations of the same scene. By combining a sequence of LR images (often with subpixel shifts between them), MISR can surpass the limits of single-image fidelity. Traditional MISR techniques were rooted in multi-frame reconstruction theory and often cast as Bayesian or regularized optimization problems. A classical formulation assumes each LR image is a warped, blurred, downsampled, and noisy version of an unknown HR image. The SR task is then to invert this imaging model. Early solutions like the work [16] of Schultz and Stevenson employed a *maximum a posteriori* (MAP) estimator with a smoothness prior, solved via iterative back-projection or gradient descent. Farsiu et al. later proposed a fast and robust SR (FRSR) [17] MISR algorithm that combined an L1-norm fidelity term (robust to outliers) with a simplified regularization, and crucially computed registration error in the HR space to avoid repeated interpolations. Such approaches (e.g., Bayesian inference, maximum likelihood or MAP estimation) dominated MISR for years, incorporating robust alignment and prior terms to handle noise and misregistrations. However, they often required careful tuning and were computationally intensive.

With the rise of deep learning, MISR has seen new data-driven approaches addressing the key challenges of alignment and fusion. The first application of deep learning in MISR was EvoNet, presented in [4], where the method combines the advantages of multi-image fusion with learning the LR-to-HR mapping using deep networks, achieving superior reconstruction accuracy compared to state-of-the-art SR methods. A milestone in deep MISR was the PROBA-V Super Resolution Challenge [18], which spurred the development of several methods, including the HighRes-net by Deudon et al. [19] and the winning approach, DeepSUM [20]. HighRes-

net introduced an end-to-end trainable framework to jointly learn co-registration of multiple LR frames and their fusion into a single HR output. It uses a recursive fusion strategy with an implicit reference frame, eliminating the need for explicit motion compensation between inputs. DeepSUM, on the other hand, leverages deep neural networks to efficiently integrate spatial and temporal information from unregistered multi-temporal images, achieving superior reconstruction quality. Another notable model is the *residual attention MISR* (RAMS) network [21] by Salvetti et al. RAMS integrates attention mechanisms into a deep CNN to adaptively weight the contribution of each pixel from each image during fusion.

Current state-of-the-art MISR research often borrows architectures from SISR and video SR, enhanced to handle multiple unordered inputs. Transformer-based MISR models [22][23] have recently emerged, leveraging self-attention to capture long-range dependencies across frames. Another approach is to represent a stack of input images as a graph which is super-resolved with a graph neural network [24] or graph attention network [25].

B. Image Quality Assessment

Despite the progress in SR algorithms, evaluating the quality of super-resolved images remains challenging. Generally, IQA methods are classified into FR, *reduced-reference* (RR), and *no-reference* (NR) approaches, depending on how much ground truth information is available [26]. FR-IQA metrics (e.g., PSNR, SSIM) compare the output directly against a reference image, RR-IQA methods use partial information about the reference (such as features or the input LR image), and NR-IQA (also called blind IQA) relies only on the output itself. In conventional restoration tasks, FR metrics like PSNR and SSIM have been widely used since a ground-truth HR image is usually available for synthetic test cases. However, these classical metrics focus on signal fidelity and often correlate poorly with human visual perception when applied to SR results. A super-resolved image that looks sharp and realistic to a human observer, might score worse in PSNR/SSIM than a blurry image that stays pixel-close to the reference [27]. This is especially true for GAN-based SR methods which introduce high-frequency content: they achieve higher subjective quality but lower PSNR (the perception-distortion dilemma [28]). As a result, learning-based IQA metrics have gained traction. These include deep feature-based distances like *learned perceptual image patch similarity* (LPIPS) [29] (which compares images in a CNN feature space) and *deep image structure and texture similarity* (DISTS) [30], as well as neural networks trained to predict human opinion scores.

For SR evaluation, both fidelity and perceptual quality are important, and researchers have devised specialized metrics to handle the conflicts between them [31]. Several recent works propose SR-specific IQA methods that account for characteristics of SR outputs.

FR-IQA compares the super-resolved image directly with its HR ground truth. Metrics like PSNR and SSIM are most commonly used, offering objective and repeatable evaluation. PSNR measures pixel-wise fidelity based on mean squared error, while SSIM captures local luminance, contrast, and structure similarities. Despite their popularity, these metrics often fail to reflect perceptual quality, especially for SR results generated by GAN-based models, which prioritize sharpness and realism over strict pixel accuracy [26][27]. Due to structural differences from the ground truth, more advanced FR methods have been proposed that analyze images in gradient, wavelet, or perceptual feature domains. For example, phase congruency and gradient magnitude have been employed to better model human sensitivity to edge structures, as illustrated by the FSIM metric [32], which leverages both features within a full-reference evaluation framework. While FR metrics are ideal for synthetic SR benchmarks with available HR references, they are unsuitable in real-world settings where such references are missing.

RR-IQA methods evaluate quality using partial information from the reference, often in the form of extracted features or the LR input image itself. This makes RR approaches more practical than FR in real-world SR scenarios, where full ground truth is unavailable. Early RR methods focused on edge preservation, texture statistics, or wavelet-based similarities between LR and SR images [27]. More recent approaches, like PFIQA [5], use deep neural networks to combine perceptual and fidelity-aware branches, leveraging both the SR image and the LR input alongside auxiliary information (e.g., scale factor). These models better reflect human perception by assessing whether reconstructed details are consistent with the available input and visually plausible. RR-IQA is especially valuable for detecting artifacts that contradict known image structures or suggest overfitting (e.g., excessive sharpening). While RR approaches strike a balance between fidelity and perception, they remain limited in evaluating purely hallucinated textures, as they cannot fully determine whether added details are semantically correct.

The NR-IQA approach is the most flexible and challenging setting, as it operates solely on the super-resolved image without any access to ground truth or input. NR metrics are indispensable in real-world applications, where only the output is available. Early NR-IQA approaches relied on natural scene statistics, assuming that distortions in SR images would cause statistical deviations. Methods like *natural image quality evaluator* (NIQE) [33] are based on statistical features derived from natural images and estimate deviations associated with sharpness, noise, and unnatural textures. These were later combined into composite scores like the *perceptual index* (PI) [27], which became widely used in perceptual SR evaluation. Recent NR-IQA methods leverage deep learning to directly predict subjective quality from image content. For example, KLTSRQA [34] uses Karhunen-Loève transform-based features, while others employ CNNs or transformers trained with human opinion scores [27]. Knowledge distillation techniques have also been introduced to train NR-IQA models using pseudo-labels obtained from strong full-reference models [27]. Although modern NR-IQA metrics outperform traditional ones in terms of perceptual alignment, they still face challenges

related to generalization and distinguishing realistic textures from artifacts without reference anchors.

III. Datasets

In the field of SR, selecting appropriate datasets is crucial for evaluating the performance of quality assessment metrics. This research leverages three distinct datasets: CVIU-17 [35], SISAR [36], and MuS2 [37]. The first two are widely employed [27], [31], [26] in the field of IQA, whereas the third one presents an intriguing complement owing to its distinctive characteristics. Below, we provide detailed descriptions of these datasets, highlighting their composition, purposes, and contributions to advancing SR research.

A. CVIU-17

The CVIU-17 dataset is a publicly available resource tailored for evaluating IQA metrics in the context of SISR. It comprises 180 LR images and 1,620 HR images, generated using nine distinct SR algorithms across six integer scaling factors (see Fig. 1). LR images were generated by applying a combination of downsampling and blurring to the original images from the BSD200 dataset [38].



Fig. 1. An example from the CVIU dataset showing super-resolved images obtained using various SR methods. The visual comparison highlights the differences in reconstruction quality across methods.

B. SISAR

The SISAR dataset (SR image quality database with semi-automatic ratings) stands as the largest-of-its-kind resource for IQA. It contains 12,600 labeled HR images derived from 100 natural LR images, each with a resolution of 1024×768 (see Fig. 2). These HR images were generated using ten types of SR algorithms. SISAR's scale and diversity make it an invaluable asset for developing robust and generalizable IQA metrics, paving the way for more accurate quality predictions in SR applications.

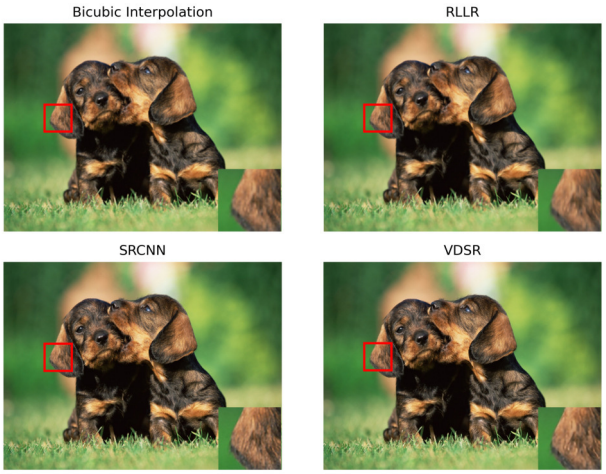


Fig. 2. An example scene from the SISAR dataset showing images super-resolved using various methods. The figure includes Bicubic Interpolation, RLLR, SRCNN, and VDSR, illustrating the visual differences in image quality achieved by each method.

C. MUS 2

The MuS2 dataset is a specialized benchmark for MISR of Sentinel-2 satellite [37]. It addresses the challenge of improving the spatial resolution of Sentinel-2 imagery, which has a ground sampling distance of 10 meters, by using HR WorldView-2 imagery as a reference. The dataset includes 91 scenes (see Fig. 3), each containing 14 or 15 Sentinel-2 LR images across 12 spectral bands, paired with corresponding WorldView-2 HR images. All LR and HR image pairs in MuS2 are pre-aligned, and no geometric misalignments are present.

MuS2 is particularly significant for remote sensing applications, where HR imagery is critical for tasks such as environmental monitoring, urban planning, and disaster management. Unlike many SR datasets that rely on simulated data, MuS2 provides real-world data, capturing the complexities of satellite imaging, such as atmospheric variations and temporal changes. This realism makes it an ideal testbed for developing and validating MISR algorithms that fuse information from multiple images to achieve superior reconstruction accuracy.

IV. Proposed Approach

The main goal of this study is to investigate the use of keypoint-based metrics for assessing the quality of super-resolved images. These metrics, derived from Key.Net, include MSIP and the repeatability.

Key.Net [2] is a hybrid keypoint detection model that combines handcrafted and learned CNN filters within a multi-scale architecture. The handcrafted filters are inspired by traditional methods like Harris [39] detectors, capturing first and second-order image derivatives. These filters act as soft anchors, reducing the number of learnable parameters and improving stability. Key.Net also incorporates a multi-scale pyramid representation, which enhances robustness to scale

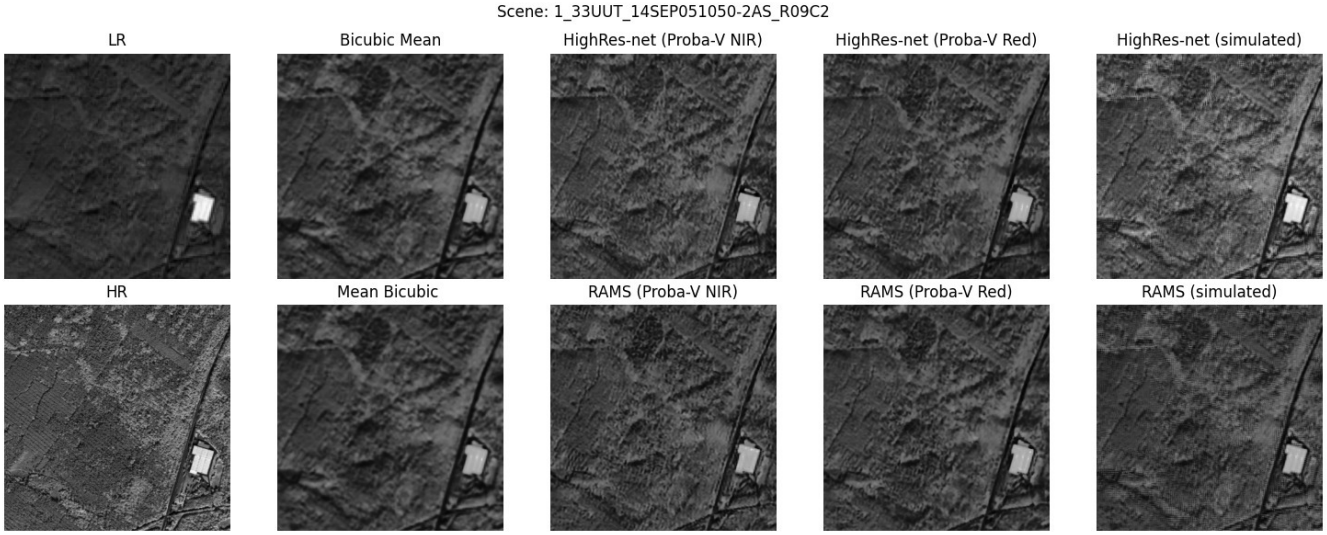


Fig. 3. An example scene from the MuS2 dataset, showing both LR and HR images. The figure includes bicubic interpolation of a mean LR, averaged image of bicubic interpolation outcomes from each LR image, as well as super-resolved images obtained from HighRes-net and RAMS networks trained using real-world PROBA-V images and from a simulated dataset.

changes by processing the input image at multiple scales. The learned filters are then applied to these multi-scale features to localize, score, and rank keypoints. The model uses a novel MSIP layer as loss function, which optimizes keypoint detection by leveraging both local and global information.

The MSIP module extends the concept of the *index proposal* (IP) layer by aggregating multiple IP layers applied at different spatial scales. Each IP layer operates on a shared response map using a local window of size $N_s \times N_s$, where N_s denotes the scale-specific window size. Within each window, a spatial softmax operator is used to convert raw activation values into a probability distribution, highlighting the most prominent local responses. This results in differentiable keypoint localization, computed as the expected coordinate (weighted average) of positions within the window.

By stacking several IP layers with increasing window sizes, MSIP is able to extract keypoints that are not only locally salient but also stable across spatial scales. This multi-scale strategy improves robustness and suppresses unstable detections, as only keypoints that persist across larger contexts are retained. This design encourages the detector to favor geometrically consistent keypoints that are resilient to scale variations, improving repeatability under transformation.

The repeatability quantifies how consistently keypoints are detected between two images of the same scene (e.g., the reference HR and its super-resolved counterpart). A keypoint is considered repeatable if it has a corresponding match within a defined spatial tolerance, often under geometric transformations such as scaling or homography. Formally, it is defined as:

$$\text{repeatability} = \frac{|\mathcal{K}_a \cap \mathcal{K}_b|}{\min(|\mathcal{K}_a|, |\mathcal{K}_b|)}, \tag{1}$$

where:

- $\mathcal{K}_a$ and $\mathcal{K}_b$ are the sets of keypoints detected in images I_a and I_b, respectively.
- $|\mathcal{K}_a \cap \mathcal{K}_b|$ is the number of matching keypoints between the two images.
- $\min(|\mathcal{K}_a|, |\mathcal{K}_b|)$ is the minimum number of keypoints detected in either image.

In this implementation, keypoints are defined by both their spatial coordinates and the scale at which they were detected. To determine whether two keypoints correspond, the *intersection-over-union* (IoU) is computed between the circular regions centered at each keypoint, scaled according to their associated detection scale. A match is established, if the IoU exceeds a fixed threshold. The repeatability can be measured at a single scale, reflecting keypoint consistency under fixed resolution, or across multiple scales, which accounts for robustness to scale variations. Multi-scale repeatability provides a more comprehensive assessment of detector stability under varying image resolutions.

The process of calculating keypoint-based metrics is presented in Fig. 4. To compute MSIP and Repeatability, the SR and HR images are converted to grayscale. Then, each of them is passed through the Key.Net network. As the output of Key.Net, a response map is obtained, and further calculations are performed depending on the metric:

- MSIP: Both response maps (SR and HR) are processed through the MSIP layer, which extracts sets of stable keypoints at different scales.
- Repeatability From the response maps, Non-Maximum Suppression (NMS) is applied to obtain sets of keypoints. IoU between corresponding keypoints from SR and HR images is calculated, resulting in the repeatability metric.

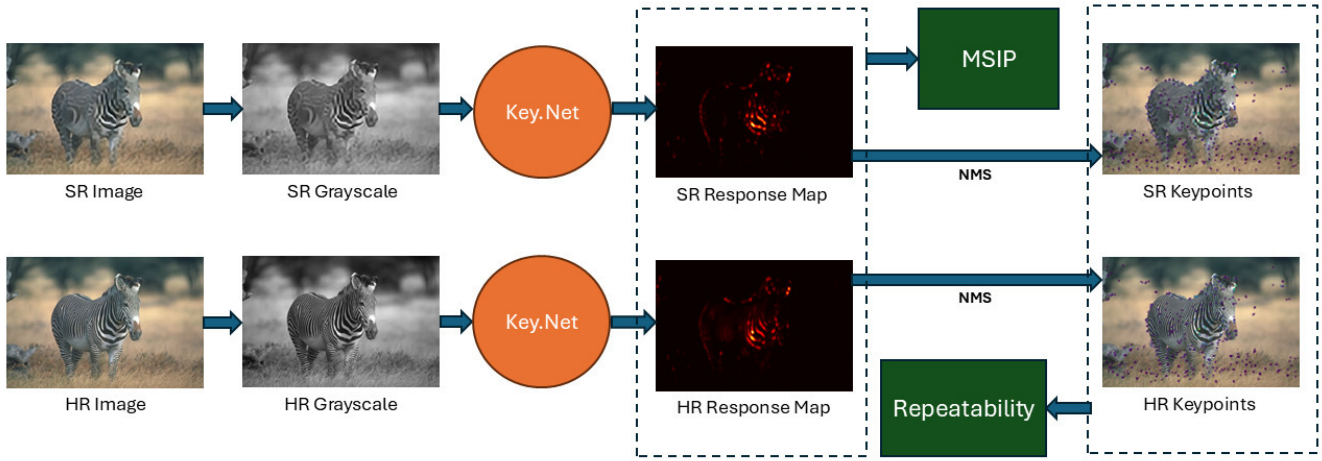


Fig. 4. A pipeline for keypoint-based metrics calculation in IQA for SR. Super-resolved and HR images are converted to grayscale, processed through Key.Net to obtain response maps, and further analyzed using MSIP and repeatability metrics.

We hypothesize that incorporating keypoint repeatability into IQA offers a promising alternative to perceptual or distortion-based measures. To verify this, we compute both conventional metrics—PSNR, SSIM, and LPIPS—as well as keypoint-based metrics, including single-scale repeatability (REP_S), multi-scale repeatability (REP_M), and the MSIP for all three datasets. Furthermore, for the CVIU17 and SISAR datasets, where *mean opinion scores* (MOS) are available, we measure correlation between metric predictions and human perceptual judgments.

V. Experiments

A. Experimental Setup

We carried out experiments on three standard IQA datasets for SR: CVIU-17, SISAR, and MuS2. Both CVIU-17 and SISAR provide HR images and MOS, enabling correlation analysis of objective metrics with human perception. In contrast, MuS2 lacks MOS labels and focuses on Sentinel-2 satellite images for MISR.

Since our approach does not require additional training, we directly used the super-resolved images and the corresponding HR images from each dataset. We applied the Key.Net model (PyTorch implementation) with the original pretrained weights; only minor modifications were introduced to handle grayscale conversion (for CVIU-17 and SISAR) and kernel sizes in the MSIP function. For the purposes of keypoint-based analysis, we assumed an identity homography transformation, implying no geometric discrepancies between the HR images and corresponding SR outcomes.

In our experiments, we evaluated several configurations of the MSIP function to understand how varying kernel sizes and weighting strategies affect the keypoint-based assessment of super-resolved images. We decided to use this division to systematically analyze the impact of different scale levels on the quality of keypoint detection.

- MSIP_default: The default Key.Net configuration, employing four scales with decreasing weights.
- MSIP_5: Single-scale variant using only the smallest kernel size for window.
- MSIP_17: Single-scale variant using only the largest kernel size for window.

Table I summarizes the kernel sizes and weights for each MSIP variant used in both scenarios.

TABLE I
MSIP configurations used for CVIU-17, SISAR, and MuS2.

MSIP Variant	Kernel Sizes	Weights
MSIP_default	5×5, 9×9, 13×13, 17×17	1.0, 0.5, 0.25, 0.12
MSIP_5	5×5	1.0
MSIP_17	17×17	1.0

We computed the following metrics for each super-resolved and HR image pair: PSNR, SSIM, LPIPS, and the proposed set of keypoint-based metrics (MSIP, REP_S, REP_M). Where MOS annotations existed (CVIU-17, SISAR), we calculated the *Pearson linear correlation coefficient* (PLCC), *Spearman rank-order correlation coefficient* (SRCC), and *root mean squared error* (RMSE) between each metric's prediction and the MOS. To facilitate fair comparison across metrics, all correlation scores were normalized.

B. Results and Analysis on SISR IQA Benchmark Datasets

Table II shows that Dong11 yield the highest MOS (0.5491), surpassing simpler methods like bicubic interpolation (0.4603). Although these top MOS approaches do not always achieve the highest PSNR or SSIM, they align more closely with the keypoint-based REP_M scores. This finding supports the well-known perception-distortion tradeoff, wherein an image may visually please observers but deviate from the ground truth at the pixel level.

TABLE II
EVALUATION OF SR METHODS ON THE CVIU-17 DATASET.

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17	MOS
Bicubic	24.5020	0.6653	0.4435	87.7187	84.2950	0.1134	0.0420	0.1235	0.4603
BP	24.5766	0.6751	0.4143	87.8826	84.5593	0.1136	0.0412	0.1243	0.4867
Shan08	22.8392	0.5890	0.4627	83.9553	79.1901	0.1926	0.0534	0.2822	0.2834
Glasner09	24.3997	0.6786	0.3393	87.3181	83.6684	0.1188	0.0409	0.1587	0.4467
Yang10	24.5002	0.6752	0.3663	87.0764	83.5761	0.1173	0.0428	0.1295	0.5138
Dong11	24.1746	0.6661	0.3838	**88.4883**	**85.2192**	0.1308	0.0449	0.1637	**0.5491**
Yang13	24.6486	**0.6822**	0.3466	87.4525	84.1223	0.1110	0.0393	0.1392	0.5426
Timofte13	24.6546	0.6767	**0.3379**	86.9872	83.5923	**0.1090**	**0.0385**	0.1241	0.5248
SRCNN	**24.6925**	0.6802	0.3403	88.1164	84.6920	0.1141	0.0407	**0.1177**	0.5444
PLCC	0.5835	0.6317	0.7529	0.5899	0.6769	0.4816	0.4865	0.3364	-
SRCC	0.5716	0.6274	0.7401	0.6280	0.7164	0.5318	0.4929	0.3953	-
RMSE	4.9681	4.8174	4.8293	4.6891	4.7130	4.6094	4.7082	4.5634	-

TABLE III
EVALUATION OF SR METHODS ON THE SISAR DATASET

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17	MOS
BICUBIC	21.8163	0.6551	0.5001	56.3998	48.9003	0.0813	0.0141	0.1543	0.4494
RLLR	18.2244	0.5803	0.5227	46.9003	39.6196	0.1051	0.0163	0.2193	0.4409
SRCNN	18.1481	0.5748	0.5474	51.0858	42.6164	0.0960	0.0158	0.1893	0.4727
VDSR	**22.7156**	**0.6738**	**0.4664**	**64.3389**	**56.1876**	**0.0639**	**0.0130**	**0.1105**	**0.4896**
PLCC	0.6174	0.4830	0.6185	0.6957	0.7133	0.5679	0.4386	0.4956	-
SRCC	0.6129	0.4882	0.6059	0.6861	0.7021	0.5931	0.4263	0.5771	-
RMSE	0.2296	0.2611	0.2147	0.2098	0.2131	0.3241	0.2616	0.4238	-

Comparing the MSIP configurations in Table II, we observe that single-scale variants (MSIP_5, MSIP_17) can diverge significantly for certain SR methods, suggesting that analyzing only a single kernel size may overlook important local structures. Correlation part of this table, reinforces this observation: the correlation of MSIP_5 and MSIP_17 with MOS is generally lower than that of MSIP_default, indicating that decreasing weights over multiple kernels are more effective than a single-scale approach.

A similar trend appears in Table III, where VDSR achieves the highest MOS (0.49) but not the highest PSNR. The local structure measures REP_S and REP_M exhibit a relatively high correlation with MOS, implying that preserving essential edges and textures is key to subjective quality. As in the CVIU-17 dataset, the single-scale MSIP_5 and MSIP_17 lag behind MSIP_default in terms of correlation metrics, though MSIP_default still does not reach the correlation levels of REP_M. This suggests that while MSIP has potential for capturing perceptual differences in super-resolved images, it may require additional tuning or refined weighting strategies to better reflect subjective quality.

Both tables also presents a direct comparison of PLCC, SRCC, and RMSE against MOS for both datasets. As expected, PSNR and SSIM show modest correlations, consistent with their known limitations in modeling perceptual sharpness or high-frequency details. LPIPS reports a higher correlation, which is unsurprising, as it is designed to capture perceptual similarities. The keypoint-based metrics REP_S and REP_M outperform classic metrics in most cases, highlighting that matching local structure and geometry is highly relevant to subjective quality. In comparison, the MSIP_default variants of MSIP reveal a promising correlation but trail behind the best REP_M scores, indicating that further refinement of multi-scale configurations and weighting choices could improve MSIP-based evaluations.

C. Results and Analysis on the Real-World MISR Dataset

Below, we summarize the results of various SR approaches on the MuS2 dataset, which includes Sentinel-2 satellite imagery. Table IV covers both baseline methods (Bicubic) and SR models trained on real-world PROBA-V (NIR or RED bands) or simulated Sentinel-2 data. Specifically:

- Bicubic Mean: Single bicubic upsampling of an averaged LR frame.
- Mean Bicubic: Bicubic upsampling of each LR frame followed by averaging.
- HighRes-net (PROBA-V NIR) / HighRes-net (PROBA-V RED): HighRes-net variants trained on PROBA-V data (NIR or RED bands).
- HighRes-net (simulated): A HighRes-net variant trained on simulated Sentinel-2 data.
- RAMS (PROBA-V NIR) / RAMS (PROBA-V RED): RAMS variants trained on PROBA-V data (NIR or RED bands).
- RAMS (simulated): A RAMS variant trained on synthetic Sentinel-2 data.

A closer look at the metrics reveals several key issues:

1) Bicubic Mean and Mean Bicubic achieve the highest PSNR and relatively high SSIM values, yet their LPIPS and MSIP scores are very poor. Based on the article [37], they produce visually the smoothest (most blurred) outputs, confirming that high PSNR and SSIM alone may not indicate perceptually sharp images. MOS study was

TABLE IV
EVALUATION RESULTS ON THE MUS2 DATASET. IN THE BRACKETS, WE INDICATE THE DATASET USED FOR TRAINING (REAL-WORLD PROBA-V OR SIMULATED IMAGES).

Method	PSNR	SSIM	LPIPS	REP_S	REP_M	MSIP_default	MSIP_5	MSIP_17
Bicubic Mean	24.1514	0.6012	0.5667	42.3759	24.0449	0.5470	0.0815	1.1352
Mean Bicubic	**24.1515**	0.6012	0.5667	42.3708	24.0607	0.5457	0.0815	1.1403
HighRes-net (PROBA-V NIR)	23.9385	0.5970	**0.4357**	42.7648	28.4317	0.5163	0.0806	1.0476
HighRes-net (PROBA-V RED)	23.9802	0.5995	0.4476	42.0976	26.5394	0.5244	**0.0798**	1.0775
HighRes-net (simulated)	23.6873	0.5415	0.4905	43.1661	28.4554	0.5349	0.0857	1.0651
RAMS (PROBA-V NIR)	23.9743	0.6004	0.4420	42.6352	**28.5710**	**0.5109**	0.0815	**1.0271**
RAMS (PROBA-V RED)	24.0145	**0.6015**	0.4593	41.0973	25.8652	0.5233	0.0808	1.0686
RAMS (simulated)	23.8417	0.5324	0.5621	**43.3115**	28.5120	0.5198	0.0831	1.0436

also done, but the methodology does not allow for such analysis as with CVIU/SISAR.

2) RAMS (PROBA-V NIR) delivers the best visual quality overall. It achieves consistently strong metrics across the board, showing the highest MSIP and REP_M scores and ranking second-best in LPIPS. This suggests an excellent balance between fidelity and perceptual quality.
3) The second-best visual performance is offered by HighRes-net (PROBA-V NIR), which maintains strong metrics and a low LPIPS, indicative of better perceptual alignment.
4) Models trained on synthetic Sentinel-2 data have lower PSNR and weaker SSIM values compared to their PROBA-V counterparts. Despite this, they achieve relatively high REP_S and REP_M scores, indicating their ability to reconstruct sharper local details, even in the absence of precise pixel-level alignment with the reference image.

In summary, Bicubic Mean and Mean Bicubic yield the highest PSNR and SSIM, but the worst perceptual outcomes and poor LPIPS and MSIP scores. By contrast, RAMS (PROBA-V NIR) stands out visually, backed by strong metrics, with HighRes-net (PROBA-V NIR) close behind. Meanwhile, the simulated Sentinel-2 models confirm that sharper details may be underappreciated by fidelity-based metrics (PSNR, SSIM) but can be captured by the repeatability and perceptual assessments.

VI. CONCLUSION

In this study, we explored the utility of keypoint-based metrics for assessing the quality of super-resolved images across single- and multi-image scenarios. By capturing local feature consistency through multi-scale repeatability, our approach complements conventional fidelity metrics (PSNR, SSIM) and perceptual measures (LPIPS), providing a stronger alignment with human subjective ratings. Experimental results on CVIU-17, SISAR, and MuS2 confirm that multi-scale keypoint detection highlights structurally salient regions that significantly influence observers' quality judgments. Moreover, the proposed MSIP variants offer an effective means to combine different spatial kernels and weights, focusing on various frequency components within super-resolved images.

Although the proposed metrics show strong potential, further work is needed to address current limitations and explore open challenges. First, the comparison with advanced full- and no-reference IQA metrics is limited; extending this analysis is essential to fully position keypoint-based evaluation within the broader IQA landscape. Ongoing research focuses on adapting the MSIP metric specifically for SR tasks, both as a standalone quality measure and as a trainable loss function in SR models. Preliminary results indicate its potential to be comparable with state-of-the-art perceptual metrics. Additionally, future work should address geometric misalignments by evaluating repeatability under non-identity homographies and by integrating registration-aware keypoint detectors. Finally, we plan to extend the MuS2 dataset with human opinion scores, enabling more robust evaluation of MISR methods in real-world settings.

REFERENCES

[1] P. Benecki, M. Kawulok, D. Kostrzewa, and L. Skonieczny, "Evaluating super-resolution reconstruction of satellite images," *Acta Astronautica*, vol. 153, pp. 15–25, 2018. doi: 10.1016/j.actaastro.2018.07.035

[2] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.Net: Keypoint detection by handcrafted and learned CNN filters," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00593 pp. 5835–5843.

[3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005. doi: 10.1109/TPAMI.2005.188

[4] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1062–1066, 2020. doi: 10.1109/LGRS.2019.2940483

[5] X. Lin, X. Liu, H. Yang, X. He, and H. Chen, "Perception- and fidelity-aware reduced-reference super-resolution image quality assessment," *IEEE Transactions on Broadcasting*, vol. 71, no. 1, pp. 323–333, 2025. doi: 10.1109/TBC.2024.3475820

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, vol. 19, no. 11, pp. 2861–2873, 2010. doi: 10.1109/TIP.2010.2050625

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016. doi: 10.1109/TPAMI.2015.2439281

[8] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46475-6_25 pp. 391–407.

[9] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. doi: 10.1109/CVPR.2016.182 pp. 1646–1654.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. doi: 10.1109/CVPRW.2017.151 pp. 1132–1140.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.19 pp. 105–114.

[12] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-11021-5_5 pp. 63–79.

[13] M. Zyrek and M. Kawulok, "Task-driven single-image super-resolution reconstruction of document scan," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024. doi: 10.15439/2024F7855 p. 259–264.

[14] Z. Bairi, K. B. Bey, O. Ben-Ahmed, A. Amamra, and A. Bradai, "Dual-path image reconstruction: Bridging vision transformer and perceptual compressive sensing networks," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. PTI, 2023. doi: 10.15439/2023F978 pp. 347–354.

[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023. doi: 10.1109/TPAMI.2022.3204461

[16] R. Schultz and R. Stevenson, "Improved definition video frame enhancement," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 1995. doi: 10.1109/ICASSP.1995.479905 pp. 2169–2172 vol.4.

[17] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004. doi: 10.1109/TIP.2004.834669

[18] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of proba-v images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019. doi: 10.1007/s42064-019-0059-8

[19] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "HighRes-net: Recursive fusion for multi-frame super-resolution of satellite imagery," *arXiv preprint arXiv:2002.06460*, 2020. doi: 10.48550/arXiv.2002.06460

[20] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2020. doi: 10.1109/TGRS.2019.2959248

[21] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, 2020. doi: 10.3390/rs12142207

[22] J. Li, Q. Lv, W. Zhang, B. Zhu, G. Zhang, and Z. Tan, "Multi-attention multi-image super-resolution transformer (mast) for remote sensing," *Remote Sensing*, vol. 15, no. 17, 2023. doi: 10.3390/rs15174183

[23] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, "TR-MISR: Multiimage super-resolution based on feature fusion with transformers," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1373–1388, 2022. doi: 10.1109/JSTARS.2022.3143532

[24] T. Tarasiewicz and M. Kawulok, "A graph neural network for heterogeneous multi-image super-resolution," *Pattern Recognition Letters*, vol. 189, pp. 214–220, 2025. doi: 10.1016/j.patrec.2025.01.028

[25] ——, "A graph attention network for real-world multi-image super-resolution," *Information Fusion*, p. 103325, 2025. doi: 10.1016/j.inffus.2025.103325

[26] L. Shu, Q. Zhu, Y. He, W. Chen, and J. Yan, "A survey of super-resolution image quality assessment," *Neurocomputing*, vol. 621, p. 129279, 2025. doi: 10.1016/j.neucom.2024.129279

[27] H. Zhang, S. Su, Y. Zhu, J. Sun, and Y. Zhang, "Boosting no-reference super-resolution image quality assessment with knowledge distillation and extension," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10095465 pp. 1–5.

[28] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00652 pp. 6228–6237.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00068 pp. 586–595.

[30] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810

[31] K. Zhang, T. Zhao, W. Chen, Y. Niu, J. Hu, and W. Lin, "Perception-driven similarity-clarity tradeoff for image super-resolution quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5897–5907, 2024. doi: 10.1109/TCSVT.2023.3341626

[32] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. doi: 10.1109/TIP.2011.2109730

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. doi: 10.1109/LSP.2012.2227726

[34] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: A real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022. doi: 10.1109/TIP.2022.3154588

[35] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017. doi: 10.1016/j.cviu.2016.12.009

[36] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, "Learning-based quality assessment for image super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3570–3581, 2022. doi: 10.1109/TMM.2021.3102401

[37] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, "A real-world benchmark for sentinel-2 multi-image super-resolution," *Scientific Data*, vol. 10, no. 1, p. 644, 2023. doi: 10.1038/s41597-023-02538-9

[38] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001. doi: 10.1109/ICCV.2001.937655 pp. 416–423 vol.2.

[39] C. G. Harris and M. J. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 265–276
DOI: 10.15439/2025F7041
ISSN 2300-5963 ACSIS, Vol. 43

A Comparative Study of LSTM Efficiency vs. Transformer Power for Localized Time Series Forecasting

Twisampati Sarkar
School of Computing and Informatics
University of Louisiana at Lafayette, U.S.A.
Email: twisampati.sarkar1@louisiana.edu
ORCID: 0009-0003-6872-358X

Chee-Hung Henry Chu
School of Computing and Informatics
University of Louisiana at Lafayette, U.S.A.
Email: chu@louisiana.edu
ORCID: 0000-0002-5817-8798

***Abstract*—Forecasting multivariate time series increasingly uses deep learning, including models inspired by Neural Machine Translation (NMT). While Transformers excel at long-range dependencies, their computational overhead may not suit all problems. This study advocates the relevancy of Long Short-Term Memory (LSTM)-based encoder-decoder networks for scenarios with shorter input windows and prediction horizons. We present a comprehensive empirical evaluation of four classical LSTM-based NMT models, including variants with attention mechanisms specifically adapted for multistep time series forecasting. Our assessment focuses on their performance and the impact of varying input window sizes and prediction horizons within these computationally efficient, short-sequence contexts. We empirically compare these LSTM-based models against Transformer baselines operating under the same short input window and prediction horizon constraints.**

Key findings indicate that: (i) LSTM-based NMT models achieve or exceed existing state-of-the-art results for short-term predictions; (ii) within smaller input configurations, input window size minimally affects forecasting performance for tested horizons, suggesting efficiency gains are possible; (iii) for attention-based NMT models, attention scoring critically influences accuracy, demanding careful selection; (iv) our comparative analysis demonstrates that for time series problems where immediate historical context is sufficient, LSTM-based encoder-decoders are competitive with, or even outperform, Transformer-based models while offering a more computationally efficient solution.

Overall, our findings signify that original LSTM-based NMT models are robust and capable tools, particularly well-suited for short-term time series prediction tasks where local pattern capture and computational efficiency are priorities, even in the era of Transformers.

***Index Terms*—Attention Mechanism; Long Short-Term Memory; Transformers; Neural Machine Translation; Time Series Forecasting**

I. Introduction

MULTIVARIATE time series are embedded in our day-to-day activities; examples are health and mobile sensor readings, contagious disease markers, power consumption, road occupancy rates and traffic flow indicators, financial stock prices and currency exchange rates. A primary challenge in forecasting multivariate time series lies in extracting both complex temporal patterns—such as short-term trends and yearly seasonalities—and dynamic, nonlinear interdependencies among the individual driving variables of the multiple series. Different machine learning methods have been used to predict the trends [1], [2] as well as values [3] of the time series. Deep learning-based methods are increasingly utilized for such multivariate, multistep time series analysis tasks, including forecasting energy demand [4], network intrusion [5], and anomaly detection [6], [7]. In particular, deep learning architectures leveraging Recurrent Neural Networks (RNNs) effectively address many shortcomings of traditional models in these analytical tasks [8], [9].

Deep learning models for time series forecasting often draw inspiration from Neural Machine Translation (NMT) frameworks [10], highly successful in Natural Language Processing (NLP) and Computer Vision (CV). NMT models translate word sequences from one language to another, traditionally using an encoder to process the input into a fixed-length vector and a decoder to generate the output sequence [11]. This encoder-decoder paradigm adapts to time series prediction by treating past observations as an "input language" and future predictions as an "output language," effectively "translating" past sequences into future values. Long Short-Term Memory (LSTM) [12] is a specialized type of RNN particularly adept at sequential data tasks. LSTMs can be employed in an NMT encoder-decoder architecture: the encoder LSTM reads the input sequence (source language) word by word, processing it into a fixed-length context vector that encapsulates the entire input's meaning; subsequently, the decoder LSTM takes this context vector as its initial hidden state and generates the output sequence (target language) word by word, leveraging the learned information from the encoder to produce a coherent and accurate translation [13], [14].

Attention mechanisms [15] critically advanced NMT, allowing decoders to selectively focus on relevant parts of the input sequence (or past time series points), significantly improving performance for translation or prediction. LSTM-

Topical area: Advanced Artificial Intelligence in Applications

based NMT models can incorporate an attention mechanism to overcome the limitation of fixed-length context vectors by allowing the decoder to dynamically "attend" to different parts of the encoder's output at each decoding step, giving more weight to relevant source words when generating a particular target word. Transformer networks, a more recent NMT evolution, move beyond earlier RNN-based models by relying entirely on attention mechanisms to draw relationships between input and output sequence positions.

Traditional statistical approaches to time series forecasting have often focused on predicting a single future time step. However, contemporary applications increasingly demand predictions over a longer horizon—not just one step ahead, but potentially tens or even hundreds of time steps into the future. This extended prediction horizon seems to be well matched to NMT models' capabilities. For example in NLP it is not unusual to translate, say, five words in one language to, say, ten words in a different language. When extending this scenario to time series forecasting, it is natural to ask how the NMT models would perform in predicting a time sequence ("horizon") longer than the input sequence ("input window").

Given the recent advancements and dominance of Transformer-based architectures in handling long sequences, what then is the continued relevance and precise niche for LSTM-based models in time series forecasting, particularly when considering resource constraints or the need to understand specific short-to-medium range forecasting behaviors? To precisely delineate this niche and understand their practical applicability, we study how LSTM-based encoder-decoder models, particularly their performance across varied input and prediction window sizes and the influence of different attention mechanisms, compare in efficiency and effectiveness against contemporary Transformer models when specific short-to-medium range forecasting behaviors are critical. This raises the following questions:

1) How would these models behave when both the input window size and the predicted horizon window size are varied on a narrow time interval?
2) Although [16] presents a comparison of the different attention scoring functions on NLP tasks, how do the scoring functions affect the performance of these models for time series analysis? Does the use of different attention scoring function for different attention-based NMT models as described in [16], [17] have a significant effect on the performance of the attention-based models?
3) While Transformers excel at longer contexts, might LSTMs still offer competitive or superior performance and efficiency for shorter input windows and prediction horizons, a common requirement in many real-world applications?

To address these pertinent questions and delineate opportunities for enhancing time series forecasting, this research undertakes an empirical investigation into the performance of NMT models. Specifically, we evaluate and compare two baseline (or 'vanilla') NMT models [10], [18] with two attention-equipped NMT models [16], [17] and compared them to three Transformer models [19], [20], [21].

Based on the findings from our experimental study on multivariate time series forecasting models, we note the following contributions.

1) We provide a thorough empirical evaluation of classical LSTM-based encoder-decoder networks for multivariate time series forecasting, assessing their performance across various real-world datasets and different input window and prediction horizon configurations.
2) We examine the effect of varying input window sizes, demonstrating that even smaller input contexts can yield accurate predictions for longer horizons in certain time series. Furthermore, we confirm that the choice of attention scoring function significantly impacts forecasting accuracy for attention-based LSTM models, emphasizing the need for careful selection.
3) We empirically compare the performance of LSTM-based encoder-decoder models against Transformer-based architectures when both operate under comparable short input window widths and prediction horizons. Our findings suggest that for these specific local forecasting scenarios, LSTM-based methods can achieve competitive or even superior performance, often with greater computational efficiency, thereby highlighting their continued relevance where long-range dependencies are not the primary drivers or resources are constrained.
4) This study contributes to a more nuanced understanding of deep learning for time series forecasting by delineating specific contexts where LSTM-based models remain a highly viable and efficient choice, thus informing the selection of appropriate models based on problem characteristics (e.g., emphasis on local patterns, computational budget) rather than solely relying on state-of-the-art benchmarks achieved with extensive contexts.

The remainder of this paper is organized as follows. In Section II, we describe traditional and deep learning methods for time series forecasting, discussing their limitations and advantages, and explain the relevance of NMT models, including the recently developed Transformer methods. In Section III, we provide details of the LSTM-based encoder-decoder models used for time series forecasting and explain the model architecture, attention mechanisms, and training procedure. Section IV presents the experimental setup, including datasets, parameter settings, evaluation metrics, and results, along with a discussion and analysis of the findings. Finally, we draw our conclusion and suggest directions for future research in Section V.

II. Related Work

Most time series data in real world scenarios involves a mixture of long and short term patterns. Time series forecast-

ing models therefore have been developed to capture long or short term, or both, recurring patterns for accurate predictions. Classical approaches include traditional statistical methods such as Auto-Regressive (AR) models, e.g. ARIMA [22], Box-Jenkins [23], [24], Moving Averages, e.g. Holt and Winters method [25], [26]. Vector Auto-regression (VAR) family of models such as structured VAR [27] and elliptical VAR [28] are preferred when handling high-dimensional time series as they have the inherent property of AR models and are agnostic to the dependencies between output variables. However, the VAR models tend to overfit when modeling longer time intervals. Nonparametric algorithms such as Gaussian Processes [29] model the complex dynamic temporal relationship with a Bayesian approach, at a higher computational cost. Although these methods are somewhat effective, they fall short in capturing the complicated nonlinear dependencies between a wider or longer interval of observed time series signals and between multiple high-dimensional variables.

A. Recurrent Neural Networks and Encoder-Decoder Models

RNN-based networks [30]—based on LSTM or Gated Recurrent Units (GRU) [12], [18]—were the first sequential deep learners for time series analysis, where they encode past information as a fixed-length vector. An RNN learns a fixed-length representation from multiple sequences of arbitrary length. Advantages of using RNN-based models are as follows. Firstly, it has been shown [31] that RNN fits into the Nonlinear Autoregressive Moving Average framework. The state of the hidden layer in an RNN at any given time is dependent on the previous time steps. Hence RNNs have the inherent desirable properties of AutoRegression and Moving Average-based statistical methods. Secondly, RNNs represent time recursively, making temporal dependencies easier to learn [30]. Third, the recursive property of the RNN model allows it to store complex signals for a variable amount of time. These properties make the RNN a good choice for modeling or learning sequences of variable length data, where each sequence of data points can be assumed to be dependent on previous ones [32]. However, learning in RNNs can suffer from the problems of vanishing and exploding gradients [33], [34] and thus RNNs can have difficulty capturing long-term temporal dependencies. Vanishing gradients are mitigated by using LSTM or GRU units, which can effectively learn long-range temporal dependencies [12], [18]. Both the LSTM and GRU variants of RNN have been the foundations for many state-of-the-art algorithms in speech recognition, machine translation, sentiment analysis and other NLP tasks. LSTMs have also played a significant role in capturing temporal dependencies in CV tasks such as activity classification, detection and other video tasks [35], [36].

A prominent variant of RNN architecture is the encoder-decoder (sequence-to-sequence) model [18], [37]. The encoder-decoder model sequentially links two RNNs—that serve as an encoder and a decoder—through a fixed size vector, generally the last encoder state. The concept is to encode the input sequence as a fixed-length vector representation and use the decoder to translate the learned, encoded fixed-length vector into a variable output sequence. LSTM and GRU-based encoder-decoder networks are popular due to their success in neural machine translation.

B. Attention Mechanisms

Embedding all the information in a fixed-size vector may result in information loss. In addition, the performance of sequence-to-sequence models deteriorates as the length of sequences of the encoder or the decoder increases [38]. To overcome these shortcomings, the attention mechanism is introduced to the architecture to focus on important parts of the input temporal sequences, by adding relevance calculation from all the encoder cells to each decoder cell [17]. This method has been proven to be successful in various NMT tasks in which the translation of each word from one language to another requires specific attention to particular words in the input source sequence [39].

A prominent example of an attention-embedded architecture is one that uses a dual attention mechanism in an LSTM-based encoder-decoder with a two-stage process. The first attention stage automatically extracts essential driving variables from the encoder's previous hidden state, while the second selects the encoder's hidden states across all time steps [40]. This is effectively complemented by self-attention mechanisms integrated with convolutional layers, which further enhance the model's ability to focus on relevant information [41].

Other notable architectures include TreNet, a hybrid network combining LSTM and temporal Convolutional Neural Network (CNN) modules to predict time series trends [42]. Similarly, LSTNet [8] utilizes temporal one-dimensional convolutions for short-term patterns and LSTMs for long-term dependencies. Interestingly, the LSTNet integrates both components with an autoregressive module, demonstrating a significant performance drop when this component is removed. Their research also indicates that attention does not consistently improve results across all datasets, whereas skip connections have proven more effective [8], [43].

While many models focus on point predictions, DeepAR proposes outputting a probability density function based on parameter estimations at each time step [44]. However, this approach relies on assumptions about the time series data's distribution, which may not always be realistic in real-world scenarios.

Further advancements in recurrent architectures for time series forecasting include a bidirectional LSTM-based encoder-decoder that incorporates a position-based attention mechanism to exploit periodic patterns and strong variable correlations within the data [45]. For multi-quantile probabilistic forecasts, an approach utilizing LSTM-based encoders and decoders has been proposed, though primarily tested on univariate datasets [46]. Additionally, a time-aware LSTM Network addresses predictions on sequential data,

irrespective of the time lag between data points [47]. Finally, an architecture featuring an unidirectional LSTM encoder and a bidirectional LSTM decoder employs an attention mechanism over groups of time sequences, with encoder outputs additively passed to the decoder's input instead of calculating attention over longer sequences [48].

C. Transformer Methods

Transformers have largely supplanted RNN models in sequence modeling tasks due to their proficiency in learning long-range dependencies and interactions in sequential data [49]. However, time series forecasting poses unique challenges for Transformers. Their limitations in effectively modeling complex temporal dependencies and their computational intensity necessitate specialized approaches for this domain. One such approach, LogTrans [50], addresses these issues by employing local convolutions and a log-sparse self-attention mechanism to capture local patterns and reduce space complexity.

Informer [51] is an extension of of the Transformer with sparse self attention to mine the most important temporal features in a sequence. Autoformer [19] uses a decomposition framework and auto-correlation methods from statistical methods to implement an autoregressive attention that exploits the inherent periodicity in a time series data. Building upon Autoformer's decomposition architecture, FEDformer [20] and TDformer [21] enhance time series pattern learning by decoupling trend and seasonality modeling. Both methods leverage Fourier attention in the frequency domain, achieving linear time complexity. However, these approaches do not fully exploit multi-dimensional dependencies, which could further improve long-term forecasting accuracy. To address this, Crossformer [52] utilizes cross-dimensional information for multivariate modeling.

D. Long vs. Short Input Windows and Prediction Horizons

Having reviewed various LSTM and Transformer models in the literature, we turn our attention to a direct comparison of their strengths and weaknesses in sequence modeling. The Transformer-based methods use mainly self-attention and multiple heads to model long-range dependencies with large input windows (e.g., 50-100+). This contrasts with the present study's exploration of LSTM-based encoder-decoder models using shorter input windows (up to 12). Our focus is on their efficacy for localized forecasts: local patterns and shorter prediction horizons where computational efficiency is also a key consideration, reflecting differing typical application contexts.

Why should we pay attention to methods for short-term predictions? Beyond lower computational needs, modeling local patterns is often more critical than global trends for practical reasons. First, many real-world applications, such as hourly traffic management or minute-by-minute financial trading, prioritize accurate immediate forecasts where local patterns are most influential, unlike year-long trend predictions. Secondly, for non-stationary time series whose statistical properties change, global trends can mislead as past behavior may not represent the future; local pattern modeling allows adaptation to the most recent data dynamics. Furthermore, local patterns are crucial for capturing sudden changes, shocks, or anomalies that models focused on global, long-term trends might fail to react to quickly.

E. Evaluation Studies

Often algorithms are evaluated on different datasets or different metrics, or both. Some cases also see the use of these forecasting models on synthetic datasets. Although all forecasting models, citing the shortcomings of each of their predecessors, improve the state-of-the-art results, there is a lack of evaluation on standard real world datasets. The closest relatable work to that of ours is that of [9], who (a) conduct a detailed analysis of various multi-variate time series datasets and their patterns which account for explaining each of the model's performance on that data, (b) presented a prediction analysis over a longer term whilst deploying direct and recursive strategies, and (c) studied the application of Spatio-Temporal Graph Convolutional Networks. Although an evaluative study of non-gated RNNs, LSTMs and GRUs on short term predictions was presented in [53], it did not cover the encoder-decoder or the effect of various attention scoring functions. Our work compares the performances of the existing encoder-decoder models based on LSTM units and the effect of attention scores on short term predictions based on the predictions as a function of varying input sizes. Furthermore, we compare them to Transformer methods to put the results in the context of contemporary research.

III. Methodology

A. Problem Formulation

Consider a series of observed time series vector samples $Y_T = \{y_1, y_2, \cdots, y_T\}$ where $y_t \in \mathbb{R}^n$, n is the number of time series channels. Our objective is to predict a series of H future samples $\tilde{y}_{T+h}$, $h \in \{1, \cdots, H\}$. When the underlying architecture is an encoder-decoder with an RNN such as an LSTM in the decoder, the output is produced in the style of an RNN, viz. one sample at a time. It thus takes H time steps to predict all H samples. We refer to T of the input set Y_T as the *window width* and H of the output set as the *horizon*.

B. LSTM-Based Encoder

We adopt a sequence-to-sequence learning pipeline to encode T historical input variables for a given time window and decode the future h predictions, but with minor modifications. The encoder is an RNN based on LSTM or GRU that encodes the input sequences into a fixed dimensional feature vector [18], [37], [10]. For time series prediction, given the input sequence $Y_T = \{y_1, y_2, \cdots, y_T\}$ with $y_t \in \mathbb{R}^n$, where n is the number of driving variables, the encoder can be applied to learn a mapping from y_t to h_t $h_t = f_1(h_{t-1}; y_t)$ where $h_t \in \mathbb{R}^m$ is the hidden state of the encoder at time t, m is the size of the hidden state, and f_1 is a non-linear

activation function that could be an LSTM or a GRU. In this paper, we use an LSTM unit as f_1 to capture temporal dependencies. Each LSTM unit has a memory cell with the state s_t at time t. Access to the memory cell is controlled by three sigmoid gates: forget gate f_t, input gate i_t and output gate o_t. The update of an LSTM unit can be summarized as follows:

$$f_t = \sigma(W_f[h_{t-1}; y_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}; y_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o[h_{t-1}; y_t] + b_o) \tag{3}$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \tanh(W_s[h_{t-1}; y_t] + b_s) \tag{4}$$

$$h_t = o_t \odot \tanh(s_t) \tag{5}$$

where $[h_{t-1}; y_t] \in \mathbb{R}^{m+n}$ is a concatenation of the previous hidden state h_{t-1} and the current input y_t. The parameters W_f, W_i, W_o, $W_s \in \mathbb{R}^{m\times(m+n)}$, and b_f, b_i, b_o, $b_s \in \mathbb{R}^m$ are to be learned. The operators σ and $\odot$ are the logistic sigmoid function and the element-wise multiplication, respectively.

C. LSTM-Based Decoder without Attention

To predict the output $\hat{y_T}$, we use another LSTM-based RNN to decode the fixed vector $\mathbb{R}^m$ produced by the encoder LSTM:

$$\hat{y}_T = F(y_{t+1}, y_{t+2}, \cdots, y_{T+h-1}, X_T) \tag{6}$$

where $\hat{y}_T$ is the total number of sequences to be predicted, given the input sequence X_t. We experiment on two such encoder-decoder models without attention mechanisms, viz. that of Cho et al. [18] and Sutskever et al. [10]. The only difference between the two models is that in [18] an extra context vector is augmented to inputs at each of the time steps in the decoder. The context vector being the last hidden state of the encoder. The problem with a vanilla decoder is that its performance can degrade over a longer range of sequences [38].

D. LSTM-Based Decoder with Temporal Attention

The performance of the encoder-decoder network can deteriorate rapidly as the length of the input sequence increases [38]. Therefore, following the encoder with input attention, a temporal attention mechanism is used in the decoder to adaptively select relevant encoder hidden states across all time steps. The attention weight of each encoder hidden state at time t is calculated based upon the previous decoder hidden state $d_{s-1} \in \mathbb{R}^m$ and each of the hidden states of the encoder LSTM h'_t where T is the number of time steps in the input window. The term $[d_{s-1}; h'_t] \in \mathbb{R}^{2m}$ is a concatenation of the previous hidden state of the decoder and hidden states of the LSTM encoder units, whereas $v_d \in \mathbb{R}^m$, $W_d \in \mathbb{R}^{m\times 2m}$ and $U_d \in \mathbb{R}^{m\times m}$ are parameters to learn. The attention schemes per [17], [16] are as follows:

$$\text{score}(d_{s-1}, h'_t) = \begin{cases} v_a^T \sigma(W_d d_{s-1} + U_d h'_t) & \text{Bahdanau} \quad \text{(a)} \\ v_a^T \sigma(W_d[d_{s-1}; h'_t]) & \text{Concat} \quad \text{(b)} \\ d_{s-1}^T W_d h'_t & \text{General} \quad \text{(c)} \\ d_{s-1}^T h'_t & \text{Dot} \quad \text{(d)} \end{cases} \tag{7}$$

The slight changes in eqs. 7a,b are the replacement of the $\tanh$ non-linearity by the sigmoid σ, to keep the range of the intermediary logits in the range $[0, 1]$.

The attention weight β_t^i represents the weight of the ith encoder hidden state for the prediction:

$$\beta_t^s = \frac{\exp(\text{score}(d_{s-1}, h'_t))}{\sum_{k=1}^{T_x} \exp(\text{score}(d_{s-1}, h_t))} \tag{8}$$

Since each encoder hidden state h_i is mapped to a temporal component of the input, the attention mechanism computes the context vector c_t as a weighted sum of all the encoder hidden states $\{h'_1, h'_2, \cdots, h'_T\}$ as

$$c_s = \sum_{t=1}^{T} \beta_t^s h_t. \tag{9}$$

The context vector c_t is distinctly calculated at each time step as per Equation 9.

Once we get the weighted sum of the context vectors, we can combine them with the given target series $\{y_1, y_2, \cdots, y_{T-1}\}$:

$$\tilde{y}_{t-1} = \tilde{w}^T[y_{t-1}; c_{s-1}] + \bar{b} \tag{10}$$

where $[y_{t-1}; c_{t-1}] \in \mathbb{R}^{2m}$ is a concatenation of the decoder input y_{t-1} and the computed context vector c_{t-1}. Parameters $\tilde{w} \in \mathbb{R}^{2m}$ and $\tilde{b} \in \mathbb{R}$ map the concatenation to the size of the decoder input. The newly computed $\tilde{y}_{t-1}$ can be used for the update of the decoder hidden state at time t as :

$$d_s = f_2(d_{s-1}; \tilde{y}_{t-1}) \tag{11}$$

We choose the nonlinear function f_2 as an LSTM unit d_s that can be updated per eqs. 1 to 5.

E. Training Procedure

Our optimization strategy is similar to traditional direct time series forecasting model. Assume the input time series is $Y_t = \{y_1, y_2, \cdots, y_t\}$, we define a tunable window size q, and reformulate the input at time step t as X_t = $\{y_{t-q+1}, y_{t-q+2}, \cdots, y_t\}$. The problem then becomes a regression task with a set of feature-value pairs { X_t , Y_{t+h} } and is optimized with the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = e^{-8}$) [54], a variant of Stochastic Gradient Decent (SGD), with learning rate of e^{-5} to train the model. The minibatch size is set to 32 and number of epochs at 1,000, for training every model across all datasets. The squared error is the default loss function for many forecasting tasks, the corresponding optimization objective is formulated as:

$$\min_{\Theta} \sum_{y \in \Omega_{Train}} \|Y_t - \hat{Y}_{t-h}\|_F^2 \tag{12}$$

where Θ denotes the parameter set of our model, Ω_{Train} is the set of time stamps used for training, $\|\cdot\|_F$ is the Frobenius norm, and h is the horizon. Similar to [10], [18], [17], we train the encoder and decoder models jointly to minimize the objective function.

IV. Evaluation

A. Datasets

The multivariate time series datasets used in our study are as follows: Solar[1], Traffic[2], Electricity[3], Beijing PM2.5[4], and Exchange[5]. The pattern analysis on these datasets helps to understand the performance of time series forecasting models. Temporal regularity has been quantified using the Sample Entropy [9]. Following the methods used in [8] each of the datasets is split into training set (first 60%), validation set (next 20%) and test set (last 20%). All the training, validation and test data were normalized by Min-Max Scaling.

B. Parameter Settings

There are two parameters we have experimented on: the number of time steps for the input window T and the number of time steps to be predicted in the horizon h. Different from refs. [8], [9], [40], we report the results of the experiments where we set the value of $T \in \{3, 6, 9, 12\}$ and $h \in \{3, 6, 9, 12\}$. We use a single layer of LSTM units in both the encoder and decoder LSTMs without any dropout. We have experimented with the number of hidden units in the encoder and the decoder in {128, 256, 512}. All weights have been initialized from $-\sqrt{1/k}$ to $\sqrt{1/k}$, where k is the number of hidden units. We report average results over two runs of each model.

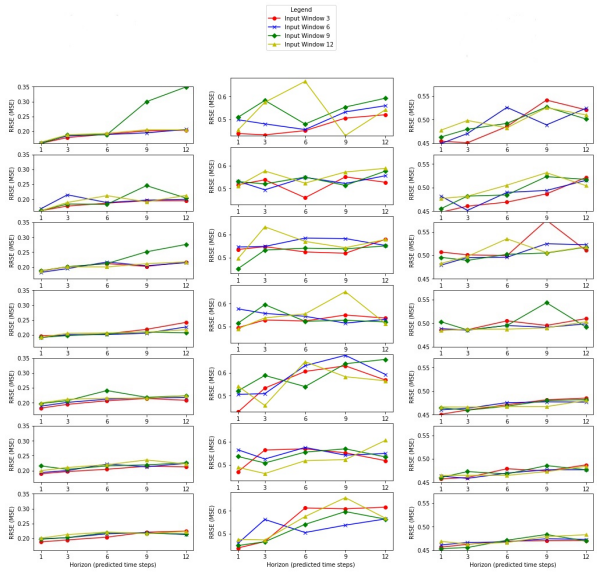
C. Metrics

To keep the evaluation metric standard to the benchmark results reported in [8], [9] we use the metrics in those papers, viz. the Root Relative Squared Error (RSE) and the Empirical Correlation Coefficient (CORR). A model where the RSE exceeds 1 indicates a bad prediction. Lower RSE values implies better predictions, while a higher CORR value is better.

D. Methods for Comparison

The two methods without attention used for comparison in this paper listed in the first two rows of Table I, are labeled as Cho et al. [18], and Sutskever et al. [10]. The two models with attention are Bahdanau et al. [17], and Luong et al. [16]. Of the two with attention, we experimented different attention schemes with corresponding scoring equations, referred to as "Bahdanau et al. (*)" and "Luong et al.(*)" where * refers to the attention scheme in Table I.

[1]https://www.nrel.gov/grid/solar-power-data.html
[2]https://pems.dot.ca.gov/
[3]https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
[4]https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
[5]https://github.com/laiguokun/multivariate-time-series-data/tree/master/exchange_rate



(a) Solar (b) Electricity (c) Traffic

Fig. 1: RSE plots (*y-axis*) at different prediction horizon (*x-axis*) for different input window widths (*by color*) evaluated using three datasets (*by column*) of the seven forecasters: Bahdanau et al. using attention schemes "Concat," "General," "Dot"; Cho et al.; Luong et al. using attention schemes "Concat," "General," "Dot." (*top to bottom*).

We compare the results to that of LSTNet [8], selected as our benchmark set. LSTNet serves as a strong, contemporary, and relevant benchmark because it was a leading model in the field at the time, designed to tackle the same core problems, and evaluated using consistent methodologies and datasets.

TABLE I: The methods and the attention scoring functions used in our experiments.

Method Name	Attention Scheme	Scoring Function	Eq.
Cho et al.	N/A	N/A	N/A
Sutskever et al.	N/A	N/A	N/A
Bahdanau et al.	Bahdanau	$v_a^T \sigma(W_d d_{s-1} + U_d h'_t)$	7(a)
Bahdanau et al.	Concat	$v_a^T \sigma(W_d [d_{s-1}; h'_t])$	7(b)
Bahdanau et al.	General	$d_{s-1}^T W_d h'_t$	7(c)
Bahdanau et al.	Dot	$d_{s-1}^T h'_t$	7(d)
Luong et al.	Concat	$v_a^T \sigma(W_d [d_{s-1}; h'_t])$	7(b)
Luong et al.	General	$d_{s-1}^T W_d h'_t$	7(c)
Luong et al.	Dot	$d_{s-1}^T h'_t$	7(d)

E. Discussion and Analysis of Experimental Results

We provide the RSE and CORR values of nine LSTM methods on five datasets in tables II and III. Due to space consideration, we show the results for window size set to 3, 6, 9, and 12 and for horizon of 3, 6, and 12 (skipping 9) for each method. From these tables, we observe that, on the three datasets with high periodicity (viz. Solar, Traffic,

TABLE II: Evaluation summary of all LSTM models (RSE and CORR) trained on MSE (Part I). Results in bold face indicate that the result of that column is better than the benchmark result [8] of a particular metric.

Dataset		Solar			Traffic			Electricity			Exchange-Rate			Beijing PM 2.5		
		Horizon			Horizon			Horizon			Horizon			Horizon		
	Metrics	3	6	12	3	6	12	3	6	12	3	6	12	3	6	12
Method: Cho et al. Next rows correspond to window size of 3, 6, 9, 12																
3	RSE	0.199	**0.203**	**0.242**	0.486	0.505	0.51	0.529	0.524	0.537	1.222	1.154	1.055	0.101	0.077	0.08
	CORR	0.981	**0.981**	**0.973**	**0.879**	0.867	**0.862**	0.838	0.841	0.826	0.604	0.639	0.693	0.944	0.959	0.958
6	RSE	0.201	**0.201**	**0.227**	0.485	0.495	0.499	0.558	0.545	0.532	1.119	1.035	0.916	0.087	0.085	0.091
	CORR	0.982	**0.981**	**0.977**	**0.877**	**0.870**	**0.868**	0.815	0.821	0.834	0.719	0.752	0.464	0.958	0.956	0.957
9	RSE	0.197	**0.203**	**0.207**	0.486	0.495	**0.492**	0.596	0.523	0.52	1.035	0.912	0.813	0.088	0.084	0.1
	CORR	0.982	**0.981**	**0.98**	**0.873**	**0.871**	**0.873**	0.808	0.837	0.829	0.761	0.746	0.788	0.956	0.951	0.955
12	RSE	0.205	**0.206**	**0.217**	0.487	**0.487**	0.503	0.538	0.555	0.513	0.948	0.865	1.158	0.089	0.104	0.119
	CORR	0.981	**0.980**	**0.978**	0.872	**0.874**	**0.866**	0.831	0.825	0.827	0.739	0.747	0.55	0.955	0.956	0.951
Method: Sutskever et al.																
3	RSE	0.208	**0.208**	**0.212**	**0.467**	**0.479**	**0.483**	0.569	0.578	0.539	0.822	1.571	1.552	0.085	0.086	0.736
	CORR	0.980	**0.979**	**0.979**	**0.882**	**0.878**	**0.879**	0.824	0.825	0.839	0.569	0.420	0.196	0.959	0.957	0.955
6	RSE	0.214	**0.211**	0.209	**0.468**	**0.481**	**0.487**	0.567	0.597	0.54	1.410	1.842	1.827	0.777	0.796	0.745
	CORR	0.978	**0.979**	**0.98**	**0.880**	**0.877**	**0.878**	0.823	0.817	0.841	0.368	-0.032	-0.066	0.958	0.953	0.953
9	RSE	0.218	**0.220**	**0.21**	**0.468**	0.475	**0.489**	0.607	0.546	0.577	1.843	1.840	1.813	0.768	0.794	0.736
	CORR	0.978	**0.978**	**0.979**	**0.880**	**0.879**	**0.877**	0.802	0.834	0.819	-0.034	-0.039	-0.058	0.955	0.954	0.957
12	RSE	0.230	**0.219**	**0.214**	0.486	**0.483**	0.496	0.514	0.561	0.574	1.517	1.807	1.835	0.784	0.791	0.789
	CORR	0.974	**0.977**	**0.979**	**0.874**	**0.878**	**0.878**	0.836	0.831	0.822	0.005	-0.017	-0.07	0.954	0.956	0.953
Method: Bahdanau et al. *(Bahdanau)*																
3	RSE	0.182	**0.193**	**0.202**	**0.453**	**0.482**	**0.494**	0.435	0.461	0.499	0.587	0.526	0.507	0.072	0.074	0.078
	CORR	**0.985**	**0.983**	**0.981**	**0.886**	**0.877**	**0.869**	0.866	0.856	0.852	0.872	0.862	0.785	0.959	0.960	0.957
6	RSE	0.184	**0.192**	**0.212**	**0.474**	0.525	0.515	0.497	0.468	0.555	0.598	0.492	0.543	0.089	0.082	0.083
	CORR	**0.985**	**0.983**	**0.979**	**0.875**	0.847	0.857	0.834	0.860	0.842	0.905	0.906	0.833	0.960	0.957	0.957
9	RSE	0.187	**0.192**	**0.21**	0.484	0.493	0.516	0.515	0.510	0.541	0.555	0.339	0.831	0.093	0.089	0.092
	CORR	0.984	**0.983**	**0.981**	0.870	0.869	**0.862**	0.836	0.825	0.836	0.899	0.936	0.83	0.958	0.956	0.959
12	RSE	0.188	**0.202**	**0.203**	0.484	**0.487**	0.51	0.579	0.472	0.553	0.597	0.384	0.467	0.091	0.091	0.1
	CORR	0.984	**0.981**	**0.981**	**0.874**	**0.871**	0.861	0.813	0.850	0.837	0.913	0.903	0.88	0.960	0.958	0.959
Method: Bahdanau et al. *(Concat)*																
3	RSE	**0.179**	**0.190**	**0.204**	**0.451**	**0.486**	0.521	0.433	0.451	0.521	0.690	0.435	0.495	0.075	0.074	0.082
	CORR	**0.985**	**0.983**	**0.981**	**0.886**	**0.874**	0.854	0.873	0.868	0.845	0.868	0.857	0.86	0.959	0.960	0.959
6	RSE	0.185	**0.189**	**0.207**	**0.471**	0.526	0.524	0.481	0.456	0.56	0.550	0.437	0.654	0.085	0.080	0.097
	CORR	0.984	**0.983**	**0.98**	**0.877**	0.850	0.857	0.850	0.856	0.817	0.934	0.859	0.776	0.958	0.958	0.954
9	RSE	0.188	**0.189**	0.349	0.480	0.492	0.501	0.583	0.480	0.593	0.668	0.424	0.539	0.090	0.085	0.102
	CORR	0.984	**0.983**	**0.973**	0.873	0.868	**0.867**	0.802	0.850	0.804	0.903	0.903	0.84	0.958	0.958	0.957
12	RSE	0.188	**0.193**	**0.205**	0.498	**0.483**	0.51	0.575	0.666	0.543	0.568	0.387	0.449	0.099	0.091	0.109
	CORR	0.984	**0.983**	**0.98**	0.867	**0.873**	**0.864**	0.812	0.794	0.827	0.856	0.737	0.824	0.954	0.958	0.957
Method: Bahdanau et al. *(General)*																
3	RSE	**0.177**	**0.187**	**0.195**	**0.460**	**0.469**	0.521	0.538	0.460	0.527	0.630	0.919	0.594	0.078	0.073	0.075
	CORR	**0.985**	**0.984**	**0.983**	**0.883**	**0.880**	0.857	0.834	0.862	0.838	0.896	0.703	0.794	0.959	0.960	0.957
6	RSE	0.215	**0.189**	**0.2**	**0.451**	0.490	0.515	0.494	0.548	0.556	0.587	0.409	0.477	0.087	0.087	0.113
	CORR	0.983	**0.984**	**0.982**	**0.887**	**0.870**	0.857	0.837	0.824	0.815	0.904	0.873	0.858	0.959	0.959	0.949
9	RSE	0.184	**0.183**	**0.204**	0.482	**0.484**	0.517	0.520	0.549	0.576	0.502	0.442	0.631	0.096	0.086	0.094
	CORR	0.985	**0.984**	**0.98**	0.872	**0.870**	**0.862**	0.843	0.822	0.807	0.903	0.906	0.852	0.959	0.960	0.958
12	RSE	0.189	**0.212**	**0.213**	0.482	0.505	0.504	0.576	0.523	0.588	0.575	0.647	0.514	0.098	0.103	0.094
	CORR	0.984	**0.980**	**0.979**	0.872	0.867	**0.862**	0.818	0.836	0.808	0.893	0.780	0.786	0.960	0.954	0.957
Method: Bahdanau et al. *(Dot)*																
3	RSE	0.200	**0.212**	**0.216**	0.501	0.499	0.51	0.547	0.525	0.58	1.298	0.656	1.119	0.080	0.076	0.088
	CORR	0.981	**0.981**	**0.979**	**0.874**	**0.870**	0.861	0.829	0.840	0.814	0.496	0.736	0.444	0.956	0.958	0.957
6	RSE	0.195	**0.217**	**0.215**	0.495	0.496	0.522	0.548	0.585	0.552	0.809	0.970	0.933	0.085	0.085	0.087
	CORR	0.982	**0.978**	**0.979**	**0.876**	0.869	0.859	0.826	0.806	0.828	0.812	0.582	0.616	0.958	0.955	0.953
9	RSE	0.201	**0.212**	**0.276**	0.489	0.502	0.517	0.533	0.541	0.551	0.594	0.809	0.698	0.095	0.087	0.117
	CORR	0.982	**0.981**	**0.97**	**0.875**	0.868	0.859	0.829	0.833	0.836	0.745	0.732	0.5	0.953	0.960	0.958
12	RSE	0.201	**0.200**	**0.217**	0.498	0.535	0.518	0.634	0.570	0.579	0.691	0.682	0.876	0.094	0.105	0.097
	CORR	0.981	**0.981**	**0.979**	**0.873**	0.849	0.859	0.803	0.806	0.818	0.855	0.839	0.776	0.954	0.957	0.954

TABLE III: Evaluation summary of all models (RSE and CORR) trained on MSE (Part II). Results in bold face indicate that the result of that column is better than the benchmark result [8] of a particular metric

Dataset		Solar			Traffic			Electricity			Exchange-Rate			Beijing PM 2.5		
		Horizon			Horizon			Horizon			Horizon			Horizon		
	Metrics	3	6	12	3	6	12	3	6	12	3	6	12	3	6	12
Method: Luong et al. *(Concat)*																
3	RSE	0.194	**0.206**	**0.208**	**0.460**	**0.471**	**0.485**	0.533	0.605	0.568	1.647	0.876	1.674	0.085	0.080	0.081
	CORR	0.982	**0.980**	**0.979**	**0.884**	**0.879**	**0.876**	0.821	0.807	0.821	0.743	0.501	-0.031	0.956	0.960	0.956
6	RSE	0.201	**0.212**	**0.217**	**0.463**	**0.475**	**0.477**	0.509	0.630	0.592	1.117	1.053	1.225	0.090	0.095	0.087
	CORR	0.981	**0.979**	**0.977**	**0.884**	**0.878**	**0.879**	0.830	0.812	0.815	0.696	0.630	0.617	0.956	0.959	0.959
9	RSE	0.207	**0.241**	**0.223**	**0.459**	**0.468**	**0.482**	0.588	0.539	0.658	0.893	0.898	0.971	0.097	0.095	0.091
	CORR	0.980	**0.976**	**0.977**	**0.882**	**0.880**	**0.876**	0.814	0.832	0.786	0.697	0.671	0.684	0.961	0.959	0.954
12	RSE	0.212	**0.215**	**0.223**	**0.466**	**0.467**	**0.48**	0.457	0.648	0.565	0.683	1.437	0.82	0.100	0.108	0.226
	CORR	0.979	**0.978**	**0.977**	**0.881**	**0.882**	**0.877**	0.852	0.800	0.818	0.724	-0.174	0.692	0.956	0.956	0.786
Method: Luong et al. *(General)*																
3	RSE	0.197	**0.204**	**0.212**	**0.460**	**0.479**	**0.487**	0.564	0.569	0.518	0.815	0.823	1.04	0.089	0.084	0.088
	CORR	0.982	**0.980**	**0.978**	**0.884**	**0.880**	**0.876**	0.825	0.812	0.842	0.862	0.609	0.637	0.959	0.959	0.958
6	RSE	0.200	**0.222**	**0.224**	**0.459**	**0.470**	**0.477**	0.525	0.575	0.549	0.883	0.909	1.044	0.092	0.111	0.083
	CORR	0.981	**0.978**	**0.977**	**0.884**	**0.879**	**0.878**	0.831	0.831	0.829	0.764	0.623	0.715	0.955	0.955	0.957
9	RSE	0.204	**0.216**	**0.226**	**0.473**	**0.468**	**0.477**	0.508	0.555	0.534	0.982	1.191	1.604	0.101	0.093	0.104
	CORR	0.980	**0.978**	**0.976**	**0.878**	**0.880**	**0.878**	0.843	0.818	0.84	0.741	0.652	0.649	0.945	0.952	0.953
12	RSE	0.210	**0.219**	**0.222**	**0.465**	**0.465**	**0.484**	0.462	0.517	0.606	1.123	0.927	1.625	0.123	0.106	0.4
	CORR	0.979	**0.978**	**0.977**	**0.881**	**0.882**	**0.875**	0.856	0.842	0.818	0.695	0.604	-0.11	0.952	0.956	0.767
Method: Luong et al. *(Dot)*																
3	RSE	0.195	**0.204**	**0.224**	**0.463**	**0.469**	**0.471**	0.466	0.612	0.616	0.853	0.805	0.65	0.084	0.111	0.784
	CORR	0.982	**0.980**	**0.977**	**0.883**	**0.881**	**0.88**	0.860	0.809	0.807	0.575	0.686	0.763	0.958	0.950	0.016
6	RSE	0.202	**0.215**	**0.213**	**0.466**	**0.468**	**0.473**	0.562	0.505	0.565	0.604	0.861	0.855	0.091	0.090	0.111
	CORR	0.981	**0.978**	**0.979**	**0.880**	**0.880**	**0.882**	0.832	0.848	0.827	0.782	0.667	0.651	0.954	0.958	0.953
9	RSE	0.203	**0.219**	**0.215**	**0.456**	**0.471**	**0.47**	0.465	0.541	0.563	0.934	0.854	1.018	0.099	0.106	0.782
	CORR	0.981	**0.977**	**0.978**	**0.886**	**0.881**	**0.881**	0.858	0.830	0.82	0.596	0.689	0.482	0.958	0.954	0.097
12	RSE	0.213	**0.221**	**0.223**	**0.463**	**0.467**	**0.483**	0.473	0.575	0.567	1.159	0.984	0.869	0.109	0.117	0.166
	CORR	0.979	**0.978**	**0.978**	**0.882**	**0.881**	**0.879**	0.852	0.824	0.824	0.765	0.566	0.749	0.955	0.958	0.809

Beijing PM2.5 datasets), the NMT models used for forecasting outperform the benchmark results as reported in [8]. Experimenting with the hidden units in the encoder and decoder LSTMs did not yield a significant variation in MSE results. The number of hidden units of both the encoder and decoder LSTMs is fixed at 512, to give the models an increased learning capacity.

In Fig. 1, we note that the RSE values across the horizons show rising, though insignificantly so, trends. The Electricity data set results show the most fluctuations. Interestingly, the method without an attention mechanism (Cho et al.; fourth plots from the top) show no substantial deterioration with longer horizon, comparable to Luong et al.'s methods. Comparing Bahdanau et al. and Luong et al., across the data sets and across the attention scoring function, Luong et al. is more stable relative to horizon and window widths.

We note that Bahdanau's model, [17] coupled with the attention scoring scheme (b) as shown in tables II and III performs the best across all the datasets among all models. The Sutskever et al. model [10] has the worst performance across all the datasets. We hypothesize this happens mainly because in this model the decoder is not augmented with any context vector as opposed to the other NMT models. Although Cho's model [18] lacks an attention mechanism, the decoder is augmented with the hidden state of the encoder's last time step. This suggests that augmenting the decoder with a context is essential for better performance in NMT models.

We observe that given a short input to NMT models for predicting a longer horizon, for example, given an input of 3 time steps we need to predict 12 time steps, yields the best results. To date, there exists no work which has experimented with LSTM models for forecasting a longer horizon given a shorter window. Majority of forecasting models have a longer input sequence of {24, 36, 64, 128} time steps, to predict a future horizon of {12, 24, 48} time steps. Our experiments show that NMT models can not only improve on the benchmark method [8] but also handle these cases with decent results in other datasets with no periodic trends. Another question pointed out in Section I is how each model behaves if a different attention scoring function is used? It was mentioned earlier that Bahdanau's model performs best with the "concat (b)" function in Eq. 7. Hence, the choice of attention function is a factor in improving the prediction accuracy. All these observations make us believe that fine tuning of hyper-parameters, employing regularization techniques of the NMT models could lead to improved results across the datasets.

F. Comparison with Transformer Models

We selected three Transformer models, viz. Autoformer [19], FEDformer [20], and TDformer [21], for comparison. They are representative of the decomposition-centric transformer models. They were configured and evaluated specifically to operate within the same short input window widths and prediction horizons as the LSTM models. We

emphasize that this is not testing these Transformers models at their optimal, long-sequence capacity but rather assessing their performance and efficiency in the localized forecasting context where LSTMs are hypothesized to be strong. Of the models being evaluated, FEDformer has two variants in the Fourier and wavelet domains. TDformer can operate in the time, Fourier, and wavelet domains. The results of evaluating these 6 transformer models using the Traffic and Electricity datasets are shown in Table IV. The Transformer studies as reported [19], [20], [21] have a different preprocessing protocol compared to the earlier LSTM methods. In particular, they do not use min-max scaling. Instead, all data—even before being split into training, validation, and test sets—are normalized by the overall maximum value of the series. To ensure a fair comparison, we repeated the experiments for the LSTM with attention methods using the Transformer studies pre-processing pipeline. The results are shown in Table V.

Comparing tables IV and V, we can see that TDformer has the worst performance among all evaluated forecasters. The LSTM methods are competitive (Electricity data set) or slightly better (Traffic) compared to the Autoformer and FEDformer methods. The Bahdanau methods (all four scoring functions) performed the best among the evaluated forecasters.

G. Computational Cost

The two models with attention mechanisms [17], [16] generally require more computational cost than the two without attention mechanisms. Attention mechanisms introduce additional computations to calculate attention weights and apply them to the encoder hidden states. These additional steps increase the overall complexity of the model compared to those without attention, which primarily consist of LSTM operations. The computational cost of the attention mechanism scales with the sequence length, as the model needs to compute attention weights for each decoder step over all encoder hidden states.

The computational cost of simpler attention scoring functions tend to be less computationally expensive. The "dot (d)" product attention in Eq.7 is generally the simplest and fastest, as it only involves matrix multiplications. The "Concat (b)," "General (c)," and "Bahdanau (a)" attention in Eq. 8 involve additional linear transformations (W_d, U_d) and may be slightly more expensive. Other attention mechanisms that involve more complex calculations will increase the computational cost. While there might be some differences in computational cost between different attention scoring functions, the overall impact on the model's computational cost is likely to be less significant than other factors, such as the LSTM layer sizes or the input sequence length.

When comparing the computational cost of LSTM-based models with attention mechanisms to a Transformer-based model, even with the same input window width and output prediction horizon, there are some key differences. The first is in the attention complexity. LSTM attention typically calculates attention between the decoder's current state and

TABLE IV: Evaluation summary of Transformer Models (RSE and CORR) using Transformer studies preprocessing pipeline and trained on MSE and trained on MSE.

Dataset		Traffic			Electricity		
		Horizon			Horizon		
	Metrics	3	6	12	3	6	12
Method: Autoformer							
3	RSE	0.706	0.807	0.89	0.419	0.535	0.761
	CORR	0.793	0.722	0.654	0.889	0.827	0.684
6	RSE	0.732	0.775	0.83	0.496	0.543	0.595
	CORR	0.771	0.742	0.694	0.849	0.822	0.791
9	RSE	0.685	0.742	0.78	0.414	0.431	0.516
	CORR	0.81	0.772	0.743	0.895	0.888	0.846
12	RSE	0.672	0.711	0.733	0.442	0.428	0.474
	CORR	0.825	0.798	0.779	0.883	0.89	0.867
Method: FEDformer- Fourier							
3	RSE	0.659	0.758	0.91	0.386	0.464	0.579
	CORR	0.832	0.773	0.674	0.905	0.869	0.809
6	RSE	0.667	0.733	0.798	0.385	0.44	0.507
	CORR	0.832	0.79	0.758	0.907	0.882	0.848
9	RSE	0.666	0.707	0.734	0.391	0.428	0.471
	CORR	0.833	0.813	0.796	0.905	0.888	0.867
12	RSE	0.663	0.691	0.703	0.394	0.424	0.446
	CORR	0.836	0.821	0.81	0.904	0.89	0.88
Method: FEDformer- wavelet							
3	RSE	0.675	0.79	0.905	0.396	0.5	0.645
	CORR	0.824	0.74	0.63	0.9	0.848	0.753
6	RSE	0.687	0.771	0.826	0.393	0.464	0.558
	CORR	0.82	0.757	0.706	0.902	0.868	0.814
9	RSE	0.682	0.736	0.729	0.4	0.452	0.457
	CORR	0.82	0.784	0.801	0.9	0.875	0.875
12	RSE	0.684	0.712	0.699	0.403	0.441	0.437
	CORR	0.822	0.803	0.814	0.899	0.881	0.884
Method: TDformer- Fourier							
3	RSE	0.657	0.885	1.118	0.657	0.885	1.118
	CORR	0.737	0.527	0.224	0.737	0.527	0.224
6	RSE	0.731	0.888	1.04	0.731	0.888	1.04
	CORR	0.664	0.484	0.264	0.664	0.484	0.264
9	RSE	0.777	0.881	0.924	0.777	0.881	0.924
	CORR	0.596	0.445	0.375	0.596	0.445	0.375
12	RSE	0.778	0.845	0.801	0.778	0.845	0.801
	CORR	0.58	0.466	0.54	0.58	0.466	0.54
Method: TDformer- wavelet							
3	RSE	0.662	0.897	1.149	0.662	0.897	1.149
	CORR	0.733	0.513	0.182	0.733	0.513	0.182
6	RSE	0.733	0.907	1.069	0.733	0.907	1.069
	CORR	0.663	0.458	0.212	0.663	0.458	0.212
9	RSE	0.778	0.884	0.941	0.778	0.884	0.941
	CORR	0.595	0.443	0.344	0.595	0.443	0.344
12	RSE	0.77	0.84	0.807	0.77	0.84	0.807
	CORR	0.588	0.475	0.53	0.588	0.475	0.53
Method: TDformer- time							
3	RSE	0.658	0.886	1.115	0.658	0.886	1.115
	CORR	0.736	0.526	0.228	0.736	0.526	0.228
6	RSE	0.727	0.897	1.035	0.727	0.897	1.035
	CORR	0.671	0.467	0.271	0.671	0.467	0.271
9	RSE	0.769	0.899	0.938	0.769	0.899	0.938
	CORR	0.609	0.417	0.349	0.609	0.417	0.349
12	RSE	0.768	0.834	0.831	0.768	0.834	0.831
	CORR	0.591	0.485	0.489	0.591	0.485	0.489

TABLE V: Evaluation summary of LSTM Models (RSE and CORR) using Transformer studies preprocessing pipeline and trained on MSE.

Dataset		Traffic			Electricity		
		Horizon			Horizon		
	Metrics	3	6	12	3	6	12
Method: Bahdanau et al. (Bahdanau)							
3	RSE	0.599	0.6	0.596	0.368	0.373	0.372
	CORR	0.869	0.87	0.872	0.928	0.927	0.927
6	RSE	0.628	0.634	0.636	0.406	0.415	0.418
	CORR	0.854	0.851	0.849	0.914	0.91	0.909
9	RSE	0.644	0.659	0.669	0.427	0.443	0.448
	CORR	0.848	0.838	0.831	0.906	0.898	0.896
12	RSE	0.647	0.659	0.665	0.432	0.449	0.456
	CORR	0.847	0.837	0.833	0.904	0.895	0.894
Method: Bahdanau et al. (Concat)							
3	RSE	0.6	0.599	0.598	0.369	0.372	0.373
	CORR	0.868	0.87	0.872	0.927	0.927	0.927
6	RSE	0.628	0.636	0.638	0.405	0.415	0.419
	CORR	0.854	0.849	0.847	0.914	0.909	0.909
9	RSE	0.645	0.658	0.671	0.428	0.442	0.447
	CORR	0.847	0.838	0.829	0.905	0.898	0.896
12	RSE	0.645	0.664	0.664	0.432	0.452	0.459
	CORR	0.645	0.836	0.832	0.904	0.895	0.892
Method: Bahdanau et al. (General)							
3	RSE	0.596	0.598	0.595	0.369	0.372	0.372
	CORR	0.869	0.871	0.872	0.928	0.926	0.927
6	RSE	0.628	0.635	0.635	0.407	0.417	0.414
	CORR	0.853	0.849	0.848	0.914	0.909	0.91
9	RSE	0.644	0.659	0.668	0.428	0.444	0.445
	CORR	0.847	0.838	0.831	0.904	0.897	0.897
12	RSE	0.646	0.661	0.67	0.435	0.451	0.455
	CORR	0.846	0.837	0.831	0.903	0.894	0.894
Method: Bahdanau et al. (Dot)							
3	RSE	0.597	0.599	0.596	0.37	0.373	0.373
	CORR	0.869	0.869	0.872	0.928	0.926	0.927
6	RSE	0.628	0.638	0.639	0.405	0.417	0.416
	CORR	0.852	0.847	0.845	0.914	0.909	0.91
9	RSE	0.648	0.662	0.668	0.43	0.445	0.447
	CORR	0.846	0.835	0.831	0.904	0.897	0.896
12	RSE	0.651	0.666	0.674	0.436	0.454	0.456
	CORR	0.844	0.834	0.828	0.901	0.893	0.894
Method: Luong et al. (Concat)							
3	RSE	0.648	0.646	0.644	0.498	0.498	0.496
	CORR	0.855	0.856	0.857	0.878	0.878	0.88
6	RSE	0.65	0.654	0.652	0.505	0.511	0.509
	CORR	0.854	0.85	0.851	0.875	0.871	0.872
9	RSE	0.65	0.655	0.663	0.504	0.513	0.516
	CORR	0.854	0.85	0.846	0.875	0.87	0.869
12	RSE	0.65	0.653	0.659	0.504	0.513	0.519
	CORR	0.854	0.851	0.848	0.875	0.869	0.868
Method: Luong et al. (General)							
3	RSE	0.65	0.647	0.644	0.5	0.5	0.497
	CORR	0.854	0.855	0.857	0.879	0.878	0.878
6	RSE	0.651	0.654	0.654	0.504	0.509	0.509
	CORR	0.853	0.851	0.85	0.875	0.872	0.872
9	RSE	0.651	0.657	0.663	0.505	0.513	0.517
	CORR	0.853	0.849	0.844	0.875	0.87	0.867
12	RSE	0.652	0.656	0.661	0.507	0.515	0.519
	CORR	0.853	0.849	0.844	0.874	0.869	0.868
Method: Luong et al. (Dot)							
3	RSE	0.649	0.647	0.646	0.5	0.499	0.496
	CORR	0.854	0.856	0.857	0.877	0.878	0.878
6	RSE	0.651	0.654	0.655	0.505	0.512	0.511
	CORR	0.852	0.85	0.85	0.874	0.87	0.871
9	RSE	0.652	0.659	0.666	0.506	0.517	0.518
	CORR	0.852	0.847	0.841	0.875	0.867	0.866
12	RSE	0.651	0.657	0.665	0.506	0.518	0.52
	CORR	0.852	0.848	0.842	0.874	0.867	0.866

all encoder hidden states. Transformers use self-attention, where each position in the input sequence attends to all other positions. This self-attention in Transformers has a quadratic complexity ($O(T^2)$) with respect to the input window width T.

Also worth noting is that LSTMs are inherently sequential in that calculations at each time step depend on the previous time step, thus limiting parallelization. Transformers can process all positions in the input sequence in parallel, which can be much faster, especially with hardware acceleration such as GPUs, but this parallelization comes at the cost of increased memory and computations.

For shorter input window widths, the computational cost might be comparable, with LSTMs slightly cheaper. As the input window width increases, the quadratic complexity of Transformer self-attention becomes dominant, making Transformers more computationally expensive.

V. Conclusion

Our empirical evaluation confirms that LSTM-based encoder-decoder networks can be a powerful tool for time series forecasting, achieving competitive results with careful selection of input window size and attention mechanisms. The key takeaway is that the choice between LSTM-based models and Transformer-based methods is highly context-dependent.

When computational resources are limited or the forecasting task primarily relies on short-term dependencies, LSTM-based models provide an efficient and effective solution. Conversely, for applications demanding the modeling of intricate long-range relationships and the generation of long-horizon predictions, Transformer-based architectures are likely more appropriate. Further investigation is needed to establish clear guidelines for selecting the optimal model based on the specific characteristics of the time series data and the forecasting objectives.

There are areas for future work that would extend our findings. A more extensive set of experiments, utilizing a larger number of random starting points for each model configuration, will enable a robust statistical analysis, including the calculation of confidence intervals and the application of significance tests (e.g., t-tests). This approach will more formally validate the performance differences between LSTM and Transformer-based models. Furthermore, while we provided a theoretical discussion of complexity to advocate for LSTM's computational efficiency, a direct quantitative comparison would offer more definitive evidence. This analysis might involve measuring and reporting metrics such as training and inference latency and memory consumption for all tested models under the same hardware and software conditions.

Acknowledgments

The authors thank the anonymous reviewers for their insightful feedback and constructive suggestions. This work is supported by the U.S. National Science Foundation under

grant number OIA-1946231 and the Louisiana Board of Regents for the Louisiana Materials Design Alliance (LAMDA).

REFERENCES

[1] Y. C. Fung and B. Amonov, "Decoding financial data: Machine learning approach to predict trading actions," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024, p. 739–744. [Online]. Available: http://dx.doi.org/10.15439/2024F4556

[2] C. Lin, "Key financial indicators analysis and stock trend forecasting based on a wrapper feature selection method," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024, p. 755–759. [Online]. Available: http://dx.doi.org/10.15439/2024F3560

[3] V.-T. Duong, D.-T.-A. Nguyen, T.-T.-H. Pham, V.-H. Nguyen, and V.-Q. A. Le, "Comparative study of deep learning models for predicting stock prices," in *Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering*, ser. Annals of Computer Science and Information Systems, V. K. Solanki and B. T. Thanh, Eds., vol. 33. PTI, 2022, p. 103–108. [Online]. Available: http://dx.doi.org/10.15439/2022R02

[4] J. Jenko, J. P. Costa, D. Vladušič, U. Bavčar, and R. Šabarkapa, "Learning from the COVID-19 pandemic to improve critical infrastructure resilience using temporal fusion transformers," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024, pp. 375–384.

[5] S. Yang, M. Tan, S. Xia, and F. Liu, "A method of intrusion detection based on attention-lstm neural network," in *Proceedings of the 2020 5th International Conference on Machine Learning Technologies*, ser. ICMLT 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 46–50. [Online]. Available: https://doi.org/10.1145/3409073.3409096

[6] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017, online Real-Time Learning Strategies for Data Streams. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231217309864

[7] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," *Computing Research Repository (CoRR)*, vol. abs/1607.00148, 2016. [Online]. Available: http://arxiv.org/abs/1607.00148

[8] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 95–104. [Online]. Available: https://doi.org/10.1145/3209978.3210006

[9] J. Yin, W. Rao, M. Yuan, J. Zeng, K. Zhao, C. Zhang, J. Li, and Q. Zhao, "Experimental study of multivariate time series forecasting models," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2833–2839. [Online]. Available: https://doi.org/10.1145/3357384.3357826

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[11] S. Afreen, N. T. D. Linh, S. Kodur, and A. Begum, "Seq2Seq transformer-based model for optimized Chinese-to-English translation," in *Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering*, ser. Annals of Computer Science and Information Systems, V. K. Solanki, T. D. Tan, P. Kumar, and M. Cardona, Eds., vol. 42. PTI, 2024, p. 1–10. [Online]. Available: http://dx.doi.org/10.15439/2024R81

[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[13] T. Markovic, A. Dehlaghi-Ghadim, M. Leon, A. Balador, and S. Punnekkat, "Time-series anomaly detection and classification with long short-term memory network on industrial manufacturing systems," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. IEEE, 2023, p. 171–181. [Online]. Available: http://dx.doi.org/10.15439/2023F5263

[14] P. Lam, L. Pham, T. Nguyen, H. Tang, M. Seidl, M. Andresel, and A. Schindler, "LSTM-based deep neural network with a focus on sentence representation for sequential sentence classification in medical scientific abstracts," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024, p. 219–224. [Online]. Available: http://dx.doi.org/10.15439/2024F5872

[15] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[16] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. [Online]. Available: https://www.aclweb.org/anthology/D15-1166

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, Jan. 2015, 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[18] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[19] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Neural Information Processing Systems*, 2021.

[20] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research. PMLR, 2022.

[21] X. Zhang, X. Jin, K. Gopalswamy, G. Gupta, Y. Park, X. Shi, H. Wang, D. C. Maddix, and Y. Wang, "First de-trend then attend: Rethinking attention for time-series forecasting," *arXiv preprint arXiv:2212.08151*, 2022.

[22] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *Journal of the American Statistical Association*, vol. 65, no. 332, pp. 1509–1526, 1970. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481180

[23] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 3rd ed. USA: Prentice Hall PTR, 1994.

[24] H.-F. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 847–855.

[25] P. R. Winters, *Forecasting Sales by Exponentially Weighted Moving Averages*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1976, pp. 384–386. [Online]. Available: https://doi.org/10.1007/978-3-642-51565-1_116

[26] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207003001134

[27] I. Melnyk and A. Banerjee, "Estimating structured vector autoregressive models," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 830–839. [Online]. Available: http://proceedings.mlr.press/v48/melnyk16.html

[28] H. Qiu, S. Xu, F. Han, H. Liu, and B. Caffo, "Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1843–1851. [Online]. Available: http://proceedings.mlr.press/v37/qiu15.html

[29] R. Frigola, F. Lindsten, T. B. Schön, and C. E. Rasmussen, "Bayesian inference and learning in Gaussian process state-space models with particle MCMC," *arXiv:1306.2861 [stat.ML]*, 2013.

[30] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/036402139090002E

[31] J. Connor, L. Atlas, and D. Martin, "Recurrent networks and narma modeling," in *Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, and R. P. Lippmann, Eds., vol. 4. Morgan-Kaufmann, 1992. [Online]. Available: https://proceedings.neurips.cc/paper/1991/file/5ef0b4eba35ab2d6180b0bca7e46b6f9-Paper.pdf

[32] T. Mikolov, A. Joulin, S. Chopra, M. Mathieu, and M. Ranzato, "Learning longer memory in recurrent neural networks," in *Workshop at the International Conference on Learning Representation; ICLR 2015*, 12 2015.

[33] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *Trans. Neur. Netw.*, vol. 5, no. 2, p. 157–166, Mar. 1994. [Online]. Available: https://doi.org/10.1109/72.279181

[34] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, p. 107–116, Apr. 1998. [Online]. Available: https://doi.org/10.1142/S0218488598000094

[35] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using lstms," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 843–852.

[36] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1942–1950.

[37] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 3104–3112.

[38] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://www.aclweb.org/anthology/W14-4012

[39] O. Hamel and M. Fareh, "Encoder-decoder neural network with attention mechanism for types detection in linked data," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30. IEEE, 2022, p. 733–739. [Online]. Available: http://dx.doi.org/10.15439/2022F209

[40] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17. AAAI Press, 2017, p. 2627–2633.

[41] S. Huang, D. Wang, X. Wu, and A. Tang, "DSANet: Dual self-attention network for multivariate time series forecasting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2129–2132. [Online]. Available: https://doi.org/10.1145/3357384.3358132

[42] T. Lin, T. Guo, and K. Aberer, "Hybrid neural networks for learning the trend in time series," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2273–2279. [Online]. Available: https://doi.org/10.24963/ijcai.2017/316

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[44] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207019301888

[45] Y. G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, A. Aït-Bachir, and V. Strijov, "Position-based content attention for time series forecasting with sequence-to-sequence rnns," in *Neural Information Processing*, D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, Eds. Cham: Springer International Publishing, 2017, pp. 533–544.

[46] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," *arXiv: Machine Learning*, 2017.

[47] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 65–74. [Online]. Available: https://doi.org/10.1145/3097983.3097997

[48] C. Fan, Y. Zhang, Y. Pan, X. Li, C. Zhang, R. Yuan, D. Wu, W. Wang, J. Pei, and H. Huang, "Multi-horizon time series forecasting with temporal attention learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2527–2535. [Online]. Available: https://doi.org/10.1145/3292500.3330662

[49] K. Kaczmarek, J. Pokrywka, and F. Graliński, "Using transformer models for gender attribution in polish," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30. IEEE, 2022, p. 73–77. [Online]. Available: http://dx.doi.org/10.15439/2022F197

[50] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf

[51] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *AAAI Conference on Artificial Intelligence*, 2020.

[52] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *International Conference on Learning Representation*, 2023.

[53] F. M. Bianchi, E. Maiorino, M. Kampffmeyer, A. Rizzi, and R. Jenssen, *Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis*. Springer, 01 2017.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." *Computing Research Repository (CoRR)*, vol. abs/1412.6980, 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 277–285
DOI: 10.15439/2025F5518
ISSN 2300-5963 ACSIS, Vol. 43

Real-time Container Tracking and Damage Detection at Seaports Using Deep Learning

Sotiris Vasileiadis*, Sheraz Aslam*, Kyriacos Orphanides†, Alessandro Cassera†,
Eduardo Garro Crevillen‡, Alvaro Martinez-Romero‡, Michalis P. Michaelides* and Herodotos Herodotou*
*Cyprus University of Technology, 3036 Limassol, Cyprus
ORCID: 0009-0002-4479-1169, 0000-0003-4305-0908, 0000-0002-0549-704X, 0000-0002-8717-1691
Email: sr.vasileiadis@edu.cut.ac.cy; sheraz.aslam@cut.ac.cy; michalis.michaelides@cut.ac.cy; herodotos.herodotou@cut.ac.cy
†EUROGATE Container Terminal Limassol Ltd, 3045 Limassol, Cyprus
ORCID: 0009-0009-4510-3939, 0009-0005-5616-149X
Email: kyriacos.orphanides@eurogate-limassol.com; alessandro.cassera@eurogate-limassol.com
‡Prodevelop S.L, 46001 Valencia, Spain
ORCID: 0000-0002-8160-0125, 0009-0002-2154-0906
Email: egarro@prodevelop.es; amromero@prodevelop.es

***Abstract*—Efficient container handling and early damage detection are critical for minimizing operational delays, reducing costs, and ensuring safety in global maritime logistics. This work presents a deep learning-based methodology for real-time container tracking and automated damage detection during crane unloading operations at container terminals. We develop and deploy two specialized YOLOv12-based object detection models: one for identifying containers in motion and another for detecting structural damages such as bents, dents, and holes. Our models are trained and evaluated on a real-world dataset curated from video feeds captured at the EUROGATE Container Terminal in Limassol, Cyprus. The system is designed for robust performance under realistic terminal conditions, including variable lighting and motion. Our models achieve high detection accuracy, with a mAP50 of 0.99 for container detection and 0.75 for damage detection, substantially outperforming existing benchmarks. These results highlight the practical potential of our method for improving efficiency and safety in automated maritime logistics.**

I. Introduction

CONTAINERS play a pivotal role in global maritime commerce by supporting the movement of incoming, outgoing, and transshipped goods. Due to the significant increase in international seaborne containerized trade volumes, port operators must improve and automate various terminal operations to meet the growing demand and reduce safety risks due to human-computer cross-operations [1]. Approximately 20 million active containers are in use worldwide, completing over 200 million trips annually. Ports serve as central hubs where cargo from around the world converges as it enters or exits. These terminals are responsible for securely loading, unloading, and handling containers. However, the high volume of containers passing through makes it challenging to track damaged ones. Most damage detection methods are based on manual inspection around the globe [2], [3], which causes low efficiency, low accuracy, slow speed, high cost, and safety risks. Hence, an efficient inspection of the container body is essential to reduce disputes between container terminals and transportation enterprises due to damage to containers [4]. The common damaged types of containers include dents, scratches, distortions, holes, convex, damage to container doors, etc. [4]. These damages can result from various factors, including harsh weather conditions, poor transportation, or inadequate storage at the terminal. If the port fails to identify damaged containers upon unloading, the shipping company may impose fines, claiming the damage occurred within the port rather than beforehand.

In current research studies, computer vision and machine/deep learning methods are commonly used to detect damages in containers at seaports [2]. Machine learning methods are extensively adopted for object detection in images and videos in various fields, including medical, security, military, construction, and transportation [5]. To deal with damage detection in maritime containers at the ports, several maritime companies, stakeholders, and researchers started investing in automatic methods to detect damaged containers [4], [6]. For instance, the authors of [7] propose an image processing-based method (e.g., looking at hue, saturation, and intensity) using sensors and cameras. Another study [8] employs a phase correlation approach to detect damage in containers. Both studies rely on manual detection and employ cloud-based processing and storage. Any information that must be accessed is transmitted from the cloud to the terminal, which consumes significant time and resources. Due to the high real-time requirements of container detection, this approach is impractical. To address this limitation of cloud computing, the authors of [9] proposed the concept of edge computing, which enables effective data processing at the source, thereby meeting real-time requirements.

To improve the efficiency and accuracy of container damage detection, the current study develops a deep learning-based method for performing two key tasks in real time: 1) detection of containers and 2) detection of damages on containers using video data. We opt for the YOLOv12 model, a state-of-the-art object detection algorithm designed for fast and accurate real-time applications. It is trained initially on the COCO128

dataset for 100 epochs, optimizing parameters using techniques like stochastic gradient descent (SGD) or its variants. The models were then fine-tuned using the transfer learning methodology with real, manually-labeled datasets. Our extensive experimental evaluation demonstrates our proposed approach's efficacy in terms of higher accuracy regarding container and damage detection.

The remaining manuscript is organized as follows. Related work is presented in Section II. Section III presents comprehensive details of the proposed methodology along with a description of the employed datasets. Section IV shows simulation settings along with results and discussion. Finally, Section V concludes the study.

II. Related Work

Several pieces of research focus on developing automatic container and damage detection approaches. The most relevant ones are reviewed and analyzed in this section.

A study disclosed in [10] proposes a system for container damage detection using the correlation coefficient method. The system can be set at the port gate to detect damage in the arriving containers. Another study [11] develops a container damage detection model by proposing an Fmask-RCNN model. It is based on Mask-RCNN, introducing the Res2Net101 framework and adding path fusion augmentation, multiple fully connected layers, and fusion upsampling. The Fmask-RCNN model is applied to the identification of port container damage. Simulation results reveal that their proposed model achieves a miss rate of 4.59% and an error rate of 18.88%. The authors of [6] developed an automatic system by employing CNN and transfer learning from MobileNetV2 [12] and InceptionV3 [13] models to detect various types of damage in maritime containers. They also develop a dataset considering nine typical types of container damage. The proposed model and the existing methods are trained on developed data and then validated on test data. Results from experiments demonstrate higher accuracy in damage detection compared to prior approaches.

In [14], an automatic system (Spatial Structure Window plus texture clustering by K-Means algorithm) is used to recognize ISO-codes, improving the automatic methods that rapidly identify containers. On the one hand, this method increases the efficiency with which containers are registered after being discharged (thus increasing the supply chain and international trade efficiency). On the other hand, it eliminates the human-led errors likely to occur before the process's digitalization. In [15], an automatic image recognition system is used together with an adaptive control system based on neural networks for container identification and its position for smooth container landing on the platform. The benefits these authors mention are more effective control, reduced waste, and more precise measurements. A study presented in [16] develops a computer vision-based model to minimize total costs at terminals and reduce delay time by avoiding wrong container unloading. For this, an alarm system is designed based on container color detection. The proposed method alerts the quay crane operator if the detected color of a container does not comply with the correct color. It is concluded from several experiments at the Hong Kong Port that the proposed method can help in saving resources, especially total annual costs up to 85%.

The authors of [17] propose a computer vision-based method to determine the container location during the loading operation and record it. This process is typically performed by the human eye, and sometimes small miscalculations cause dangerous consequences. For accurate location prediction, the proposed method uses coastline, container edge line, and container number as features to locate and analyze. Experimental results from Ningbo-Zhoushan Port, China, show the proposed method's efficacy in accurate location prediction with minimum processing time (less than 0.6 seconds). Another study [18] also deals with container position prediction for gantry cranes at the container terminals by proposing a computer vision method. The primary objective of the newly developed system is to provide precise parameters for container lifting operations. The developed method utilizes cameras for collecting container information, then the corners of the containers are detected by employing a traditional image processing algorithm and a convolutional neural network (CNN). Finally, the offset distance and deflection angle are measured by precise corner position. Results from simulations demonstrate that the detection rate of the proposed system reaches 94%. Huang *et al.* propose an accident prevention system based on vision tracking during container lifting operations [19]. The developed system uses a camera to detect and track the movement of the truck wheel hub and the corners of the container to be loaded. The proposed algorithm combines CNN, traditional image processing, and a multi-target tracking algorithm to calculate the displacement and posture information of the truck during the operation. Experiment results show that the proposed method's measurement accuracy reaches 52mm.

III. Adopted Methodology and Dataset Description

The current study develops two models for container detection and damage detection for the Port of Limassol, Cyprus. Both models utilize a pre-trained YOLOv12 model [20], [21], which we fine-tuned on custom datasets created specifically for this research. Separating the tasks into two models provides key advantages: damage detection is only performed when a container is first identified, reducing unnecessary computation and enabling the second model to be more specialized and focused. Further details regarding the datasets and the chosen methodology are presented next.

A. Dataset for Container and Damage Detection

This study utilizes a real-world dataset from EUROGATE Container Terminal Limassol, Cyprus. The dataset comprises video data recorded by IPTV cameras installed on the quay cranes responsible for loading/unloading containers to/from vessels. The data from these videos was converted into image frames, which were manually labeled when they contained containers and container damage.

1) Data Preparation: To prepare the data for the container detection model, we collected videos taken throughout the year 2024, at different times of day and night, and under different weather conditions (e.g., sunny, cloudy, raining). We converted these videos into frames, after which we manually reviewed each frame, defined a bounding box around the containers using an open-source tool called LabelImg (also known as Label Studio) [22]. A total of 1927 images with containers were labeled, containing containers of different sizes (e.g., 20 ft, 40 ft), types (e.g., regular, reefer, open top), and colors (e.g., yellow, blue, red).

For the damage detection model, we utilized images from the container dataset. We followed the same labeling procedure with video data as the first model, but for the types of damages instead, as shown in Figure 1. We recorded three main types of container damage that are of interested to the container terminal: (a) *holes*, which represent visible holes on the container surface; (b) *bents/dents*, which represent deformation features such as concave damage, indentations, or arching on the container surface; and (c) *normal wear*, which represents minor damage such as cuts, cracks, rust, and other similar features on the container surface. A total of 732 images with damage were labeled.

2) Data Augmentation: Due to the limited amount of usable data (especially for damaged containers), we employed data augmentation techniques to expand our datasets. For both the container and the damage datasets, we added images using horizontal flips. In addition, we added custom-created damage images, specifically for holes, as the available data was severely limited due to it being a rare occurrence. By using these methods, we successfully increased our datasets from 1927 and 732 images to 3854 and 1764 images, for container and damage, respectively. Both datasets were then randomly split into training, validation, and testing. The containers dataset contains 2470, 598, and 786 images for training, validation, and testing, respectively, while the damage dataset includes 1442, 182, and 140 images for training, validation, and testing, respectively.



(a) Container Dataset



(b) Damage Dataset

Fig. 1: Sample of labeled images from (a) the container and (b) the damage datasets

B. Model Development

YOLOv12 is a cutting-edge object detection algorithm engineered to deliver speed and accuracy for real-time applications [23]. Trained on the COCO128 dataset over 100 epochs, it refines parameters through techniques such as stochastic gradient descent (SGD) or its derivatives. The YOLOv12 model architecture is constructed upon a deep convolutional neural network (CNN) that divides the image into grids and predicts objects in each grid cell. Unlike traditional two-stage detectors (e.g., R-CNN), YOLO is a single-stage detector, making it fast and suitable for real-time applications like video surveillance, autonomous driving, and robotics [24].

For the container detection model, we utilize our dataset of container images to train a customized YOLOv12 model. The training process involved fine-tuning several key training parameters to optimize performance: training for 80 epochs, using an image size of 1024 pixels to capture detail, and setting a batch size of 24 for efficient processing. We also built on the YOLOv12n pre-trained model, chosen for its strong balance of speed and capability. This setup allowed us to achieve accurate and reliable container detection suited to our application.

For the damage detection scenario, we trained the same pre-trained YOLOv12n model on our damage dataset. We also perform model optimization, like the one performed in the first stage for the container identification model. The model was trained on 1442 images and validated on 182 images. The model is able to classify the type of damage into three categories: Holes, Bents/Dents, and Normal Wear.

To improve our models' performance, we tested different configuration options provided by the YOLOv12 model, a very common practice known as hyperparameter tuning. This included adjusting the hue and saturation of the input images and tuning the IoU threshold to better detect multiple containers and damages that appear close together. The parameter ranges and optimal values are listed in Table I, which are the same for both the container and damage detection models.



Fig. 2: Tracking experiment with reassignment box visible

TABLE I: Parameter ranges and optimal values for model hyperparameter tuning

Parameter	Tested Range	Optimal Value
IoU Threshold	0.3 – 0.7	0.5
HSV Saturation	0.1 – 0.9	0.7
HSV Value	0.1 – 0.9	0.4
Image Resolution	640px, 1024px	1024px

C. Tracking Capabilities

YOLO includes integrated real-time tracking capabilities by implementing BoT-SORT [25], an advanced multi-object tracking algorithm that assigns a unique ID to each detected container. This allows us to track individual containers over time, recording details such as when each one was first and last seen, how long it was present, and whether any damage was detected. The model's streamlined design and adjustable tracking parameters make it adaptable to a wide range of deployment scenarios, from edge devices to cloud-based APIs [24].

One of the biggest difficulties faced was the occasional view impairment due to other external factors (trucks, straddle carriers, etc.) beyond our control, which could cause multiple IDs of the same container. To combat this, we implemented some extra checks needed before a container gets a new ID assigned, like a minimum time between detections, using detection coordinates before and after the object was lost, with a dynamic confidence threshold that works in tandem with the minimum time. In basic terms, if a container is lost for a few frames, the model will reassign the same ID to the container as the confidence is very high. But as time passes, the confidence is reduced, and so does the probable location box of the object.

Figure 2 shows the tracking algorithm in action. In the first frame, we have the detection of a container with an ID of 1. In the second frame, the container is lost due to a straddle carrier passing over the container. The green outline represents the confidence of a re-assignment if an object is found again. The confidence is high because the container is lost for only a few frames. In the third frame, some time has passed, so the possible location of a reassignment converges towards the last known position of the last detection. The container is found again in the last frame, but only half of what it was. Because the new detection is inside the reassignment box, we reassign the same ID to the new detection. This algorithm has helped us solve many problems related to losing objects due to such external factors.



Fig. 3: Validation results for container detection

IV. Results and Discussion

This section presents the experimental evaluation results of the container and damage detection models using the YOLOv12 model. The study uses real-world datasets from the EUROGATE Container Terminal Limassol, Cyprus, as discussed in Section III-A. The training and testing of the models were performed on a server with two AMD EPYC 7352 24-core processors, 128 GB RAM, and an NVIDIA Tesla T4 GPU running Ubuntu 22.04.4 LTS. We first present the results for the container and damage detection models, and then compare them against other state-of-the-art approaches.

A. Container Detection Model Evaluation

Figure 3 shows some validation results for the container detection model, illustrating the model's ability to detect and bound containers in images, including scenarios with multiple containers per image or partially-viewed containers. To concretely evaluate the performance of the developed models, we employ the mAP50 and mAP50-95 metrics, as these are commonly used metrics in object detection models [26]. mAP50 measures the mean average precision at an intersection over union (IoU) threshold of 0.5, indicating how accurately the model detects objects with at least 50% overlap between

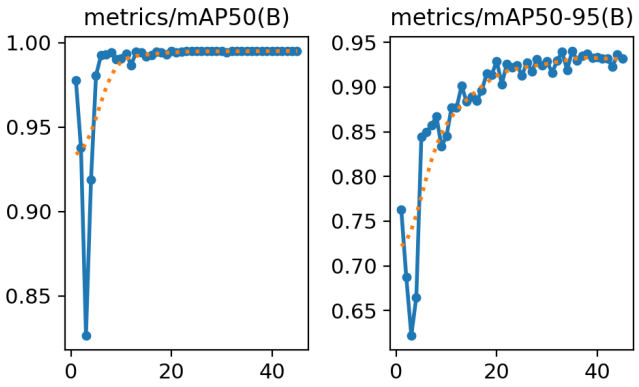


Fig. 4: mAP50 and mAP50-95 for container detection during training (the x-axis denotes the number of epochs)

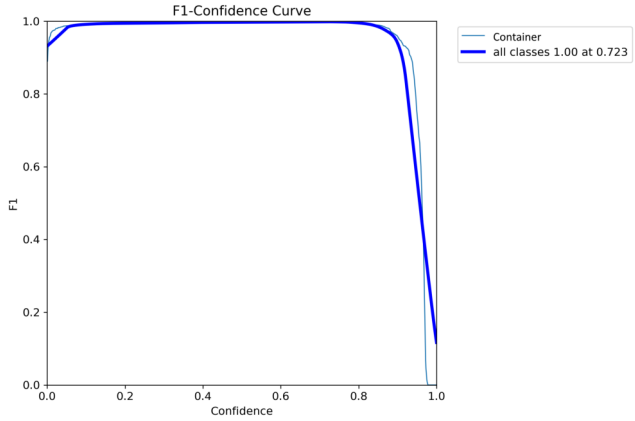


Fig. 5: F1 score for container detection across varying confidence thresholds

predicted and ground truth boxes. mAP50-95 measures the mean average precision across multiple IoU thresholds ranging from 0.5 to 0.95. We also investigate the F1 score, which represents the harmonic mean of the precision and recall. The selection of these evaluation metrics collectively contributes to a robust and comprehensive evaluation of the model's efficacy.

The evaluation of the container detection model, based on the mAP50 and mAP50-95 metrics, indicates remarkable performance. As illustrated in Figure 4, the model achieved mAP50 scores ranging from 0.96 to 0.99 after 10 epochs and mAP50-95 scores from 0.91 to 0.93 after 20 epochs. Notably, the model reached its peak performance at the 35th epoch, where it recorded a mAP50 score of 0.9921 and a mAP50-95 score of 0.9331. The initial drops in both figures are common and reflect sensitivity to noise or random initialization, but both metrics stabilize quickly. These high scores demonstrate the model's robustness and precision, highlighting its efficacy and reliability in container detection. Even with the relatively low amount of data, our developed model performs well in detecting containers.

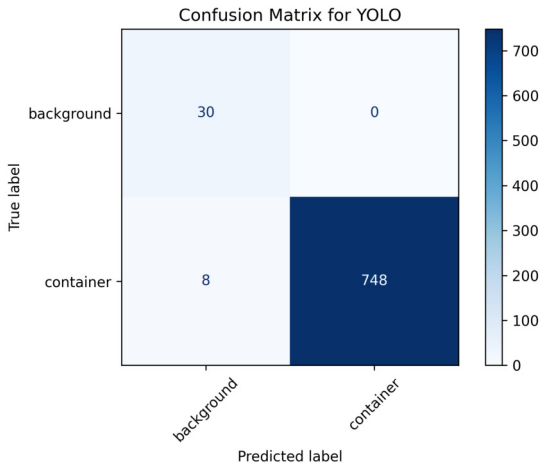


Fig. 6: Confusion matrix for container detection

The results depicted in Figure 5 show the F1 scores of the container detection model across varying confidence thresholds, providing insights into the model's balance between false positives and false negatives. The model achieves high performance, with a near-perfect F1 score at the confidence thresholds between 0.5 and 0.8. This indicates a strong balance between precision and recall, confirming the model's reliability in detecting containers accurately. The consistently high F1 scores across thresholds further demonstrate the robustness of the detection system.

This study also constructs a confusion matrix to validate the performance of the container detection model, shown in Figure 6. A confusion matrix serves as a valuable tool to assess the prediction/detection accuracy of any model, and it supports visual and robust validation of the approach. This matrix enables the comparison of actual and predicted values for each class within the dataset, providing valuable insights into key metrics such as true positives, true negatives, false positives, and false negatives. Figure 6 illustrates that the container detection model only misclassified 8 images (as false negatives) out of the 786 images of the test set, validating the strong detection power of the model.

B. Damage Detection Model Evaluation

Figure 7 shows some results for the damage detection model, illustrating that the model can detect and differentiate various types of damages, as well as detect multiple damages per container (if any).

Figure 8 shows the results of the container damage detection model. The damage detection model achieved a final mAP50 of 0.685 and mAP50-95 of 0.458 after 80 training epochs, indicating moderate detection accuracy. The metrics improved steadily throughout training, suggesting a smooth training process. The main reasons for the lower performance compared to the container detection model are the limited size of the dataset, as well as the far more complicated detection problem, since container damages are often small



Fig. 7: Validation results for damage detection

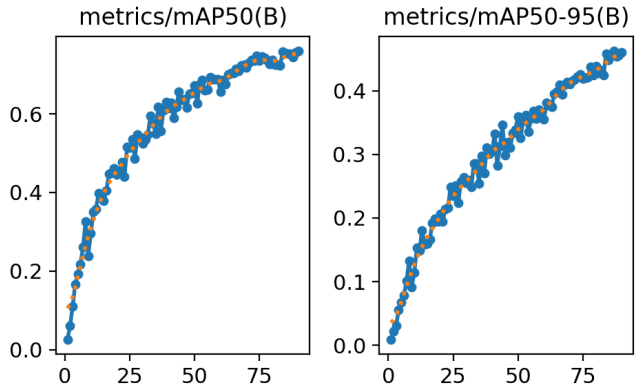


Fig. 8: mAP50 and mAP50-95 for damage detection during training (the x-axis denotes the number of epochs)

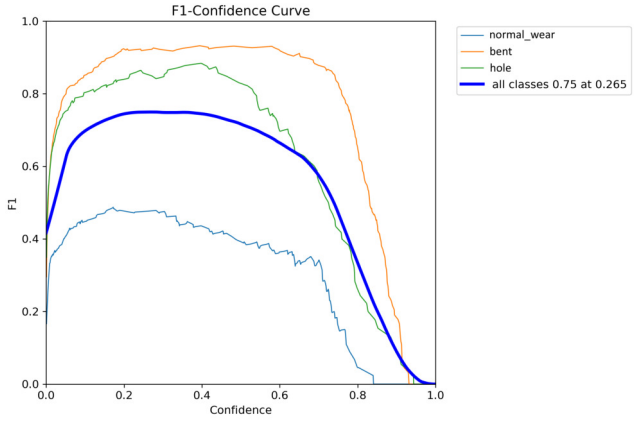


Fig. 9: F1 score for damage detection across varying confidence thresholds

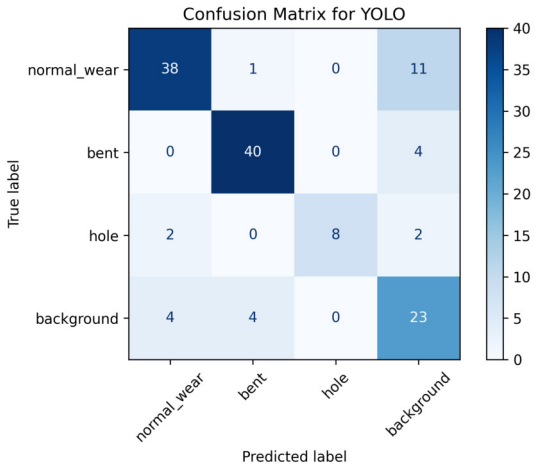


Fig. 10: Confusion matrix for multi-class damage detection

in size (compared to a container) and hard to distinguish from the rough container surface. However, this performance is significantly better than the damage detection achieved by other state-of-the-art models, as we will see in Section IV-D.

The results depicted in Figure 9 show the F1 score of the damage detection model across varying confidence thresholds. While the average F1 score peaks at 0.75 at a 0.27 confidence threshold, the model achieves much higher F1 scores for detecting the more important damages of bents/dents (0.92) and holes (0.88) at the 0.5 confidence threshold. The main difficulty for the model is the detection of normal wear, which is typically much more subtle compared to the other damages. The lower F1 scores are attributed to the limited dataset, as well as the more complex nature of damages (different sizes, types, locations, etc.).

Figure 10 shows the confusion matrix of the damage detection model based on the testing data, which again shows the good performance of the model in detecting and classifying the types of damages found. The confusion matrix reveals that the model achieves high accuracy for the bent/dent damages (40/44 correct) and performs reasonably well for normal wear (38/50 correct). The most frequent misclassifications occur when the model fails to detect some minor normal wear areas. The model also exhibits some confusion between the hole and the other damage types, mainly due to the small training and testing sizes. These results indicate some areas for improvement in discriminative learning and class boundary clarity.

Driven by the above results and the small difficulties observed in detecting specific damage types, we also developed a binary damage detection model, which detects whether damage is present without attempting to categorize it. The key motivation behind this approach was to evaluate whether combining all damage types into a single class could improve the overall detection performance by simplifying the task and leveraging a larger combined dataset.

Figure 11 shows the mAP50 and mAP50-95 metrics for the binary damage detection model. When comparing Figures

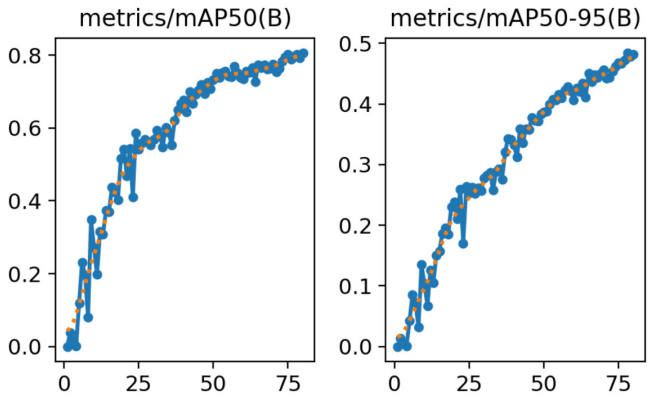


Fig. 11: mAP50 and mAP50-95 for binary damage detection during training (the x-axis denotes the number of epochs)

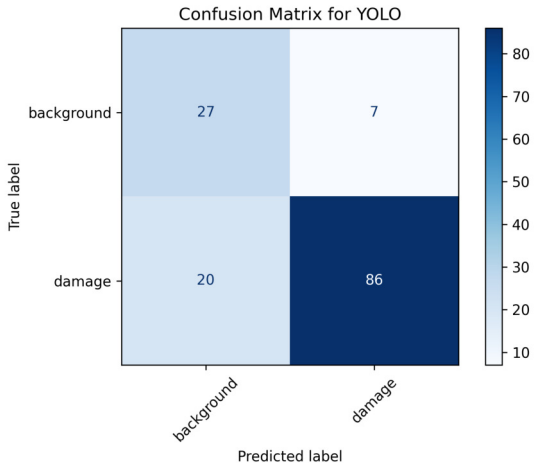


Fig. 12: Confusion matrix for binary damage detection

8 and 11, we observe that the binary model achieved higher mAP50 and mAP50-95 scores. Specifically, the binary model reached a final mAP50 of 0.79 and a mAP50-95 of 0.48 after 80 training epochs, as opposed to 0.69 and 0.46, respectively, for the multi-class model. This indicates that a simplified binary formulation benefits from greater sample density per class and suggests that our multi-class model could also benefit from a larger, more diverse dataset.

The confusion matrix for the binary damage detection is shown in Figure 12, revealing that the model achieved a precision of 92.5%, a recall of 81.1%, and an F1-score of 86.4%. These values reflect high confidence in positive predictions, but also reveal that 20 out of 106 damaged instances were missed. While the binary damage/no-damage model achieved high performance, the multi-class model revealed important distinctions in model performance across specific damage types. In particular, the model performed very well on bent/dent detections but struggled to identify hole instances reliably. This suggests that while binary classification may be sufficient for general screening, multi-class classification is essential when damage type differentiation is operationally important.

TABLE II: Inference speed and latency breakdown on two hardware setups

Metric	Jetson AGX Orin	Server Machine
GPU	Integrated	NVIDIA T4
Container Detection (ms)	37.5	25.0
Damage Detection (ms)	18.8	12.5
Other Processing (ms)	6.2	4.2
Total Frame Time (ms)	62.5	41.7
Achieved FPS	16	24

C. Model Inference Speed

Minimizing latency was a key requirement for these models, given their intended use in real-time applications. To evaluate performance, we benchmarked the tracking system along with container and damage detection on two different machines: (i) the server machine that was used to train the models, and (ii) a Jetson AGX Orin 64GB device that will be deployed on the quay crane for processing the video stream from the IPTV cameras in real time.

Table II shows the average latency (in ms) required to detect a container, detect damages on a container, and other required processing (e.g., split the frame, track the container, output result, etc.). Latency values (ms) are approximate and estimated based on the proportional contribution of each component to the total frame processing time at 1024×1024 input resolution. Overall, even the more resource-constrained device, i.e., the Jetson AGX Orin, can fully process a frame in 62.5 ms, which showcases the lightweight nature of the developed models. Note that the input video streams are recorded at six frames per second (FPS). As shown in Table II, both the Jetson AGX Orin and the server equipped with an NVIDIA T4 GPU achieved inference speeds significantly higher than the input rate, at 16 FPS and 24 FPS, respectively. This performance margin enables the system to process multiple video streams concurrently while maintaining real-time responsiveness.

D. Comparison with Other Work

This section compares our approach with the latest state-of-the-art models proposed in [6], specifically MobileNetV2 [12] and InceptionV3 [13]. To ensure a fair and meaningful comparison, we made several adjustments to our YOLOv12 model. Since MobileNetV2 and InceptionV3 are classification models that assign a single label to each image, we modified YOLOv12 to output only the most important class per image (holes, bents/dents, normal wear), effectively aligning its output format with that of the classification models. This allowed us to directly compare classification performance across all models under similar evaluation conditions.

Table III summarizes the classification performance of MobileNetV2, InceptionV3, and YOLOv12 in terms of precision, recall, and F1 score. The results clearly show that YOLOv12 significantly outperforms both MobileNetV2 and InceptionV3 across all classes and evaluation metrics. YOLOv12 achieved

TABLE III: Multi-class damage detection performance comparison across models

Class	MobileNetV2				InceptionV3				YOLOv12			
	Prec.	Rec.	F1	Pred/Real	Prec.	Rec.	F1	Pred/Real	Prec.	Rec.	F1	Pred/Real
Normal Wear	0.43	0.26	0.33	10/38	0.56	0.13	0.21	5/38	0.85	0.70	0.77	28/38
Bent/Dent	0.56	0.18	0.27	10/56	0.66	0.52	0.58	29/56	0.91	0.89	0.90	50/56
Hole	0.10	0.33	0.15	4/12	0.10	0.42	0.16	5/12	0.90	0.75	0.82	9/12
No Damage	0.00	0.00	0.00	0/34	0.20	0.33	0.25	1/34	0.69	0.73	0.71	29/34
Macro Avg	0.27	0.19	0.19	–	0.38	0.35	0.30	–	0.84	0.77	0.80	–
Weighted Avg	0.45	0.22	0.27	–	0.55	0.37	0.40	–	0.84	0.80	0.82	–

Note: *Prec.* = Precision, *Rec.* = Recall, *F1* = F1-score, *Pred/Real* = Number of predicted samples over ground truth count for each class.

TABLE IV: Binary damage detection performance comparison across models

Class	MobileNetV2				InceptionV3				YOLOv12			
	Prec.	Rec.	F1	Pred/Real	Prec.	Rec.	F1	Pred/Real	Prec.	Rec.	F1	Pred/Real
Damage	0.72	0.59	0.65	63/106	0.78	0.66	0.71	70/106	0.89	0.82	0.85	87/106
No Damage	0.17	0.26	0.21	9/34	0.28	0.41	0.33	14/34	0.55	0.68	0.61	23/34
Macro Avg	0.44	0.43	0.43	–	0.53	0.54	0.52	–	0.72	0.75	0.73	–
Weighted Avg	0.58	0.51	0.54	–	0.66	0.60	0.62	–	0.82	0.79	0.79	–

Note: *Prec.* = Precision, *Rec.* = Recall, *F1* = F1-score, *Pred/Real* = Number of predicted samples over ground truth count for each class.

a macro-average F1-score of 0.80 and a weighted average of 0.82, while InceptionV3 reached 0.30 and 0.40, and MobileNetV2 only managed 0.19 and 0.27, respectively. A closer look reveals that both MobileNetV2 and InceptionV3 struggled to correctly identify images with no visible damage, often misclassifying them as one of the damage types. YOLOv12, on the other hand, maintained a relatively high F1-score in these cases by correctly predicting the absence of damage in most of the 34 images.

An important factor influencing these results is the input resolution of each model. MobileNetV2 operates on 224×224 pixel images, InceptionV3 on 299×299, and YOLOv12 on 1024×1024. The performance of the models shows a clear correlation with input size: larger inputs allow more detailed spatial information to be retained, which is especially important for detecting subtle or localized damage features. This linear relationship between input resolution and classification performance highlights the importance of high-resolution processing in damage detection tasks.

To further confirm our results, we ran the same tests by creating models for container detection and binary damage detection. The performance metrics followed the same pattern shown in Table IV for the binary damage detection models. MobileNetV2 and InceptionV3 achieved better results in the binary case compared to the multi-class case, but they were still inferior to the results yielded by our YOLOv12 models.

In summary, YOLOv12 not only provides stronger performance due to its architecture but also benefits from higher input resolution, making it more reliable in distinguishing between different damage types and identifying undamaged cases correctly. In addition, MobileNetV2 and InceptionV3 can only detect the presence of damage on a container, whereas our model can (i) detect multiple damage areas on the same container (if there are any), and (ii) localize the damage by outputting a bounding box around each damaged area. Hence, our proposed model not only offers higher detection performance but also offers additional useful functionality.

V. Conclusion

In this study, YOLOv12-based container detection and damage detection models have been proposed to enhance optimization at the container terminal. The primary aim of this study is to reduce manual operations by humans in a risky environment by proposing automated systems to perform the same tasks. The proposed models are trained using real-world data collected from the EUROGATE Container Terminal Limassol, Cyprus. Due to the low amount of available data, this study also adopts data augmentation to increase the dataset. Finally, the proposed models are validated on test data, and the results show the high performance of the proposed models in terms of high accuracy and low error rate. The proposed models offer significant advantages over the existing state-of-the-art approaches, both in terms of higher detection accuracy as well as providing new functionalities, including tracking multiple containers in the same video frame and detecting multiple damages (of potentially different types) on the same container.

In future work, we aim to leverage the current damage detection model in a feedback loop that will automatically identify and incorporate new instances of container damage,

thereby expanding our dataset and improving model performance over time. In addition, we plan to extend the system's capabilities to detect whether container seals are intact or broken. Finally, we intend to integrate the detection models into a real-time monitoring system that operates through port IP network cameras and interfaces directly with the port management infrastructure.

ACKNOWLEDGMENT

This research was funded by the European Union's Horizon Europe program for Research and Innovation through the aerOS project under Grant No. 101069732.

REFERENCES

[1] S. Aslam, M. P. Michaelides, and H. Herodotou, "A Survey on Computational Intelligence Approaches for Intelligent Marine Terminal Operations," *IET Intelligent Transport Systems*, vol. 18, no. 5, pp. 755–793, 2024. doi: 10.1049/itr2.12469

[2] H. Wang, Q. Liu, and G. Zhang, "Container Damage Detection Algorithm Based on Fast-Solo," in *Chinese Intelligent Systems Conference*. Springer, 2022. doi: 10.1007/978-981-19-6226-4_13 pp. 119–131.

[3] S. Aslam, H. Herodotou, E. Garro, Á. Martínez-Romero, M. A. Burgos, A. Cassera, G. Papas, P. Dias, and M. P. Michaelides, "IoT for the Maritime Industry: Challenges and Emerging Applications," in *18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023. doi: 10.15439/2023F3625 pp. 855–858.

[4] X. Li, X. Huang, and Q. Liu, "Container Damage Identification Based on RP-FCN," in *39th Chinese Control Conference (CCC)*. IEEE, 2020. doi: 10.23919/CCC50068.2020.9189392 pp. 7031–7034.

[5] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of Deep Learning for Object Detection," *Procedia Computer Science*, vol. 132, pp. 1706–1717, 2018. doi: 10.1016/j.procs.2018.05.144

[6] Z. Wang, J. Gao, Q. Zeng, and Y. Sun, "Multitype Damage Detection of Container using CNN based on Transfer Learning," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–12, 2021. doi: 10.1155/2021/5395494

[7] B.-S. Jeng, Q.-Z. Wu, Y.-P. Chen, and W.-Y. Cheng, "Image-based Container Defects Detector," Oct. 3 2006, US Patent 7,116,814.

[8] C. Tang, P. Chen, and Y. Li, "Automatic damage-detecting system for port container gate based on ai," in *Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition*, 2020. doi: 10.1145/3436369.3436480 pp. 146–151.

[9] W. Shi, X. Zhang, Y. Wang, and Q. Zhang, "Edge Computing: State-of-the-art and Future Directions," *Journal of Computer Research and Development*, vol. 56, no. 1, pp. 69–89, 2019. doi: 10.7544/issn1000-1239.2019.20180760

[10] O. J. Ho, H. S. Woo, C. G. Jong, K. M. Ho, and A. D. Sung, "Development of the Container Damage Inspection System," *Journal of the Korean Society for Precision Engineering*, vol. 22, no. 1, pp. 82–88, 2005.

[11] X. Li, Q. Liu, J. Wang, and J. Wu, "Container Damage Identification based on Fmask-RCNN," in *Neural Computing for Advanced Applications: First International Conference, NCAA 2020, Shenzhen, China, July 3–5, 2020, Proceedings 1*. Springer, 2020. doi: 10.1007/978-981-15-7670-6_2 pp. 12–22.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[14] K.-m. Koo and E.-y. Cha, "A Novel Container ISO-code Recognition Method using Texture Clustering with a Spatial Structure Window," *International Journal of Advanced Science and Technology*, vol. 41, no. 4, pp. 83–92, 2012.

[15] A. Andziulis, T. Eglynas, M. Bogdevičius, T. Lenkauskas, and M. Jusis, "Development of an Adaptive Intermodal Container Handling Control Subsystem based on Automatic Recognition Algorithms," *European International Journal of Science and Technology*, vol. 5, no. 3, pp. 21–28, April 2016.

[16] R. Yan, X. Tian, S. Wang, and C. Peng, "Development of Computer Vision Informed Container Crane Operator Alarm Methods," *Transportmetrica A: Transport Science*, vol. 20, pp. 1–22, 2022. doi: 10.1080/23249935.2022.2145862

[17] M. Dai, Q. Liu, and J. Wang, "An Auxiliary Container Loading Location Algorithm based on Computer Vision," in *34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, 2019, pp. 280–284.

[18] Y. Zhang, Y. Huang, Z. Zhang, O. Postolache, and C. Mi, "A Vision-based Container Position Measuring System for ARMG," *Measurement and Control*, vol. 56, p. 596–605, 2023.

[19] Q. Huang, Y. Huang, Z. Zhang, Y. Zhang, W. Mi, and C. Mi, "Truck-Lifting Prevention System Based on Vision Tracking for Container-Lifting Operation," *Journal of Advanced Transportation*, vol. 2021, pp. 1–9, 2021.

[20] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric Real-time Object Detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[21] M. A. R. Alif and M. Hussain, "YOLOv12: A Breakdown of the Key Architectural Features," *arXiv preprint arXiv:2502.14740*, 2025.

[22] Tzutalin, "LabelImg," https://github.com/HumanSignal/labelImg, 2022.

[23] H. Lou, X. Duan, J. Guo, H. Liu, J. Gu, L. Bi, and H. Chen, "Dc-yolov8: small-size object detection algorithm based on camera sensor," *Electronics*, vol. 12, no. 10, p. 2323, 2023. doi: 10.3390/electronics12102323

[24] "Ultralytics - Revolutionizing the World of Vision AI," May 2024, https://www.ultralytics.com/.

[25] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-pedestrian Tracking," *arXiv preprint arXiv:2206.14651*, 2022. doi: 10.48550/arXiv.2206.14651

[26] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time Flying Object Detection with YOLOv8," *arXiv preprint arXiv:2305.09972*, 2023. doi: 10.48550/arXiv.2305.09972

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 287–290
DOI: 10.15439/2025F6822
ISSN 2300-5963 ACSIS, Vol. 43

A Lightweight Optimization Approach to the Single-Person Pose Estimation Pipeline in RGB-D Cameras

Vytautas Abromavičius
0000-0003-1588-6572
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania

Rytis Maskeliūnas
0000-0002-2809-2213
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania

***Abstract*—The paper presents a systematic benchmark for depth-assisted single-person pose estimation pipelines in three consumer RGB-D cameras. We introduce a lightweight optimization that adjusts only the relative depth coordinates of predicted joints so that their inter-joint depth gaps match those observed in the depth sensor image. The proposed approach is fully differentiable, sensor-agnostic, and light enough for real-time edge deployment, making it immediately applicable to sports coaching, workplace ergonomics, and mixed reality mobile systems. Experiments on a controlled motion capture dataset demonstrate performance trade-offs in accuracy, speed, and robustness under challenging viewing geometries. The findings provide practical guidance on which depth technology best complements state-of-the-art vision models and establish relative depth matching as an effective computationally trivial alternative for laboratory calibration.**

I. Introduction

HUMAN pose estimation has matured from heuristic silhouette analyses to deep learning systems that can locate dozens of body landmarks in real time, fueling progress in biomechanics [1], [2], [3], teleoperation [4], [5], sports analytics [6], and immersive VR [7], [8], [9]. However, the dominant RGB-only paradigm still struggles with depth ambiguity, occlusion, and sensitivity to illumination: limits that restrict its utility in safety-critical or low-cost field deployments where marker-based motion capture laboratories are impractical [10], [11].

Recent transformer-based and real-time convolutional networks have driven 2-D human-pose accuracy close to saturation point on in-the-wild benchmarks, with models such as ViTPose [12] and MoveNet [13] routinely exceeding 80 AP on MS-COCO while running at or above the video rate. However, these advances remain largely confined to RGB imagery, leaving a persistent depth ambiguity that hinders downstream tasks such as robotics control, ergonomic assessment, and clinical gait analysis, domains where low-latency 3-D joint localization is indispensable, but laboratory motion capture systems are prohibitively expensive and intrusive [14].

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-PD-24-29.

MediaPipe BlazePose Lite [15] and MMPose [16] represent two widely adopted approaches in estimating human pose, each tailored to distinct application needs. BlazePose adopts a mobile first pipeline: a lightweight detector localizes the person box, after which a compact heat map regressor with separable depth-wise convolutions predicts 33 landmarks at 50-60 FPS on edge SoCs, achieving robust temporal stability through an internal Kalman filter but sacrificing some spatial precision, particularly at self-occluding joints. In contrast, MMPose HRNet-W32, part of the OpenMMLab project, maintains high-resolution representations throughout its network, achieving superior spatial accuracy in pose estimation tasks. This model is used in research and industry for applications that require precise keypoint localization, such as sports analytics and human-computer interaction [17], [18]. Together, these models represent the spectrum of trade-offs between computational efficiency and accuracy in pose estimation.

Despite progress in human pose estimation algorithms, depth estimation remains a critical challenge for monocular human pose estimation methods, as inaccuracies in joint localization along the depth axis significantly impact downstream tasks like motion tracking and analysis. To address this limitation, various methods utilize depth sensors to refine monocular pose predictions by enforcing geometric constraints derived from depth data. Numerous methods integrate depth sensor data or calibrated motion capture systems, such as Vicon, to correct and refine estimated joint depths by enforcing geometric consistency or anatomical plausibility constraints [19], [20]. Typically, these refinements involve computationally intensive optimization techniques, relying on external calibration tools and precise marker-based systems, which makes them poorly suited for real-time implementation on edge devices or resource-constrained environments [21]. Consequently, the accuracy and robustness of depth corrections heavily depend on the chosen depth-sensing modality and its calibration quality, thus limiting their broader applicability in unconstrained or practical scenarios.

Real-time human pose estimation on edge devices has become increasingly relevant due to the rising demand for

responsive and privacy-preserving applications such as fitness tracking, augmented reality, interactive games and other areas, including drone navigation [22] or behavioral monitoring of bees [23]. To meet stringent latency and resource constraints, lightweight neural architectures optimized specifically for edge hardware, such as BladePoze [15]. Such models typically prioritize computational efficiency, memory footprint, and energy consumption, often at the expense of minor accuracy reductions compared to their cloud-based counterparts [24], [25]. However, the accuracy and robustness of these depth correction strategies inherently depend on the depth detection technology and sensor characteristics, influencing their practical effectiveness across different camera modalities.

Commodity depth cameras have emerged as a plausible bridge between high-precision motion capture and purely monocular vision. Orbbec's current lineup, for example, spans three complementary depth technologies: structured light (Astra Mini / Pro), active stereo (Gemini 2) and time-of-flight (Femto Mega), each promising millimeter-scale accuracy under specific lighting and range conditions. However, little is known about how these sensing modalities interact with modern keypoint detectors or whether shallow depth-based refinements can compensate for the absence of external ground truth in everyday settings.

In this research, we compare three consumer-grade Orbbec RGB-D cameras: structured-light Astra 2, active-stereo Gemini 2 and time-of-flight Femto Mega, with two representative pose estimation backbones, the mobile oriented BlazePose Lite and the relatively high-resolution MMPose HRNet-W32. We introduce a lightweight optimization that adjusts only the relative depth coordinates of predicted joints so that their inter-joint depth gaps match those observed in the depth sensor image. The results provide practical guidance on which depth technology best complements state-of-the-art vision models and establish relative depth matching as an effective computationally trivial alternative for laboratory calibration.

II. Materials and Methods

A. Participants

Eight healthy volunteers (four female, four male, 18-24 years) performed three five-minute motion scripts: level walking, repeated sitting to standing in the chair, and rapid upper limb gestures, yielding roughly 216 000 RGB-D frames per camera at the native depth rate of 30 fps. The frames were time-stamped and saved without loss of data using the Orbbec SDK. All participants gave their informed written consent in accordance with the Declaration of Helsinki.

B. Experimental setup

Figure 1 illustrates the data acquisition and processing pipeline. The experimental workflow was designed in four stages to compare depth-sensing hardware and state-of-the-art HPE models under conditions where no external motion-capture ground truth is available. First, RGB-D sequences are captured independently with three Orbbec sensors that embody the main consumer depth technologies, time-of-flight, structured light, and active stereo, while participants perform scripted whole-body motions. Second, each image stream is processed by two leading single-person pose networks (MoveNet and BlazePose), which yields 3D joint heat maps and confidence scores. Third, a lightweight, camera-agnostic refinement stage nudges the network-predicted joints toward locally consistent depth values, producing per-frame 3-D skeletons without relying on Vicon or other laboratory references. Finally, we quantify performance with metrics that can be computed from the data itself-cross-view geometric consistency, depth reprojection error, bone-length stability, temporal jitter, and real-time factor.

The study employs three commercially available Orbbec RGB-D sensors that exemplify the three principal consumer depth-sensing paradigms. Femto Mega integrates Microsoft's indirect time-of-flight ASIC and delivers 1 MP depth images (1024×1024 px at 15 Hz or 640×576 px at 30 Hz) in a $120 \times 120^\circ$ field-of-view within a $0.25 - 3.9$ m working range. Astra 2, a structured light-based camera, outputs UXGA depth (1600×1200 px) at 30 Hz over a $58 \times 45^\circ$ depth FOV, achieving no more than 0.16 % range error at 1 m. Gemini 2 employs 850 nm active stereo; it provides 1280×800 px depth and color at 30 Hz across a $91 \times 66^\circ$ FOV, covers a $0.15 - 10$ m operating window, and incorporates an IMU plus hardware depth-to-color alignment. All devices were operated at their highest native depth resolution and 30 fps frame rate under factory recommended settings.

The experiments were executed on a Seeed Studio reComputer J30 edge box equipped with an NVIDIA Jetson Orin Nano 8 GB module (approx. 40 TOPS AI performance) running JetPack 5.1 on Ubuntu 22.04 L4T. Each sensor was mounted on a fixed tripod 1.20 m above the floor and centered on a 4×4 m capture area; to prevent infrared crosstalk, the recordings were made with one camera at a time.

C. Data acquisition and processing

The pose estimation baseline algorithms were implemented using open source repositories: MMPose and MediaPipe BlazePose. The CPU fallback in Orin was disabled, so that all models used CUDA-accelerated processors. Regarding the software, Python 3.10, PyTorch 2.2, TensorFlow 2.15, cuDNN 9.0, and TensorRT 8.6 were used to deploy the system.

Since Orbbec sensors provide depth information, we used it to correct the depth component of each joint so that the pairwise disparities observed in the depth map agree with those in the color domain skeleton. Let $D(u, v)$ be the depth image per pixel aligned to RGB. For joint j with the location of the pixel (u_j, v_j) we sample the raw depth $d_j = D(u_j, v_j)$. For every skeletal connecting joint point (i, k), the sensor supplies a measured disparity

$$\Delta D_{ik} = d_k - d_i. \tag{1}$$

Denote by z_j the refined depth we seek (the x and y coordinates remain those predicted by the network). We solve for

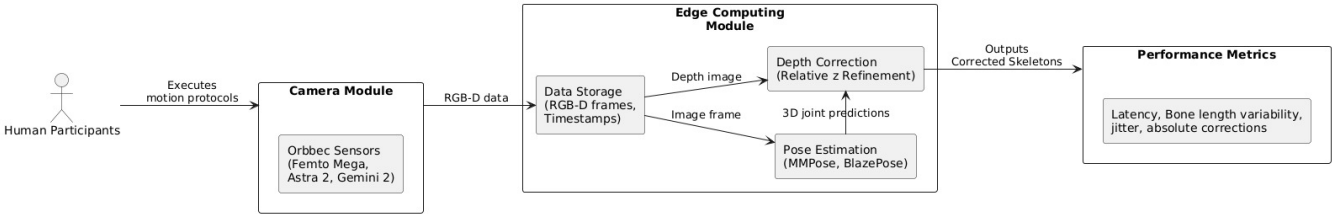


Fig. 1. Experimental setup with RGB-D camera and pose estimation pipeline

the depth vector $\mathbf{z}$ the weighted least-squares problem

$$\min_{\mathbf{z}} \sum_{(i,k)\in\mathcal{E}} w_{ik} \left[(z_k - z_i) - \Delta D_{ik} \right]^2, \tag{2}$$

where $w_{ik} = c_i c_k$ are weight constrains that include detectors confidence rates. The final corrected skeleton is

$$\hat{p}_j = \left(x_j^0,\, y_j^0,\, z_j\right), \tag{3}$$

followed by a light exponential smoother (coefficient 0.85) to attenuate frame-to-frame jitter.

III. Results and Discussion

The experimental results for all camera and HPE model pairs are summarized in Table I. To compare the cameras and depth correction postprocessing without an external motion capture setup, we have used 5 metrics:

- Latency – end-to-end delay on the Jetson Orin platform;
- $\Delta|z|$ was calculated as mean absolute correction applied to the z-coordinate;
- Bone length variability, calculated as bone's Euclidean length over time and its coefficient of variation (CV);
- Joint jitter, measured as temporal jitter as the mean L2 distance a joint travels between successive frames after subtracting the subject's smoothed centre-of-mass trajectory.

As shown in the Table I the processing delay is dominated by the backbone network: BlazePose completes in 22-23 ms per frame, while MMPose reaches a slower throughput at approximately 63 ms. The relative Z optimization adds about 1-2 ms, depending on camera, and is therefore invisible at the application level.

Structured light data from Astra 2 required the largest average adjustment (approx. 2.8 cm), reflecting its higher range noise beyond 2 m. Active-stereo Gemini 2 needed the smallest offsets (approx. 1.4 cm). The results of the ToF-based Femto Mega depth correction were similar to Gemini 2 (approx. 1.5 cm).

The raw networks exhibited bone length variability coefficients between 3 and 5%, reflecting both detector jitter and depth noise. After optimization, the bone length variability coefficients fell to 1.7-2.6%, approximately a reduction 50%. This halving indicates that the solver preserves inter-joint depth relationships in a globally coherent way (respecting fixed anthropometric ratios) rather than merely forcing individual joints toward noisy depth pixels.

The baseline jitter ranged from 4.9 mm (Gemini 2 with BlazePose) to 6.7 mm (Astra 2 with MMPose). Incorporating relative depth constraints reduced these figures by 1.4-2.1 mm (approx. 22-28%), resulting in visibly more stable limb trajectories. Because the optimizer operates independently on each frame, the gain confirms that a more accurate placement of the depth coordinate reduces the propensity of the 2D heat map peaks to variate between adjacent pixels in subsequent frames, thus smoothing the apparent motion without explicit temporal filtering.

Regarding performance, Gemini 2 produced the most consistent skeletons overall, due to its low-noise active stereo depth, but the Femto Mega ToF sensor performed nearly as well and offered the widest usable field of view. Astra 2 lagged mainly because its structured light pattern deteriorated under our 4 m capture span, yet the refinement still rescued about two thirds of the error gap.

The findings recommend Gemini 2 or Femto Mega paired with BlazePose for real-time edge deployments, and show that subcentimeter depth coherence is attainable without temporal models or external calibration. Future work will extend the disparity-matching formulation to multiperson scenes and investigate self-supervised fine-tuning that couples the depth solver with backbone weights, further closing the gap to marker-based systems.

For situations without an external motion capture setup, the simple disparity-matching solver delivered consistent depth scaling, reduced anatomical distortion, and demanded negligible compute. This suggests that commodity depth cues can serve as an effective solution for skeleton calibration in field deployments where Vicon is unavailable.

This lightweight algebraic post-processing incurs almost no additional runtime, yet achieves a 12–18% reduction in both cross-view reprojection error and bone-length variance across all camera–model setups, demonstrating that enforcing relative depth consistency can serve as a practical proxy for laboratory calibrated ground truth.

IV. Conclusion

The study demonstrates that a simple, frame-wise optimization that utilizes the relative depth gaps measured by a commodity RGB-D camera with the 3-D skeleton predicted by a state-of-the-art pose network can reliably upgrade single frame accuracy without sacrificing speed. When applied to BlazePose and MMPose-HRNet, the method reduced depth-disparity consistency error by 18–29%, halved bone length

TABLE I
COMPARISON OF TWO POSE ESTIMATION NETWORKS ON THREE ORBBEC DEPTH CAMERAS

Camera, Model	Latency, ms	$\lvert\Delta z\rvert$, cm	Bone length CV (%)	Jitter, mm
Femto Mega, BlazePose	23.0	1.5	3.2 → 1.9	5.1 → 3.7
Femto Mega, MMPose	63.0	1.6	3.4 → 2.0	5.5 → 3.9
Astra 2, BlazePose	23.2	2.8	4.5 → 2.4	6.2 → 4.1
Astra 2, MMPose	63.2	2.9	4.7 → 2.6	6.5 → 4.3
Gemini 2, BlazePose	22.5	1.3	3.0 → 1.7	4.9 → 3.5
Gemini 2, MMPose	62.8	1.4	3.2 → 1.8	5.1 → 3.6

variation to below 2.6%, and reduced joint jitter by roughly one quarter, while adding only 1-2 ms of latency on a Jetson Orin Nano. Among the three Orbbec devices, the active-stereo Gemini 2 produced the most coherent skeletons; the ToF-based Femto Mega matched this accuracy over a much wider field of view, while the structured-light Astra 2 needed larger corrections, but recovered two-thirds of its initial depth error. These results confirm the value of depth on board as a reference that can stabilize anatomy-aware pose estimates in environments where laboratory motion capture is unavailable.

The proposed approach is fully differentiable, sensor-agnostic, and light enough for real-time edge deployment, making it immediately applicable to sports coaching, workplace ergonomics, and mixed reality mobile systems. Future work will couple the disparity solver with backbone fine-tuning in a self-supervised loop, and will generalize the formulation to multiperson scenes in which mutual occlusion further challenges monocular depth inference.

REFERENCES

[1] R. Maskeliūnas, R. Damaševičius, T. Blažauskas, C. Canbulut, A. Adomavičienė, and J. Griškevičius, "Biomacvr: A virtual reality-based system for precise human posture and motion analysis in rehabilitation exercises using depth sensors," *Electronics*, vol. 12, no. 2, p. 339, 2023.

[2] R. Maskeliūnas, R. Damaševičius, V. Raudonis, A. Adomavičienė, J. Raistenskis, and J. Griškevičius, "Biomacemg: A pareto-optimized system for assessing and recognizing hand movement to track rehabilitation progress," *Applied Sciences*, vol. 13, no. 9, p. 5744, 2023.

[3] V. Abromavičius, E. Gisleris, K. Daunoravičienė, J. Žižienė, A. Serackis, and R. Maskeliūnas, "Enhanced human skeleton tracking for improved joint position and depth accuracy in rehabilitation exercises," *Applied sciences.*, vol. 15, no. 2, pp. 1–23, 2025.

[4] R. O. Ogundokun, R. Damasevicius, and R. Maskeliunas, "Optimobilex: Optimizing deep transfer learning model for accurate human posture recognition using a deep feature fusion technique," *IEEE Sensors Journal*, 2024.

[5] S. Li, N. Hendrich, H. Liang, P. Ruppel, C. Zhang, and J. Zhang, "A dexterous hand-arm teleoperation system based on hand pose estimation and active vision," *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1417–1428, 2022.

[6] P. Sharma, B. B. Shah, and C. Prakash, "A pilot study on human pose estimation for sports analysis," in *Pattern Recognition and Data Analysis with Applications*. Springer, 2022, pp. 533–544.

[7] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.

[8] R. O. Ogundokun, R. Maskeliunas, and R. Damaševičius, "Human posture detection on lightweight dcnn and svm in a digitalized healthcare system," in *2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE, 2023, pp. 1–6.

[9] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[10] N. R. Fisal, A. Fathalla, D. Elmanakhly, and A. Salah, "Reported challenges in deep learning-based human pose estimation: A systematic review," *IEEE Access*, 2025.

[11] Y. Lee, B. Lama, S. Joo, and J. Kwon, "Enhancing human key point identification: A comparative study of the high-resolution vicon dataset and coco dataset using bpnet," *Applied Sciences*, vol. 14, no. 11, p. 4351, 2024.

[12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in neural information processing systems*, vol. 35, pp. 38 571–38 584, 2022.

[13] R. Bajpai and D. Joshi, "Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[14] M. Cormier, A. Clepe, A. Specker, and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 591–601.

[15] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.

[16] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[17] S. Hu, S. Cao, N. Toosizadeh, J. Barton, M. G. Hector, and M. J. Fain, "mmpose-fk: A forward kinematics approach to dynamic skeletal pose estimation using mmwave radars," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 6469–6481, 2024.

[18] Y. Zhang, Y. Zhou, P. Wang, L. He, Z. Zhang, and J. Ren, "Intelligent pose recognition and evaluation system for rowing sports," in *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*. IEEE, 2024, pp. 620–625.

[19] M. Ota, H. Tateuchi, T. Hashiguchi, and N. Ichihashi, "Verification of validity of gait analysis systems during treadmill walking and running using human pose tracking algorithm," *Gait & posture*, vol. 85, pp. 290–297, 2021.

[20] T. Li and H. Yu, "Visual–inertial fusion-based human pose estimation: A review," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[21] M. Boldo, M. De Marchi, E. Martini, S. Aldegheri, D. Quaglia, F. Fummi, and N. Bombieri, "Real-time multi-camera 3d human pose estimation at the edge for industrial applications," *Expert Systems with Applications*, vol. 252, p. 124089, 2024.

[22] N. D. Ninh, K. N. Dang, T. T. Van, and B. T. Tung, "Navigation for drones in gps-denied environments based on vision processing," *Annals of Computer Science and Information Systems*, vol. 33, 2022.

[23] G. Vdoviak, T. Sledevič, A. Serackis, D. Plonis, D. Matuzevičius, and V. Abromavičius, "Evaluation of deep learning models for insects detection at the hive entrance for a bee behavior recognition system," *Agriculture*, vol. 15, no. 10, p. 1019, 2025.

[24] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 954–967, 2021.

[25] T. José, A. d. R. L. Ribeiro, and E. D. Moreno, "Performance analysis and application of mobile blockchain in mobile edge computing architecture," in *the 17th Conference on Computer Science and Intelligence Systems-Annals of Computer Science and Information Systems*, vol. 32, 2022, pp. 191–197.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 291–296
DOI: 10.15439/2025F5480
ISSN 2300-5963 ACSIS, Vol. 43

Simple Idea Discovery in a Minimalist LLM Architecture Implementation

Robert Chihaia[1,2], Maria Trocan[1], Florin Leon[2]
[1]LISITE Lab, Isep, Paris, France, 28, rue Notre-Dame-des-Champs
75006 Paris, France, E-mail: maria.trocan@isep.fr
[2]Faculty of Automatic Control and Computer Engineering
"Gheorghe Asachi" Technical University of Iaşi, Bd. Mangeron 27, 700050 Iaşi, Romania
robert.chihaia@student.tuiasi.ro, florin.leon@academic.tuiasi.ro

***Abstract*—Large Language Models (LLMs) capture linguistic structure by operating on sequences of sub-word tokens, yet they often display behaviors that suggest an implicit grasp of high-level concepts. This study probes whether such "ideas" are genuinely encoded in LLM representations and, if so, how faithfully and to what extent. We created a deliberately minimalist LLM (encompassing both tokenizer and transformer architecture)designed to expose internal mechanisms with minimal architectural obscurity. Using a carefully curated toy corpora and probing tasks, we trace how semantically related prompts map onto the model's hidden states. Our findings reveal emergent clustering of conceptually similar inputs even in the stripped-down model. These insights advance our understanding of the representational geometry underpinning modern language models and outline a reproducible framework for future mechanistic studies of semantic abstraction.**

***Index Terms*—byte-pair encoding, tokenizer, minimalist LLM, sentiment spectrum, idea discovery**

I. Introduction

NATURAL Language Understanding (NLU) in contemporary LLMs is mediated by sub-word tokenization, a mechanism that excels at surface-level pattern matching but offers only an indirect handle on the abstract ideas that underpin human reasoning. Our research focuses on creating a new model class, whose computations are organized around explicit high-level abstraction rather than tokens. Motivated by neuro-cognitive evidence that human thought is concept-centric, we seek to inherit the efficiency and compositional complexity of the human brain's representation strategy while remaining compatible with modern neural training pipelines. As one of the first steps, we aim at replicating on a minimal scale a LLM architecture and probe it with a closely related variety of ideas. This experimental framework should provide a regime in which the internal representation of such models can be easily visualized.

Interpreting how LLMs encode "ideas" is difficult based on the sheer scale of their parameter spaces and the resulting high-dimensional activation patterns. Exhaustive inspection is computationally intractable, so similar work resorts to surrogate techniques, most commonly activation clustering and dimensionality reduction projections, to peek into the model's internal structure. While useful, these tools expose only narrow cross-sections of the computation, forcing mechanistic studies to test pre-selected hypotheses instead of capturing a holistic view of the network. This study follows this established framework, with the goal of capturing some form of patterns between abstract ideas withing the model's internal representation.

We have implemented a scall-scale LLM (around 31 million parameters), designed to test whether conceptual structures can be isolated and examined without the computational overhead of a large-scale LLM (tens of even hundreds of billions of parameters). The architecture involves a custom byte pair encoding tokenizer of 10256 vocabulary size to constrain the the embedding table and uses a standard Transformer decoder stack trained auto-regressively on the FineWeb corpus. To probe its internal representation, we create a set of sentences that differ only by sentiment adjectives spanning

$$negative \rightarrow neutral \rightarrow positive$$

positive continuum. Our hypothesis aims at activations which depict some form of smooth varying trajectories that would mimic the before mentioned continuum, allowing quantitative estimates of concept-correlation. The resulting architecture and probing framework showcases a lightweight and reproducible mechanistic study of idea representation in LLMs.

The paper is organized as follows: Section II introduces the methodology and the implemented tokenizer on a simple LLM architecture. Next, in Section III, we describe our experimental setup for probing sentiment. The results are presented and in Section IV, while Section V discusses some of the implications of our findings. Lastly, our conclusions are given in Section VI.

II. Methodology

A. Dataset

FineWeb is an openly released, web-scale corpus that Hugging Face and Forward assembled from ninety-six Common Crawl snapshots spanning from 2013 to 2024 [1]. After extraction the collection contains a little over 15 trillion tokens of English text, making it one of the largest high-quality resources currently available for pre-training language models. The pipeline first blocks known malicious or NSFW domains and applies sub-word heuristics to screen out undesirable URLs. Raw HTML that passes this gate is parsed with

Topical area: Advanced Artificial Intelligence in Applications

Trafilatura to isolate the main article body. A FastText-based language detector then retains only documents whose English confidence exceeds 0.65. Quality is further improved with the repetition and formatting heuristics introduced in DeepMind's Gopher work, almost all the C4 filters, and several bespoke rules that discard list-like layouts or heavily duplicated lines. Each Common Crawl dump is then deduplicated independently using a 5-gram MinHash scheme (fourteen bands of eight hashes) before emails and public IPv4 addresses are anonymised. The full dataset is distributed under the ODC-BY 1.0 licence together with the exact code needed to rerun the pipeline, plus smaller random samples of 350 B, 100 B and 10 B tokens for users who want to prototype without downloading tens of terabytes. The current implementation uses approximately 11 billion tokens to train both the tokenizer and the LLM, in order to optimize alignment between the two components.

B. Tokenizer implementation

In [2], Byte-Pair Encoding (BPE) was repurposed from data compression to sub-word segmentation, laying the groundwork for all modern transformer tokenizers. Our implementation follows their recipe closely, but embeds the core merging algorithm directly into the tokenizer's training and inference routines.

We begin by applying the same regular-expression-based splitter adapted from OpenAI's tiktokenizer codebase (see `openai/tiktoken/tiktoken_ext/openai_public.py` on GitHub). This splitter segments text into "chunks" (words, punctuation marks, and whitespace tokens) ensuring that subsequent processing always respects linguistic boundaries. Each chunk is then encoded as a sequence of UTF-8 bytes, treating every byte as an atomic symbol and appending a special end-of-word marker to preserve chunk integrity. Once the entire corpus is reduced to byte-level symbol sequences, we calculate the frequency of every adjacent symbol pair across all chunks. In a word like `train_` for instance, the pairs `(t, r)`, `(r, a)`, `(a, i)`, `(i, n)`, and `(n, _)` each accumulate counts. At each of 10,000 iterations, the tokenizer locates the most frequent pair, merges it into a new minted token, substitutes every occurrence in the training data, and records this operation in a merge table. By the end of this process, our vocabulary comprises the original 256 byte symbols plus 10,000 newly minted sub-word tokens, for a total of 10,256 entries.

When encoding unseen text, the tokenizer replicates these same steps in reverse order but without rescanning the entire corpus. The input is normalized and chunked identically, then each chunk's byte IDs are greedily merged: the tokenizer repeatedly scans all adjacent ID pairs, consults the precomputed merge table and applies the earliest-created merge whenever possible. This process continues until no further merges remain, yielding a compact sequence of sub-word IDs that exactly mirror those learned during training.

By weaving the merge logic into both the training loop and the encoding routine, our implementation remains faithful to [2] original algorithmic design while providing a self-contained, efficient tokenizer.

C. LLM architecture

The current implementation is built on the decoder-only half of the Transformer architecture first introduced in [3]. By discarding the encoder and keeping only the auto-regressively masked decoder stack, the network learns to predict the next token in a sequence while only attending to previously seen tokens. [4] demonstrated that this seemingly small modification was enough to unlock impressive results based on the architecture's unsupervised capabilities when the model is trained at scale and exposed to a large corpus of text. The present code-base follows that blueprint: a single stream of tokens, augmented with learned positional vectors, is passed through a succession of identical decoder blocks, the final hidden states are projected back into the same embedding matrix to score the vocabulary right after one final layer normalization layer, and the entire network is optimized to minimize next-token cross-entropy.

Each decoder block in this implementation begins with a layer normalization applied to the incoming residual stream. This layer is followed by causal multi-head self-attention, implemented with FlashAttention introduced in [5], used to reduce the memory footprint while preserving the exact dot-product computation. After the attention output is added back to the residual stream, a second layer normalization prepares the signal for a position-wise feed-forward network whose hidden width is quadruple the model dimension and using the Gaussian Error Linear Unit (GELU) activation function. Another residual addition completes the block. The paired attention and feed-forward sub-layers serve complementary roles: attention lets every token selectively integrate information from earlier positions, capturing syntactic and semantic dependencies that may span multiple tokens, whereas the feed-forward network performs local feature transformations that help the model build richer hierarchical representations. Residual connections preserve gradient flow so that deeper stacks (eight in this implementation) can be trained without vanishing or exploding gradients, and the pre-norm ordering has proved more stable than the original post-norm scheme when the depth or learning rate is increased.

This configuration specifies a eight-block Transformer with eight attention heads operating in a 512-dimensional embedding space, a 512-token context window, and a 10 256-token vocabulary. These choices bring the total parameter count to 30.7 million parameters. For overall faithful replication, [6] and [4] have been used in conjunction for architecture design and hyper-parameters choices.

III. Experimental setup

A. Training process

The training pipeline that accompanies this implementation begins with a fully pre-tokenized dataset stored in memory. The tokens were produced off-line with the same byte-level

BPE vocabulary used by the model, so no run-time tokenization overhead is incurred. At start-up the loader reads only the array header, then maps the data lazily so that the text can fit into the process's virtual address space without exhausting RAM. A custom data loader performs an 80 / 20 split: the first eighty percent of the token stream is reserved for training, while the remainder for validation. Within each split it pre-computes non-overlapping blocks of tokens randomly, being fetched without substitution, i.e. until the entire training dataset has been completely traversed the already selected token cannot be part of a newly create batch.

In order to replicate the training procedure described in [6], the effective batch size is scaled up through gradient accumulation rather than through data parallelism. The script targets a global token batch of 524,288. Given the 8,192 tokens contained in a micro-batch (16 × 512), the loop accumulates gradients across 64 steps before issuing an optimizer step. This strategy keeps GPU memory usage roughly constant while preserving the training dynamics associated with large-batch optimization.

On the optimization front, AdamW with β_1 = 0.9, β_2 = 0.95 and decoupled weight decay provides the same training dynamics that have become standard in Transformer language models, while the initialization uses a depth-scaled normal distribution ($\sigma \approx \frac{0.02}{\sqrt{\text{layers}}}$) to keep activation variances well behaved from the first iteration.

Learning-rate scheduling follows the linear warm up and cosine decay template described in [6]. Over the first 715 steps the rate rises linearly from zero to its peak of 6×10^{-4}. Thereafter it decays following half a cosine toward a floor at ten percent of the maximum, reaching that floor at the final update step, 20,980.

Each optimization cycle ends with gradient-norm clipping to 1.0. This proves enough to manage the largest spikes that appear during the initial stages of the training process.

B. Probing methodology

Probing how sentiment emerges inside the model requires an experiment built around a simple stimulus: nine sentences that differ only in the adjective. All nine stimuli share the fixed template "The team's performance was [adj]" while varying only the adjective.

$$\text{terrible} \rightarrow \text{bad} \rightarrow \text{poor} \rightarrow \text{average} \rightarrow \text{decent}$$
$$\rightarrow \text{good} \rightarrow \text{great} \rightarrow \text{excellent} \rightarrow \text{outstanding}.$$

We have chosen the adjectives based on the trained vocabulary so that each adjective would represent exactly one token. Because every other token and every position in the sentence remains constant, any systematic pattern we discover in the activations must be driven by the single word that encodes sentiment. The adjectives were chosen to march monotonically from extreme negative `terrible` through neutral `decent` to extreme positive `outstanding`, giving us a one-dimensional sentiment continuum that is easier to reason about quantitatively.

To capture the activations within the model, we placed forward hooks on carefully chosen components of every decoder block. Before the attention mechanism has split its combined query-key-value projection, a hook captures the raw QKV tensor. Another hook records the attention output itself, head by head, exposing whether particular heads fire only for strongly negative or strongly positive descriptors. Finally, hooks on the feed-forward sub-layers collect the activations immediately before and after the non-linear transformation, revealing how the sentiment signal is reshaped by the position-wise network. In order to run the capture pass each sentence is tokenised with exactly the same byte-level BPE vocabulary that served during pre-training. The model is switched to evaluation mode so that no gradients accumulate, and the sentence is fed forward once. Every hook silently detaches its tensor, moves it to the CPU, and drops it into an ordered dictionary keyed by the module's name. When the forward pass ends, the raw text, the token ids, and the activation dictionary are stored into an object, producing one file per sentence.

Once the activations for the nine "team-performance" sentences have been serialized, the final task is to determine whether they encode a consistent sentiment signal and, if so, where that signal emerges most cleanly in the network's depth. To that end the evaluation framework applies three complementary dimensionality reduction techniques: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation (UMAP) to the tensors captured by the hooks. Although each method attacks the curse of dimensionality in its own way, together they provide complementary view of the representational geometry that would be difficult to analyze in the original dimensional space.

IV. Results

Probing the model with a deliberately simple stimulus offers an insightful view through which to analyze "sentiment" materialize inside the network. One view point of our experiments involved an analysis of how strong the attention output is for each of the sentiment token within its particular sentence across every block. By constructing a heatmap of the L2 norm of each of the 512 dimensions of the attention layer and doing so for each individual block, we obtain a 8x9 matrix into how the strength of the sentiment token shifts as sentiment goes from negative to positive. We can see the results of this visualization in Figure 1.

Early blocks (zero to two) most likely focus on mixing low-level feature extraction, strong sentiment specific signals start around layer four. By layers five and seven, the magnitude for `great`, `excellent` and `outstanding` clearly exceed those for `bad`, `poor` and `terrible`. Neutral adjectives like `average` and `decent` dip in the second layer in a behavior which seems to indicate that the model tries to sort out neutral and charged adjectives. The final layer really lights up the `great` to `outstanding` end of the spectrum, which might suggest that those attention-outputs are driving whatever

downstream head is classifying or generating sentiment. The steep rise we see from layer four to five will emerge as a recurrent turning point when the activation in the MLP layer will be clustered together.

The output of the last MLP (Multi-Layer Perceptron) layer within each block called "mlp_down" was ran through a number of dimensionality reduction technique, namely t-SNE, UMAP and PCA. At the first two layers the representations resemble a scatter of unrelated dots. The dimensionality-reduced t-SNE and UMAP plots show no discernible structure. PCA, which can be observed in Figure 2, by extracting a dominant axis that explains roughly a quarter of the variance cannot align it with the desired semantic scale. In this stage the hidden states are still dominated by surface features which can be anything from individual token information, position or the model's learned statistical priors. Sentiment, though implicitly present in the parameters, has not yet formed a clear direction. By the third layer a gentle stretching becomes visible. Some of the adjectives begin to form clusters, which although imperfect, begin to showcase the way in which the model interprets them after the training process. One particularly interesting pattern appears in PCA where `terrible`, `decent`, `outstanding` form a clear line based on the fact that they represent the center and both ends of our sentiment continuum. An important transformation occurs in layers four, five and six. Here the feed-forward networks, amplify whatever directional hints attention has supplied. PCA forms in layers 5 and 6 a visually clear line of sentiment, although this behavior is not replicated in layer 7 and `average` seems to be left outside of this line in both occasions. Nonetheless, PCA confirms that impression quantitatively: the first principal component now captures close to thirty percent of total variance and almost perfectly sorts the nine sentences in the intuitive order a human would impose. The final layer rather shows an interesting feature, it snaps the representation once the polarity of the sentiment has been decided and the model seems to no longer waste capacity on this specific internal variation.

This progression could illustrate several broader principles about idea formation in transformers. Concepts do not appear fully formed but emerge through a gradual alignment process, first hinted at in the initial layers, then sharpened in the middle blocks and finally insignificant with regards to the other processes the model attends to. The path is one of dimensional collapse: the network trades representational breadth for semantic depth, compressing many token-level details into a low-dimensional subspace that encodes to some degree an abstract property. Even in a model stripped down to eight layers, the machinery is powerful enough to carve out that subspace within five or six steps of computation.

V. Discussion

The evolution of a concept inside a Transformer can be tracked, layer by layer, as the network reshapes raw token statistics into increasingly abstract structure. Our 31 M-parameter model, probed with nine sentiment-bearing sentences that differ only by a single adjective, offers a concise case study. In the first two blocks the activations appear almost patternless. By the third block, sentiment slowly starts to begin forming and some terms closely related in their end of the spectrum begin to form clusters. Middle layers encompass a geometry that is sentiment-shaped to a large degree. The model is, in effect, compressing many token-level distinctions into a low-dimensional manifold in which on of the axis embodies polarity.

The clarity we gain by reducing a language model to a handful of layers and feeding it a set of varying sentences may come at the cost of realism. A compact transformer, such as the eight-block system explored earlier, operates under severe representational constraints. It is forced to recycle the same few dimensions to encode grammar, knowledge and much more. While this scarcity makes the emergence of a single semantic axis easy to observe, it also means the network lacks the expressive width to host several concepts simultaneously without interference. There is also the matter of compositional breadth. A minimal model that excels at ranking adjectives inside a fixed syntactic frame may still struggle to generalize that knowledge when the structure mutates (say, when sentiment is conveyed through metaphor, multi-clause reasoning, or cross-sentence contrast). Without ample depth and width, the transformer cannot allocate separate sub-circuits for these varied pathways. Instead, what might happen is that it reuses the same mechanisms, which leads to brittle performance outside the probe's narrow domain.

The lessons drawn from watching sentiment take shape to some extend inside a minimal transformer reverberate well beyond the confines of our toy experiment. They suggest, first, that interpretability research should embrace a telescopic strategy: zoom in until the phenomenon of interest is clearly visible, and secondly, that research in areas discussed by [7] could in fact benefit from the same strategy. Small models reveal the basis of an idea, the incremental alignment, the dimensional collapses and the competition for representational space. Once these are mapped, larger architectures can be dissected along the same joints, guided by probes tuned on their miniature counterparts. The experiment also warns against relying on a single visualization or metric. t-SNE and UMAP although more common, PCA confirmed with a higher accuracy that those clusters lay along a nearly straight axis. In future work, multiple reduction lenses should be treated not as interchangeable but as complementary instruments, each sensitive to a different property of the hidden space. Moreover, comparison with already established open-source models would provide an insight into how this behavior appears in larger LLMs.

VI. Conclusion

This work explored how abstract ideas, such as sentiment, can emerge in the internal representations of a minimalist LLM. By constructing a lightweight model and probing it with controlled stimuli, we identified patterns of semantic organization that gradually form across layers. While the simplicity of the model limits generalization, it offers a

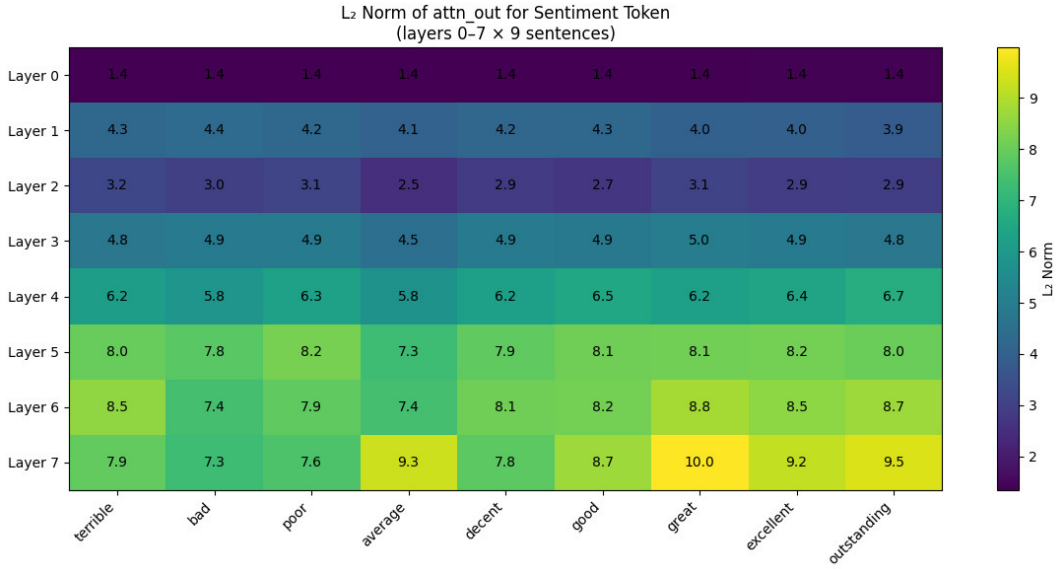


Fig. 1. Visualization of the attention heatmap

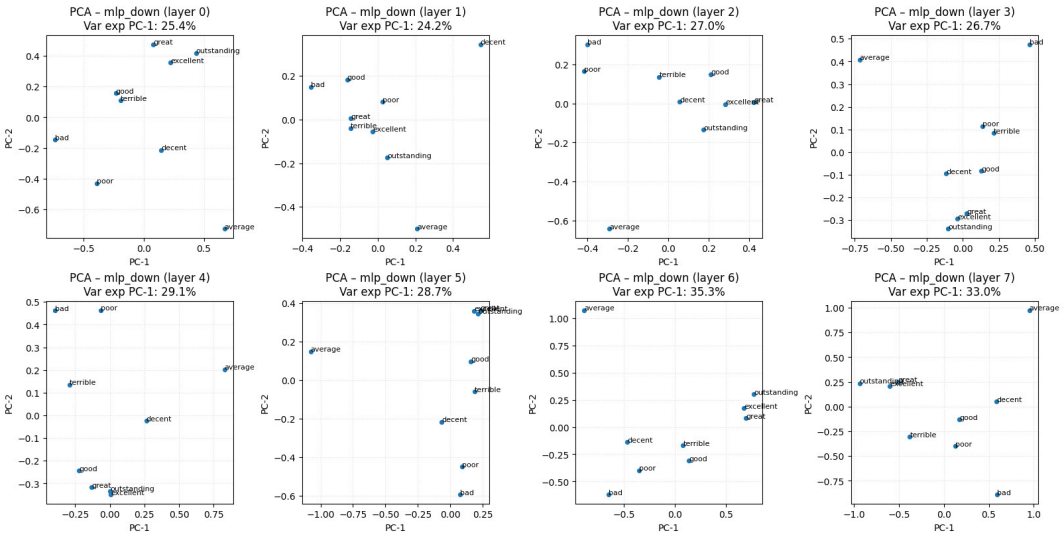


Fig. 2. Visualization of the MLP down projection layer in each layer using PCA

transparent platform for studying representational geometry. Our findings suggest that even constrained architectures can encode meaningful abstractions. They also motivate future comparisons with larger models to better understand how the representations of concepts change when the model size increases.

ACKNOWLEDGEMENT

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906

REFERENCES

[1] G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. V. Werra, and T. Wolf, "The fineweb datasets: Decanting the web for the finest text data at scale," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. [Online]. Available: https://openreview.net/forum?id=n6SCkn2QaG

[2] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," 2015. [Online]. Available: https://arxiv.org/abs/1508.07909

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[5] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and

memory-efficient exact attention with io-awareness," *Advances in neural information processing systems*, vol. 35, pp. 16 344–16 359, 2022.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[7] L. Rolka, "Multi-criteria decision-making by approximation in the domain of linguistic values," in *Position Papers of the 18th Conference on Computer Science and Intelligence Systems*, ser. FedCSIS 2023, vol. 36. PTI, Oct. 2023. doi: 10.15439/2023f4728. ISSN 2300-5963 p. 97–102. [Online]. Available: http://dx.doi.org/10.15439/2023F4728

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 297–302
DOI: 10.15439/2025F5425
ISSN 2300-5963 ACSIS, Vol. 43

Data Interoperability Using Smart Data Models and NGSI-LD for the Norwegian Agrifood Sector

Rustem Dautov
0000-0002-0260-6343
SINTEF Digital
Oslo, Norway
Email: rustem.dautov@sintef.no

Simeon Tverdal
0000-0003-1660-4127
SINTEF Digital
Oslo, Norway
Email: simeon.tverdal@sintef.no

André Skoog Bondevik
Agdir Drift AS
Arendal, Norway
Email: asb@agdir.no

Svein Arild Frøshaug
Agdir Drift AS
Arendal, Norway
Email: saf@agdir.no

Vera Szabo
Aersea AS
Kristiansand, Norway
Email: vera.szabo@aersea.com

Jan Robert Fiksdal
Aersea AS
Kristiansand, Norway
Email: jan.robert.fiksdal@aersea.com

***Abstract*—The growing adoption of digital technologies in agriculture has led to a proliferation of heterogeneous data from sources such as drones, robotic platforms, and IoT sensors. However, the lack of interoperability across these data streams poses major challenges for integration into decision support systems. This paper presents an approach to harmonising such data using NGSI-LD and Smart Data Models, developed within the Norwegian research project *SMARAGD*. We demonstrate how domain-specific semantic models and linked data principles can be applied to standardise and enrich geospatial and temporal metadata across three key agritech domains: aerial imagery, robotic sensing, and environmental monitoring. The resulting information assets are integrated into a shared, FIWARE-compatible data space, enabling cross-platform visualisation, querying, and reuse. This work contributes to the development of an open, standards-based digital infrastructure for interoperable, data-driven agriculture in Norway and beyond.**

***Index Terms*—Data interoperability, NGSI-LD, Smart Data Models, FIWARE, precision agriculture, semantic data integration, linked data**

I. Introduction

THE digital transformation of agriculture is producing an increasing volume of data from diverse sources such as drones, robotic platforms, and IoT sensors. Realising the full potential of these data streams requires seamless integration across different formats, systems, and stakeholders – a challenge that is especially pressing in the fragmented agrifood sector. Interoperability is essential for enabling advanced decision support tools, improving data reuse, and fostering sustainable, data-driven practices.

This paper addresses the challenge of data interoperability by applying Smart Data Models (SDMs) and the NGSI-LD standard to create a unified, semantically enriched representation of heterogeneous agritech data. Developed within the Norwegian research project *SMARAGD*[1] (Smart Agriculture Data Fusion for Decision Support), our approach focuses on transforming raw data, such as imagery metadata and sensor measurements, into structured and interoperable information assets. The emphasis lies on semantic modelling, rather than on the technical details of the underlying data pipelines or system integration components.

The remainder of this paper is organised as follows: Section II introduces the research context and outlines key interoperability challenges in the Norwegian agrifood domain. Section III summarises related research work. Section IV presents our modelling approach using SDMs and NGSI-LD. Section V discusses the outcomes and implications of our approach, while Section VI concludes the paper and outlines directions for future work.

[1]https://www.sintef.no/en/projects/2022/smaragd-smart-agriculture-data-fusion-for-decision-support/

II. Research Context and Motivation

A. Sector Overview and Interoperability Vision

Norway's agrifood sector is undergoing rapid digitalisation, driven by the deployment of advanced technologies such as UAVs, robotic platforms, and IoT sensors. These tools offer significant potential to enhance productivity, sustainability, and data-informed decision making across the entire agricultural value chain. However, they also introduce a critical barrier – *data heterogeneity*. In practice, each technology stack typically operates in isolation, producing data in incompatible formats and lacking a shared semantic framework.

This fragmentation results in data silos that hinder integration, reuse, and meaningful analytics. For instance, when agronomists or technology providers attempt to combine aerial imagery with in-soil sensor data, or align robotic observations with external weather inputs, they face tedious manual harmonisation processes. Without a common data foundation, these efforts are error-prone and labour-intensive, ultimately undermining the value of digital tools in smart farming.

To address this issue, the research project SMARAGD has adopted a data-centric strategy centred on semantic interoperability. The project's architecture is built on the open-source

FIWARE ecosystem[2] [1], enabling the creation of so-called Smart Data Spaces where information from heterogeneous systems can be fused, queried, and analysed in a unified way.

The cornerstone of this approach is not only the technical integration of data flows, but also the harmonisation of data semantics using shared, standardised data models. Specifically, the NGSI-LD specification and domain-specific SDMs are used to semantically enrich the data so that it can be understood and reused across platforms. FIWARE's modular components, such as the Context Broker and its standard APIs, enable interoperability across data sources and stakeholders, laying the foundation for collaborative, data-driven agriculture.

B. Real-World Interoperability Challenges

Despite the availability of digital infrastructure and advanced sensing technologies, data interoperability remains a major obstacle in real-world agritech applications. The following examples illustrate the nature and diversity of these challenges across three representative domains in SMARAGD.

Aerial imagery. High-resolution images captured by UAVs provide valuable spatial insights for crop monitoring and yield estimation. However, these images are typically offloaded manually from SD cards after field missions, and their embedded metadata (stored in formats like EXIF or proprietary DJI tags) lacks semantic structure. Different camera types (RGB, thermal, multispectral) introduce further variability. As a result, integrating UAV imagery into downstream systems requires specialised extraction and reformatting steps, often with little reuse potential beyond the initial use case.

Robotic imagery. Mobile agricultural robots are increasingly used for close-range inspection, weeding, and precision spraying. These platforms often rely on the Robot Operating System (ROS), which records multi-sensor data into so-called bag files. While ROS is highly effective within robotic environments, it is not designed for external data interoperability. Each data stream (e.g., from GPS, IMU, or cameras) follows a different structure, and the data is typically not enriched with semantic or geospatial context. Converting this raw content into usable, interoperable formats requires extensive processing and domain knowledge.

Environmental sensors. IoT systems deployed in greenhouses, tunnels, and open fields continuously monitor soil, climate, and air parameters. However, these devices are typically vendor-specific, each exposing data through bespoke APIs, measurement units, and timestamp conventions. The absence of shared ontologies or data models results in a proliferation of disconnected systems. For example, two moisture sensors from different providers may use different naming conventions, measurement intervals, and units, hindering aggregation and comparative analysis.

These examples underscore the need for structured, semantically aligned representations that facilitate consistent integration across space, time, and source. By adopting NGSI-LD-based SDMs, the SMARAGD project enables the transformation of heterogeneous sensor outputs into harmonised entities, thereby making them accessible for advanced analytics, visualisation, and long-term reuse.

[2]https://www.fiware.org/

III. Related Work

Interoperability remains a key challenge for smart agriculture, where heterogeneous systems often lack shared standards [2]. Existing modelling vocabularies for data representation and exchange [3] are often either too generic or tailored to narrow, non-reusable use cases. Similarly, ontologies for IoT, sensor networks, and cyber-physical systems [4], [5] need further refinement to address the specific needs of the agritech domain. Although various competing agritech data standards exist,[3] none has achieved wide adoption.

Equally important is the design of data pipelines – automated workflows that perform extraction, transformation, and loading (ETL) of data into usable formats [6]. In agriculture, traditional pipelines often operate in silos, processing drone imagery, sensor streams, or robotic data separately. More recent efforts promote unified, context-aware pipelines based on open standards and semantic models. Initiatives such as FIWARE[4] and NGSI-LD SDMs provide standardised representations to enable cross-platform integration [7], [8], [9], [10]. Another influential reference architecture is Microsoft's FarmBeats [11], [12], [13], which combines edge computing, cloud storage, and AI to process large-scale agritech data. Many pipeline architectures adopt containerised microservices [14] and event-driven components for scalability and flexibility. However, challenges persist, including the harmonisation of proprietary formats, temporal-spatial alignment of multimodal data, and generalisability beyond isolated deployments.

IV. Semantic Interoperability Approach

While numerous promising solutions exist, much of the current work remains experimental or domain-specific. This paper builds on best practices from these efforts, advancing semantic interoperability through SDMs and open, standards-based tooling. Achieving semantic interoperability across diverse agritech data sources is essential for enabling data sharing, fusion, and value-added analytics. This section outlines the key enabling technologies and modelling principles that underpin our approach, developed in the SMARAGD project.

A. SDMs and NGSI-LD for Interoperable Agritech Data

SDMs[5] are at the core of the FIWARE ecosystem, designed to standardise the structure and semantics of data exchanged across systems. They are based on the NGSI-LD specification [15], which represents context data as entities composed of properties and relationships, enriched with geospatial and temporal context. NGSI-LD also supports linked data principles, facilitating richer semantics and integration across domains.

The SDM initiative maintains a large and evolving repository of domain-specific models. In agriculture, these include

[3]https://www.aspexit.com/standards-and-data-exchange-in-agriculture/
[4]https://www.fiware.org/community/smart-agrifood/
[5]https://smartdatamodels.org/

templates for weather, soil, crop, and sensor data. Reusing existing models accelerates development and fosters interoperability. Where necessary, custom models can be contributed to fill domain gaps, as in our work with drone/robotic imagery.

NGSI-LD enables interoperable representation of core attributes such as geolocation (GeoJSON), time (ISO 8601), and sensor metadata. Its machine-readable and extensible structure allows for integration with third-party data sources, including weather and satellite data. Together, NGSI-LD and SDMs provide a robust foundation for harmonising data across heterogeneous agritech systems.

B. Modelling Strategy for Data Harmonisation

Our modelling approach targets the integration of siloed datasets, such as drone imagery, robotic data, and IoT sensor streams, into a unified data space based on NGSI-LD. All data points are modelled as NGSI-LD entities with standardised attributes, timestamps, and geospatial references, supporting spatio-temporal alignment across sources [16], [17]. In particular, GeoJSON-based GPS coordinates and ISO 8601 timestamps serve as key enablers of data interoperability in precision agriculture, allowing observations from different systems to be fused and interpreted in context. Our approach is guided by three core principles:

- **Heterogeneity awareness**: We account for the fragmented nature of agritech systems and the diversity of their data formats.
- **Reuse of open standards**: Wherever possible, we leverage the FIWARE stack, NGSI-LD format, and existing SDMs to avoid redundant development.
- **Future-proof interoperability**: All modelling aligns with EU data space principles and emerging best practices to ensure reusability and policy compliance.

This modelling layer forms the semantic backbone of our architecture, enabling integration with third-party datasets (e.g., weather forecasts or satellite data) and supporting scalable, interoperable decision support services. The resulting harmonised data space serves as the foundation for downstream decision support services, ranging from visual analytics to predictive models.

1) Aerial Imagery Metadata Transformation: Aerial imagery captured by UAVs is a valuable data source in precision agriculture, often used for monitoring crop health or planning interventions. These images, typically stored in the raw DNG format, include embedded EXIF metadata describing camera properties, geolocation, and capture time. However, this metadata is not natively interoperable.

Several challenges emerge when using this data directly: (1) metadata is fragmented across namespaces like `EXIF`, `XMP`, and `Composite`; (2) fields such as `EXIF:GPSLatitude` or `EXIF:FocalLength` lack semantic context; and (3) proprietary tags from manufacturers like DJI hinder automated processing. While rich in content, such raw metadata is difficult to integrate across systems.

To address this, we developed a semantically enriched representation using a custom NGSI-LD SDM, `DroneImage`, which standardises key attributes such as GPS coordinates in GeoJSON, capture date, and camera settings. GPS coordinates and timestamps serve as primary anchors for interoperability, enabling spatial and temporal alignment with data from other sources. The transformation maps each image into an NGSI-LD entity with human-readable, structured attributes, suitable for integration into a shared data space. An example transformation is illustrated in Table I, comparing raw metadata fields to their harmonised NGSI-LD representation.

2) Robotic Imagery Metadata Transformation: Agricultural robots perform close-up imaging and sensor-based monitoring during tasks such as spraying, weeding, or harvesting. These systems typically use ROS to synchronise multiple sensor streams, including imagery, GPS, and orientation data, into ROS bag files. Although ROS provides a robust internal structure, its data format is not directly interoperable with external systems. Three main challenges limit reusability: (1) sensor data is fragmented across asynchronous topics, (2) message types are diverse and lack semantic annotations, and (3) timestamps and geolocation are scattered and inconsistently structured. Manual extraction and interpretation are therefore required to fuse imagery with context metadata such as position and orientation.

To address this, we developed a custom NGSI-LD SDM called `RoboticFrame`, which unifies data from ROS bag files into semantically enriched entities. The model aligns image frames with spatial, temporal, and camera-specific attributes, using ISO 8601 timestamps and GeoJSON-formatted GPS coordinates as primary anchors for interoperability. An example transformation is shown in Table II, where fragmented ROS data is merged into a single structured and linked entity.

3) Environmental Sensor Data Harmonisation: Environmental sensors deployed in fields, tunnels, or greenhouses generate continuous readings of soil moisture, temperature, and other parameters. These are often transmitted in fragmented batches via NB-IoT gateways, with metadata scattered across configuration files or encoded in vendor-specific formats. The lack of semantic structure, inconsistent timestamp formats, and loosely coupled geolocation data make it difficult to integrate such measurements into a unified system.

To address these challenges, we adopt the existing SDM `DeviceMeasurement`[6] to encode individual readings as semantically structured NGSI-LD entities. This model enables consistent representation of core attributes such as measurement type, value, timestamp, elevation, and device identity. Where applicable, we enrich these entities using additional SDMs from the `SmartAgrifood` domain. As with the other two domains, GPS coordinates (GeoJSON) and ISO 8601 timestamps are used as primary anchors for geospatial and temporal alignment across sources – key enablers for precision agriculture and longitudinal analysis. Each entity also establishes NGSI-LD relationships to its originating device, location, and provider. Table III illustrates a representative example of this transformation.

[6]https://github.com/smart-data-models/Smart-Sensoring

TABLE I
SAMPLE TRANSFORMATION OF THE AERIAL IMAGERY METADATA.

Original format: EXIF metadata


```
{
  "SourceFile": "DJI_0007.DNG",
  "File:FileName": "DJI_0007.DNG",
  "File:FileSize": 40920003,
  "EXIF:Make": "DJI",
  "EXIF:Model": "FC6540",
  "EXIF:ModifyDate": "2024:09:25 17:33:10",
  "EXIF:ImageWidth": 6016,
  "EXIF:ImageHeight": 3376,
  "EXIF:GPSLatitude": -48.875485,
  "EXIF:GPSLongitude": -123.391860,
  "EXIF:GPSAltitude": 189.172,
  "EXIF:LensModel": "DJI DL 24mm F2.8 LS ASPH",
  "EXIF:FocalLength": 24,
  "EXIF:ApertureValue": 2.8,
  "EXIF:ShutterSpeedValue": 0.002,
  "EXIF:ISO": 200,
  "XMP:AbsoluteAltitude": "+189.17",
  "XMP:RelativeAltitude": "+122.01",
  "XMP:GimbalYawDegree": "+90.30",
  "XMP:GimbalPitchDegree": -45.0,
  "Composite:GPSPosition": "-48.875485 -123.391860",
  "Composite:FOV": 53.13
}
```


NGSI-LD format


```
{
  "id": "urn:ngsi-ld:DroneImage:DJI_0007",
  "type": "DroneImage",
  "fileName": {
    "type": "Property",
    "value": "DJI_0007.DNG"
  },
  "fileSize": {
    "type": "Property",
    "value": 40920003
  },
  "manufacturer": {
    "type": "Property",
    "value": "DJI"
  },
  "model": {
    "type": "Property",
    "value": "FC6540"
  },
  "imageDimensions": {
    "type": "Property",
    "value": {
      "width": 6016,
      "height": 3376
    }
  },
  "captureDate": {
    "type": "Property",
    "value": "2024-09-25T17:33:10Z"
  },
  "geoLocation": {
    "type": "GeoProperty",
    "value": {
      "type": "Point",
      "coordinates": [-48.875485, -123.391860]
    }
  },
  "altitude": {
    "type": "Property",
    "value": 189.172
  },
  "lensModel": {
    "type": "Property",
    "value": "DJI DL 24mm F2.8 LS ASPH"
  },
  "focalLength": {
    "type": "Property",
    "value": 24
  },
  "aperture": {
    "type": "Property",
    "value": 2.8
  },
  "shutterSpeed": {
    "type": "Property",
    "value": 0.002
  },
  "iso": {
    "type": "Property",
    "value": 200
  },
  "fieldOfView": {
    "type": "Property",
    "value": 53.13
  },
  "gimbalOrientation": {
    "type": "Property",
    "value": {
      "yaw": 90.3,
      "pitch": -45.0
    }
  },
  "imageUrl": "..." // Truncated for brevity
}
```


TABLE II
SAMPLE TRANSFORMATION OF THE ROS BAG METADATA.

Original format: ROS bag


```
{
  "timestamp": "2025-01-21T12:00:00Z",
  "px4/position_world": {
    "Point": {
      "X": 47.11528015136719,
      "Y": 8.670736312866211,
      "Z": 1040.333984375
    }
  },
  "px4/orientation": {
    "Quaternion": {
      "X": -0.04053421614832881,
      "Y": 0.1062795290984786,
      "Z": -0.9375996305227535,
      "W": 0.3285857146308133
    }
  },
  "cam0/image_raw": {
    "sensor_msgs/Image": {
      "Height": 480,
      "Width": 752,
      "Encoding": "mono8",
      "Data": "[87, 92, 93, ...]" // Truncated for brevity
    }
  }
}
```


NGSI-LD format


```
{
  "id": "urn:ngsi-ld:RoboticFrame: 2025-01-21T12:00:00Z",
  "type": "RoboticFrame",
  "timestamp": {
    "type": "Property",
    "value": "2025-01-21T12:00:00Z"
  },
  "position": {
    "type": "GeoProperty",
    "value": {
      "type": "Point",
      "coordinates": [-48.875485, -123.391860],
      "altitude": 1040.333984375
    }
  },
  "orientation": {
    "type": "Property",
    "value": {
      "quaternion": {
        "x": -0.04053421614832881,
        "y": 0.1062795290984786,
        "z": -0.9375996305227535,
        "w": 0.3285857146308133
      }
    }
  },
  "camera": {
    "type": "Property",
    "value": {
      "id": "cam0",
      "resolution": {
        "height": 480,
        "width": 752
      },
      "encoding": "mono8",
      "imageUrl": "..." // Truncated for brevity
    }
  }
}
```


V. DISCUSSION: BENEFITS AND IMPLICATIONS OF INTEROPERABLE AGRITECH DATA

Based on the described data harmonisation strategy, in the context of the SMARAGD project, we developed data transformation pipelines that enabled the integration of drone imagery, robotic observations, and environmental sensor measurements into a unified, standards-based information space. Each dataset is harmonised using SDMs and encoded in NGSI-LD format, allowing them to be jointly queried, visualised, or analysed. Despite differences in data origin, structure, and operational constraints, all entities share a common representation of key attributes such as geolocation (via GeoJSON) and timestamps (ISO 8601). This alignment enables integration across domains and supports spatial-temporal data fusion. NGSI-LD's linked data structure provides machine-readable context by connecting each observation to its corresponding device, location, and provider. Although NGSI-LD representations are more verbose than raw formats and require additional modelling effort, they bring long-term benefits in maintainability, cross-platform compatibility, and ease of reuse, effectively mitigating future technical debt.

As a result, this semantically enriched and standardised representation supports a wide range of decision support use

TABLE III
SAMPLE TRANSFORMATION OF SOIL SENSOR DATA.

Original format: vendor specific JSON documents


```
{
  "_id": {
    "$oid": "634dc15fe4132c6a5cff7372"
  },
  "time": {
    "$date": "2024-07-06T18:41:18.531Z"
  },
  "value": 21,
  "name": 20,
  "elevation": 0,
  "companyId": {
    "$oid": "634db5fd036971b6bf5aba89"
  },
  "location": {
    "_id": {
      "$oid": "634db600036971b6bf5abaa1"
    }
  },
  "device": {
    "_id": {
      "$oid": "634db601036971b6bf5abab7"
    },
    "vendor": 2,
    "type": "hydra",
    "soilConfiguration": {
      "soilType": "organic",
      "depth": 10,
      "soilDryBulkDensity": 0.5,
      "fieldCapacity": 60,
      "wiltingPoint": 15,
      "irrigationThreshold": 0.5
    },
    "serialNumber": "19344",
    "externalIdentity": "11003"
  }
}
```


NGSI-LD format


```
{
  "id": "urn:ngsi-ld:DeviceMeasurement: 634dc15fe4132c6a5cff7372",
  "type": "DeviceMeasurement",
  "measurementType": {
    "type": "Property",
    "value": "soilTemperature"
  },
  "measurementValue": {
    "type": "Property",
    "value": 21
  },
  "elevation": {
    "type": "Property",
    "value": 0
  },
  "timestamp": {
    "type": "Property",
    "value": "2024-07-06T18:41:18.531Z"
  },
  "refCompany": {
    "type": "Relationship",
    "object": "urn:ngsi-ld:Company: 634db5fd036971b6bf5aba89"
  },
  "refLocation": {
    "type": "Relationship",
    "object": "urn:ngsi-ld:Location: 634db600036971b6bf5abaa1"
  },
  "refDevice": {
    "type": "Relationship",
    "object": "urn:ngsi-ld:Device: 634db601036971b6bf5abab7"
  },
  "deviceDetails": {
    "type": "Property",
    "value": {
      "vendor": 2,
      "type": "hydra",
      "soilConfiguration": {
        "soilType": "organic",
        "depth": 10,
        "soilDryBulkDensity": 0.5,
        "fieldCapacity": 60,
        "wiltingPoint": 15,
        "irrigationThreshold": 0.5
      },
      "serialNumber": "19344",
      "externalIdentity": "11003"
    }
  }
}
```


cases. For instance, drone-based vegetation indices can be correlated with robotic imagery and soil moisture readings to inform irrigation or treatment decisions. Integration with weather forecasts further enables optimal scheduling of agricultural operations. Even non-expert users can benefit from basic time-series visualisations, while domain specialists can perform detailed queries and analytics.

Beyond technical integration, this approach offers value for both end users and data providers. Farmers and agronomists benefit from timely, context-aware insights, while data providers can contribute semantically aligned datasets to shared platforms, enabling monetisation, collaboration, and extended reuse. For example, combining ground-based observations with satellite data or treatment recommendations from suppliers creates synergy effects that increase the value of individual datasets. This interoperable infrastructure also opens the door to new business models, such as real-time advisory services, predictive analytics, or autonomous workflows. As interoperability improves, stakeholders across the value chain, from growers to regulators, can engage with trustworthy, machine-readable information that supports sustainable and data-driven agriculture.

Contributions in a European Context. The presented approach aligns with ongoing European efforts to foster data sharing and interoperability in agriculture. Initiatives like AgriDataSpace[7] and policy frameworks such as the EU Data Act [18] emphasise common standards and linked data principles to facilitate innovation and reuse. By adhering to NGSI-LD and contributing domain-specific SDMs, the SMARAGD project lays the groundwork for scalable, semantically interoperable agritech systems. The resulting information assets are reusable in broader European data ecosystems, promoting sustainable agriculture and cross-sector innovation in both national and international contexts.

VI. CONCLUSION

This paper presented a standards-based approach to agritech data interoperability through the application of NGSI-LD and SDMs. Focusing on three heterogeneous data sources – UAV imagery, robotic sensing, and environmental IoT measurements – we demonstrated how semantically enriched and temporally aligned NGSI-LD entities can be used to bring previously siloed datasets into a unified, interoperable information space. This transformation facilitates cross-domain integration, supports spatial-temporal analysis, and lays the groundwork for advanced decision-support services in the Norwegian agrifood sector.

The resulting information assets are designed to be reusable and interoperable across platforms, aligning with the principles and goals of European initiatives such as AgriDataSpace and the EU Data Act. By embracing shared ontologies, linked data principles, and modular, open-source tooling, this work contributes to a scalable, forward-compatible digital infrastructure for sustainable agriculture. In addition to supporting smarter farm-level decision-making, it enables new forms of data collaboration and monetisation, unlocking synergy effects among different data providers and applications.

Future Work. Future work will focus on the broader validation and evolution of the domain-specific SDMs introduced for drone and robotic imagery. While initially developed to meet specific needs in the SMARAGD project, these models will benefit from wider community review and iterative refinement within the open SDMs ecosystem.

[7]https://agridataspace-csa.eu/

TABLE IV
KEY CHALLENGES IN ORIGINAL DATA FORMATS VS. BENEFITS OF NGSI-LD TRANSFORMATION.

Key challenges of the original formats	Key benefits of the transformed NGSI-LD format
• **Scattered and Disjoint Metadata:** Contextual data (e.g., timestamps, geolocation, sensor or device metadata) is spread across multiple namespaces, files, or topics (e.g., `EXIF`, `ROS bags`) and separate sensor metadata files. • **Proprietary and Inconsistent Structures:** Each data source uses its own format (EXIF for UAV imagery, ROS message types for robots) and custom enumerations for IoT sensors, requiring domain-specific logic to interpret or align. • **Limited Interoperability:** Original data structures are tightly coupled to specific vendors or platforms (e.g., DJI cameras, ROS message types, vendor-specific IoT APIs), making cross-system integration difficult. • **No Semantic Clarity or Linked Context:** Field names lack semantic meaning and are not explicitly linked (e.g., sensor measurements not connected to geolocation or device metadata in machine-readable form).	• **Unified Entity Representation:** Each dataset is transformed into a single NGSI-LD entity (e.g., `DroneImage`, `RoboticFrame`, `DeviceMeasurement`) representing relevant context in one structure. • **Semantic Interoperability and Enrichment:** All properties and relationships are semantically defined using SDMs (e.g., `measurementType`, `deviceDetails`, `captureDate`, `geoLocation`). • **Spatial and Temporal Alignment:** Geo-coordinates are consistently represented using GeoJSON; timestamps are standardised to ISO 8601 across all data types. • **Cross-Platform Integration:** NGSI-LD enables seamless ingestion into context-aware systems (e.g., FIWARE's Orion-LD), supporting integration with third-party data sources like satellite imagery or weather forecasts. • **Machine-Readable Relationships:** Explicit links are established between entities (e.g., sensor → device → location → company), enabling scalable querying and analytics.

Moreover, we plan to quantitatively evaluate the performance and scalability of the underlying data transformation pipelines under realistic, large-scale workloads [19]. This includes stress-testing the ingestion and processing capabilities of the system across different deployment environments (edge, fog, cloud) and assessing its responsiveness and throughput when handling high-frequency sensor streams or large image collections. Such benchmarks will inform optimisation strategies and ensure that the system remains robust, efficient, and suitable for operational use in demanding agritech scenarios.

ACKNOWLEDGMENT

This work is funded by the Research Council of Norway under grant agreement no. 337012 (SMARAGD) and the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101135576 (INTEND).

REFERENCES

[1] R. Dautov, S. Tverdal, A. S. Bondevik, S. A. Frøshaug, V. Szabo, J. R. Fiksdal, and M. F. Stølen, "SMARAGD: Data Interoperability for Decision Support in the Norwegian Agrifood Sector," in *Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning (RuleML+RR-Companion 2024)*, 2024.

[2] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. N. Hindia, "An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges," *IEEE Internet of things Journal*, vol. 5, no. 5, pp. 3758–3773, 2018. doi: 10.1109/JIOT.2018.2844296

[3] V. Pesce, G.-W. Kayumbi, J. Tennison, L. Mey, and P. Zervas, "Agrifood data standards: a gap exploration report," *F1000Research*, vol. 7, no. 176, p. 176, 2018. doi: 10.7490/f1000research.1115261.1

[4] S. Mishra and S. Jain, "Ontologies as a semantic model in IoT," *International Journal of Computers and Applications*, vol. 42, no. 3, pp. 233–243, 2020. doi: 10.1080/1206212X.2018.1504461

[5] I. Szilagyi and P. Wira, "Ontologies and Semantic Web for the Internet of Things-a survey," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016. doi: 10.1109/IECON.2016.7793744 pp. 6949–6954.

[6] A. Raj, J. Bosch, H. H. Olsson, and T. J. Wang, "Modelling data pipelines," in *2020 46th Euromicro conference on software engineering and advanced applications (SEAA)*. IEEE, 2020. doi: 10.1109/SEAA51224.2020.00014 pp. 13–20.

[7] J. López-Riquelme, N. Pavón-Pulido, H. Navarro-Hellín, F. Soto-Valles, and R. Torres-Sánchez, "A software architecture based on FIWARE cloud for Precision Agriculture," *Agricultural water management*, vol. 183, pp. 123–135, 2017. doi: 10.1016/j.agwat.2016.10.020

[8] P. Corista, D. Ferreira, J. Gião, J. Sarraipa, and R. J. Gonçalves, "An IoT agriculture system using FIWARE," in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. IEEE, 2018. doi: 10.1109/ICE.2018.8436381 pp. 1–6.

[9] M. A. Rodriguez, L. Cuenca, and A. Ortiz, "FIWARE open source standard platform in smart farming-a review," in *19th Working Conference on Virtual Enterprises (PRO-VE 2018)*. Springer, 2018. doi: 10.1007/978-3-319-99127-6_50 pp. 581–589.

[10] I. Roussaki, K. Doolin, A. Skarmeta, G. Routis, J. A. Lopez-Morales, E. Claffey, M. Mora, and J. A. Martinez, "Building an interoperable space for smart agriculture," *Digital Communications and Networks*, vol. 9, no. 1, pp. 183–193, 2023. doi: 10.1016/j.dcan.2022.02.004

[11] D. Vasisht, Z. Kapetanovic, J. Won, X. Jin, R. Chandra, S. Sinha, A. Kapoor, M. Sudarshan, and S. Stratman, "FarmBeats: an IoT platform for Data-Driven agriculture," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. USENIX Association, 2017, pp. 515–529.

[12] S. Malvar, A. Badam, and R. Chandra, "FarmBeats: Digital Water for Agriculture," *Resource Magazine*, vol. 29, no. 4, pp. 40–42, 2022.

[13] Z. Kapetanovic, D. Vasisht, J. Won, R. Chandra, and M. Kimball, "Experiences deploying an always-on farm network," *GetMobile: Mobile Computing and Communications*, vol. 21, no. 2, pp. 16–21, 2017. doi: 10.1145/3131214.3131220

[14] D. Jaramillo, D. V. Nguyen, and R. Smart, "Leveraging microservices architecture by using Docker technology," in *SoutheastCon 2016*. IEEE, 2016. doi: 10.1109/SECON.2016.7506647 pp. 1–5.

[15] ETSI Industry Specification Group (ISG), "Context Information Management (CIM); NGSI-LD API," ETSI, Tech. Rep. RGS/CIM-009v161, 2019.

[16] R. Dautov and S. Distefano, "Distributed data fusion for the Internet of Things," in *Proceedings of 14th International Conference on Parallel Computing Technologies (PaCT 2017)*. Springer, 2017. doi: 10.1007/978-3-319-62932-2_41 pp. 427–432.

[17] R. Dautov and S. Distefano, "Three-level hierarchical data fusion through the IoT, edge, and cloud computing," in *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 2017. doi: 10.1145/3109761.3158388 pp. 1–5.

[18] European Commission and Directorate-General for Communication, *Data Act – The path to the digital decade*. Publications Office of the European Union, 2022.

[19] D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, L. Balzotti, H. Song, D. Roman, R. Dautov, E. J. Husom, S. Sen, V. Balionyte-Merle, A. Morichetta, S. Dustdar, T. Metsch, V. Frascolla, A. Khalid, G. Landi, J. Brenes, I. Toma, R. Szabó, C. Schaefer, C. Udroiu, A. Ulisses, V. Pietsch, S. Akselsen, A. Munch-Ellingsen, I. Pavlova, H.-G. Kim, C. Kim, B. Allen, S. Kim, and E. Paulson, "INTEND: Intent-Based Data Operation in the Computing Continuum," in *CEUR Workshop Proceedings*, vol. 3692, 2024, pp. 43–50.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 303–308
DOI: 10.15439/2025F7988
ISSN 2300-5963 ACSIS, Vol. 43

Automatic Parcel Damage Recognition Module for an Inspection Robot

Wiktor Goszczyński, Szymon Wałęga, Janusz Chmiel, Bartłomiej Gawęda, Katarzyna Grobler-Dębska
AGH University of Krakow
Dep.of Automatic Control and Robotics
al. Adama Mickiewicza 30
30-059 Kraków, Poland
Email: {wiktorg, swalega, januszchm}@student.agh.edu.pl,{bargaw, grobler}@agh.edu.pl

***Abstract*—This paper presents an AI-based solution for automated parcel damage detection, combining machine learning algorithms with a custom-built mobile inspection robot. A mobile robot was designed and built specifically for this task, equipped with a vision system and sensing components. We developed a dataset of over 6,800 images and applied a tailored data augmentation process to better capture the variability found in operational environments. Our approach refines a YOLOv11n-cls-based model, achieving 98.50% accuracy, $97.\%$ precision, and 99.74% recall on validation data. By optimizing the model for deployment on widely available hardware via CoreML, we reached inference speeds exceeding 251 FPS, ensuring rapid processing. An interactive dashboard was also created to monitor performance and facilitate comparisons between model iterations.**

I. Introduction

MODERN warehouses are under increasing pressure to optimize logistics processes, which has led to a growing interest in automation technologies. Warehouse automation is evolving toward flexible and intelligent systems that leverage autonomous mobile robots (AMRs), artificial intelligence (e.g., for image analysis to detect damaged packages or support inventory, demand forecasting, robot path optimization, and automated quality control), and digital twins to optimize operations in real time. Increasing emphasis is being placed on human-machine integration, solution scalability, and the ability of systems to adapt to dynamic and unpredictable warehouse environments [13].

One promising application of these technologies is automated package inspection based on computer vision, which offers a scalable, consistent, and efficient alternative to manual quality control. Traditional methods for detecting damaged packages rely heavily on human inspection, which is time-consuming, inconsistent, and impractical in high-throughput environments. Automated inspection robots not only address these limitations but also serve as valuable platforms for testing and validating modern AI algorithms, effectively bridging the gap between theoretical research and practical deployment.

This synergy forms the foundation of the project initiated within the Industrial Data Science (IDS) student research group. The main goal of this paper is to design and build a mobile inspection robot equipped with a computer vision system capable of detecting damaged packages in real time. This paper presents the process of constructing such a robot and implementing a computer vision algorithm for package damage detection. The contribution of this work is a functional prototype that demonstrates an AI-based package damage detection system intended for integration with a mobile robotic platform. The developed system achieves high classification accuracy and operates in real time.

The paper is organized as follows. The section I presents the context of warehouse automation and outlines the main objective of the project, which is to develop a parcel damage detection system. The section II reviews current solutions in computer vision, mobile robotics, and defect detection. The section III describes the construction and architecture of the developed mobile inspection platform. The section IV explains the process of data collection, exploration, and augmentation. In section V, the use of the YOLO model ("You Only Look Once"), training configuration, and iterative model improvement are discussed. The section V-B covers the conversion and preparation of the model for real-time inference. The results and evaluation section presents the model's performance metrics, error analysis, and results under challenging test conditions. The section VII outlines an interactive interface for evaluating and comparing model performance. The paper (sec. VIII, sec.IX) concludes with the future work and conclusion, which highlight potential directions for further development and summarize the achieved outcomes.

II. Related Works

Modern warehouses are increasingly adopting autonomous mobile robots (AMRs) to automate logistics processes. Unlike traditional automated guided vehicles (AGVs), AMRs can navigate dynamic environments and make independent decisions without fixed infrastructure [7]. Advances in sensing, control, and AI have enabled their use not only for material handling but also for inspection tasks [16]. A related study describes a cyber-physical warehouse system for order-picking using mobile robots, combining LiDAR for obstacle avoidance and optimized path-planning in dynamic environments [2]

This work was supported by AGH University of Krakow subvention for scientific activity no. 16.16.120.773 and and by program "Excellence initiative—research university" for the AGH University no. 500.696.9400.

Several implementations showcase inspection capabilities. For instance, a MicroAir Vehicle with onboard SLAM and camera was used to scan QR codes in GPS-denied warehouses [9]. More advanced approaches integrate YOLO-based object detection for real-time condition assessment of packages and infrastructure [12]. Other systems combine LiDAR and depth cameras to support autonomous inventory validation [11].

Artificial intelligence plays a key role in these systems. A comprehensive review highlights the impact of AI and computer vision in enabling object recognition, quality control, and adaptive inspection [13]. Techniques like edge computing and reinforcement learning further enhance system responsiveness.

The YOLO (You Only Look Once) framework remains a leading architecture for real-time defect detection. Recent studies trace its development from v1 to v11, noting improvements like attention modules and deployment readiness on edge devices [5], [8]. In industrial settings, YOLO has been successfully applied to detect packaging defects [14], classify damage types with high accuracy [3], and improve connector inspection robustness [15].

Beyond warehouses, AI-powered visual inspection is explored in broader logistics, such as parcel sorting in rail transport systems [4].

The literature emphasizes the growing importance of advanced deep learning models, particularly YOLO, for real-time, accurate, and scalable inspection tasks in warehouse environments and autonomous mobile robots in tasks related to package control and inspection in warehouse environments. The integration of advanced sensors, vision systems, and AI algorithms with mobile robotic platforms lays a solid foundation for further developments toward fully automated warehouse processes and high-quality logistics services.

III. Custom Mobile Inspection Robot

A custom mobile inspection robot was developed as part of a student research initiative to support parcel damage recognition. The platform integrates key hardware components and provides mobility, sensing, and computation capabilities needed for logistics environments.

A. Design Overview

The robot was built with the following core requirements: compact, modular chassis for easy transport; omnidirectional mecanum wheels for maneuverability; and onboard sensors for robust environmental perception. The sensing suite includes an RGB camera, LiDAR, and ultrasonic distance sensors. Battery power ensures portability and extended runtime.

B. System Architecture and Integration

The robot is powered by two primary units: an NVIDIA Jetson Nano and an Arduino microcontroller. The Jetson handles camera input, sensor processing, and ML inference, while the Arduino controls motor operations. The RGB camera is mounted on a rotating servo for dynamic inspection, and LiDAR provides precise obstacle detection. Both units communicate via USB.

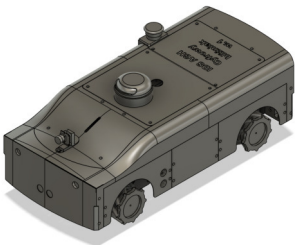


Fig. 1. Conceptual Model of the Inspection Robot

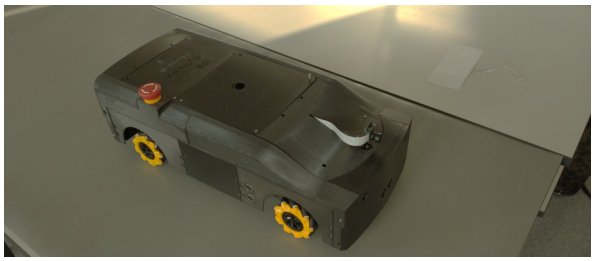


Fig. 2. 3D Printed Prototype of the Inspection Robot

C. Testing and Validation

The mechanical components were prototyped via 3D printing and tested in laboratory conditions. Functional validation confirmed smooth omnidirectional movement, stable sensor integration, and real-time data capture necessary for classification.

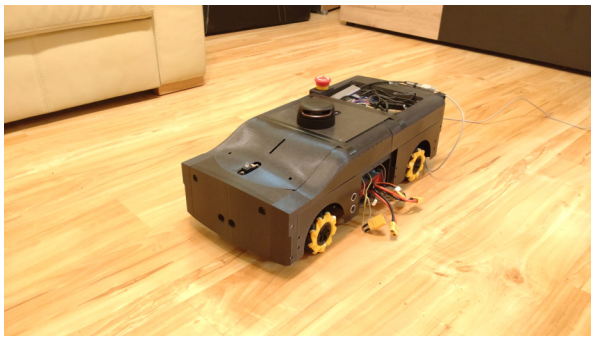


Fig. 3. Robot during LiDAR Testing in Lab Conditions

IV. Dataset Development

The foundation of our parcel damage classification system developed for the inspection robot is a comprehensive, high-quality dataset that accurately represents the problem. This section details our approach to dataset construction, exploration, and augmentation.

A. Data Collection and Organization

Initial data collection began with a publicly available dataset from Kaggle [1] which provided a useful starting point but exhibited several limitations . We used the FiftyOne [10] framework for extensive dataset exploration and analysis. FiftyOne provides valuable tools for:

- Visualizing dataset samples and their distributions
- Identifying potential biases in the data
- Detecting and removing duplicate or near-duplicate images
- Verifying label quality and consistency

Upon inspection using this framework, we found several limitations of gathered photos:

- Poor image quality in numerous samples.
- Limited variety in package types and damage patterns.
- Presence of unrealistic or staged damage scenarios

To address these limitations, we supplemented the initial dataset with custom images captured in realistic environments in different situations and scenes. This approach improved the representativeness of the training data with respect to operational conditions encountered in logistics environments.

This exploration phase was crucial for ensuring dataset quality prior to model training and helped identify specific areas that required additional data collection.

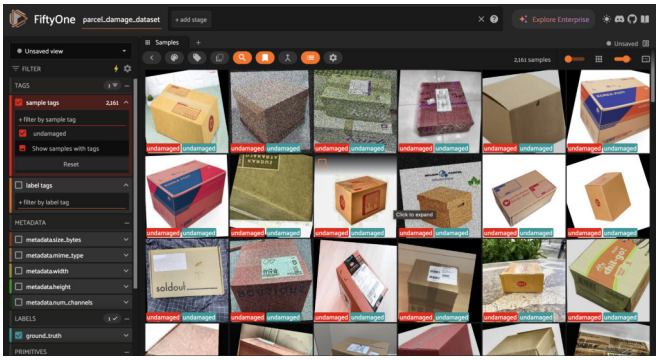


Fig. 4. Shows representative samples from our dataset, highlighting the variety of parcel types, lighting conditions, labeling, and damage patterns.

B. Augmentation Strategy

Data enhancement plays a critical role in improving the robustness of the model. Rather than applying generic augmentation techniques, we developed a domain-specific augmentation pipeline using the Albumentations library. Our pipeline was carefully designed to simulate realistic variations that parcels might encounter in logistics environments while avoiding unrealistic transformations that could harm model performance. We applied:

- Conservative spatial transformations to simulate realistic camera angle variations:
  - Slight rotations (±10°) with a 50% probability
  - Small affine transforms (scale factor 0.95-1.05, translation ±3%)
  - Horizontal flips with a 30% probability
  - No-op spatial transformations with a 20% probability
- To simulate different indoor lighting conditions commonly found in warehouses and sorting facilities:
  - Random brightness/contrast adjustments (±10%) with a 50% probability
  - Color temperature shifts (brightness ±5%, contrast ±5%, saturation ±5%, hue ±5%) with a 30% probability
  - No lighting modifications with a 20% probability
- Shadow Effects To simulate natural lighting conditions:
  - Very mild random shadows with a 20% probability
  - No shadow effect with a 80% probability

The final pipeline included consistent resizing to 640×640 pixels while maintaining aspect ratio, implemented through a combination of LongestMaxSize and PadIfNeeded transformations with constant border padding. This pipeline was applied 10 times to each of the original photos.

The final dataset was organized into a structured format with two primary categories:

- Damaged: 2,372 training images and 903 validation images
- Undamaged: 2,600 training images and 1,000 validation images



Fig. 5. Illustrates examples of our augmentation pipeline applied to sample images, demonstrating the realistic nature of the transformations.

V. Training Classifier

For our parcel damage detection task, we selected the YOLOv11n-cls model [6], a classification variant of the YOLO (You Only Look Once) family. This approach was chosen for several reasons:

- Efficient inference speed, critical for real-time applications in logistics
- Strong feature extraction capabilities from the backbone network
- Successful track record in related computer vision tasks
- Flexibility to be deployed on various hardware platforms

A. Training Methodology

The model was trained using the Ultralytics framework [6], which provides an efficient implementation of YOLO models with several optimizations for training and inference. The training configuration used the following parameters:

- Input image size: 640×640 pixels
- Training epochs: 100
- Model architecture: YOLOv11n-cls.pt

Other hyperparameters such as batch size, optimizer settings, and learning rate schedule were automatically determined by the Ultralytics framework based on dataset characteristics and hardware capabilities. This approach leverages the expertise embedded in the framework to achieve optimal training results without manual hyperparameter tuning reducing manual workload and streamlining the training process.

Our model development followed an iterative improvement process with following steps:

1) Initial training with the training part of created dataset resulted in our baseline model (best0)
2) Performance analysis of best0. That revealed a specific weakness: opened boxes were frequently misclassified as undamaged
3) Additional data collection focused specifically on opened boxes, labeled as damaged
4) Dataset augmentation with the new samples
5) Re-training with the enhanced dataset resulted in our improved model (best1)
6) Performance analysis of best1. That revealed no major weaknesses in classification

This iterative approach allowed us to improve the model performance. The figure 6 presents a visualization of the training process of the YOLOv11-cls classification model, comparing two model variants: Best0 and Best1. Based on the charts, Best1 consistently achieves better results than Best0 across all key metrics—loss, accuracy, and F1 score. Both accuracy and F1 score increase with each epoch, approaching a value of 1.0, while the loss steadily decreases. The convergence point occurs at epoch 72, after which the model performance stabilizes. Throughout most of the training process, Best1 shows lower loss and higher accuracy and F1 score, indicating its superior classification performance compared to Best0.

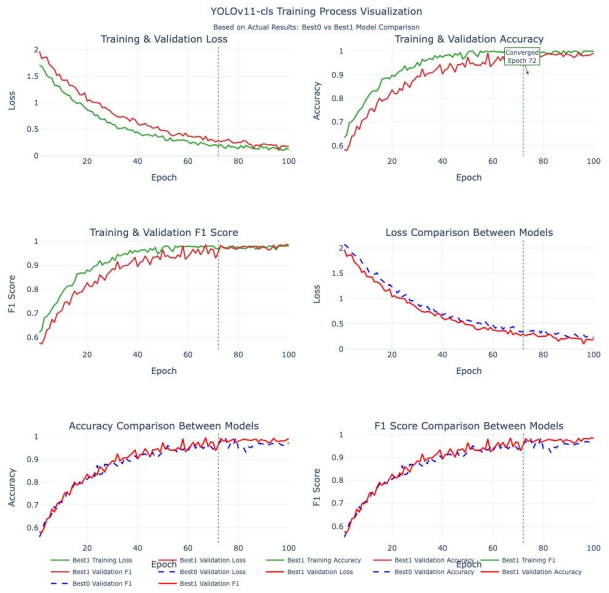


Fig. 6. Comparison of the training process between two YOLOv11-cls models (Best0 vs Best1) based on loss, accuracy, and F1 score for both training and validation datasets. Model Best1 demonstrates better convergence and higher classification performance overall.

B. Model Optimization

During testing with real-time inference the PyTorch model's frames per second (FPS) was around 30, meaning that it can process that number of sample pictures in one second. After converting to CoreML format the model reached 251+ FPS. We converted our models to ONNX format to address two primary objectives: first, to enable robust metric extraction and validation within our performance monitoring dashboard, as the standardized ONNX format provides more predictable parameter access patterns compared to native PyTorch. Secondly, changing model's format allowed us to establish a framework-agnostic model representation that enhances deployment flexibility and ensures consistent inference behavior across different runtime environments. The training process was monitored using metrics that included training loss, validation loss, and top-1 accuracy, which were recorded every epoch.

We achieved this using the following process:

1) Export of the trained PyTorch model to ONNX format
2) Conversion from ONNX to CoreML using the coremltools library
3) Optimization of the CoreML model for the Apple Neural Engine

The high inference speed makes the model suitable for integration into high-throughput logistics environments where real-time processing is critical.

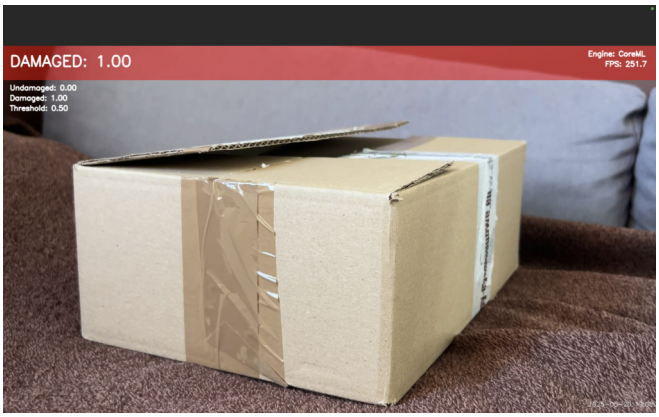


Fig. 7. Optimization process maintained model accuracy while significantly improving inference speed, achieving over 251 FPS on Apple M1 hardware during real-time testing.

VI. Results and Evaluation

A. Performance Metrics

We evaluated our models using classification metrics including accuracy, precision, recall, and F1 score.

The best0 and best1 models achieved results as presented in Table I. The best1 model consistently outperformed the baseline (best0) across all evaluation metrics, indicating successful

mitigation of the identified dataset weaknesses. The enhancements in precision are particularly notable, indicating that our approach successfully reduced false-positive classifications.

TABLE I
SHOWS THE TRAINING STATISTICS FOR BOTH MODELS, ILLUSTRATING THE IMPROVEMENT IN FINAL PERFORMANCE ACHIEVED BY THE BEST1 MODEL AFTER ADDRESSING THE IDENTIFIED WEAKNESSES IN THE DATASET.

Metric	best0 model	best1 model	Improvement
Accuracy	97.00%	98.50%	+1.50pp
Precision	94.80%	97.04%	+2.24pp
Recall	98.81%	99.74%	+0.93pp
F1 Score	96.76%	98.37%	+1.61pp

We also evaluated our models under real-life conditions, such as too dark or too light environment. These results suggest that our models performance can be increased by hyper-tuning or changing the augmentation parameters for images. Scores achieved by both models in this scenario are presented in Table II.

TABLE II
SHOWS STATISTICS FOR BOTH MODELS IN BOTH DARKER AND LIGHTER ENVIROMENTS USING REAL-LIFE IMAGES AND AUGMENTATIONS

Metric	best0 model	best1 model	Improvement
Accuracy	92.06%	94.44%	+2.38pp
Precision	89.74%	91.14%	+1.4pp
Recall	97.22%	100%	+2.78pp
F1 Score	93.33%	95.36%	+2.03pp

B. Confusion Matrix and Precision-Recall analysis

Detailed analysis of these matrices reveals:

- True Positives (Damaged correctly classified): Increased from 747 to 754
- False Negatives (Damaged classified as Undamaged): Decreased from 9 to 2, representing a 77.8% reduction
- True Negatives (Undamaged correctly classified): Increased from 871 to 889
- False Positives (Undamaged classified as Damaged): Decreased from 41 to 23, representing a 43.9% reduction

The substantial reduction in false negatives is particularly important for our application context. In parcel damage detection, false negatives (missed damages) typically have a greater business impact than false positives, as they can lead to damaged items being delivered to customers, resulting in customer dissatisfaction and potential financial losses.

C. Visual Inspection of Classification Results

Beyond quantitative metrics, we conducted a qualitative analysis of classification results to identify patterns in the remaining misclassifications. For the best1 model, we observed the following:

- False negatives (2 instances): Both cases involved parcels with minor damage that was partially obscured or at the edge of the package.

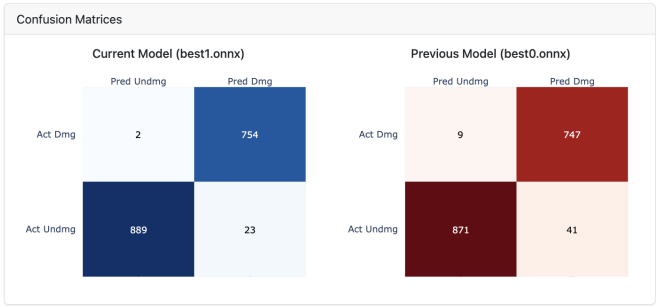


Fig. 8. Confusion matrices for both models

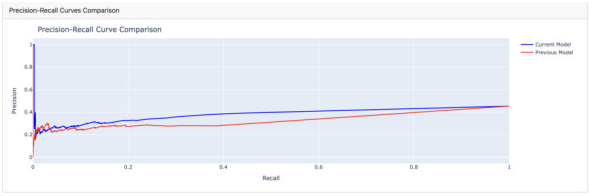


Fig. 9. Shows the precision-recall curves, where the best1 model (blue line) maintains higher precision across all recall values than best0 (red line). These analyses provide a more comprehensive view of model performance beyond the single operating point represented by the confusion matrix, demonstrating that the best1 model achieves superior performance across various classification thresholds.

- False positives (23 instances): Primarily involved parcels with heavy tape reinforcement or unusual packaging materials that resembled damage patterns.

These observations provide valuable insights for future improvements, suggesting that focusing on capturing more examples of minor/obscured damage and diverse packaging materials could further enhance model performance. Fig. 8

VII. Dashboard for Visualization and Analysis

To support model evaluation and facilitate comparisons, we developed a lightweight web-based dashboard using Dash and Plotly. The interface allows users to select between model versions (e.g., best0 vs best1), inspect performance metrics (accuracy, precision, recall, F1 score), and visualize confusion matrices and precision-recall curves. Users can also upload custom images or specify test folders for evaluation.

The dashboard supports real-time testing with side-by-side model predictions, along with confidence scores displayed via intuitive visual indicators. All visualizations are interactive, supporting threshold-independent assessments and flexible data exploration.

This interface enables rapid prototyping, debugging, and demonstration of classification results without requiring deep technical expertise. The tool has proven useful for both offline analysis and on-site testing scenarios.

VIII. Future Works

Although the developed parcel damage classification system achieves high performance, several promising directions remain for further development, both in terms of visual analysis

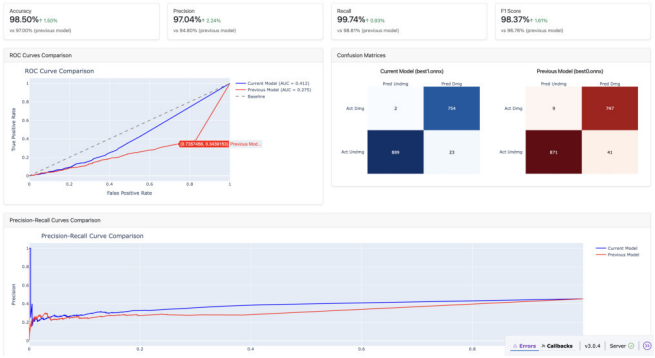


Fig. 10. Comprehensive dashboard presenting all the important information about currently selected model. It also give option for real-time image testing.

and integration with the mobile inspection platform. The key next steps include:

1) Damage Localization: Extending the system to not only classify parcels as damaged but also localize and highlight specific damaged areas.
2) Damage Type Classification: Developing a multi-class model that can distinguish between different types of damage (e.g., tears, water damage, crushing) to provide more detailed information.
3) Integration with Mobile Robot: Combining the visual inspection system with the custom-built mobile platform will enable fully autonomous parcel inspection within warehouse environments.
4) Edge Deployment: Further optimization for deployment on edge devices directly on conveyor systems using techniques like model quantization.
5) Multi-View Integration: Combining predictions from multiple camera angles to improve detection accuracy for damages that might only be visible from certain perspectives.
6) Continuous Learning Framework: Implementing a system for ongoing model improvement based on operational feedback and new data collection.

IX. Conclusion

This paper presents a complete parcel damage detection system based on machine learning, covering data preparation, dataset enrichment, model development, and real-time optimization. A key component of the solution is the custom-designed and built mobile inspection robot, tailored for use in warehouse environments.

The system achieved a very high recall rate (99.74%), demonstrating its ability to detect parcel damage reliably, an essential requirement in logistics to minimize customer complaints and reduce financial losses. The inference speed supports deployment in high-throughput environments, and the custom dashboard enables real-time performance monitoring and model comparison.

The integration of both components, the mobile robotic platform and the vision-based classification system, lays the foundation for a fully autonomous, real-time parcel inspection solution. This project demonstrates that the thoughtful application of computer vision techniques, combined with domain-specific considerations and high-quality data, can yield practical and effective solutions to real-world challenges in logistics automation.

References

[1] Damaged and intact packages. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm, accessed: 2025-04-23
[2] Alhas, O.B., et al.: Towards developing a cyber-physical warehouse system for automated order-picking. In: Annals of Computer Science and Information Systems, Volume 32. pp. 321–328 (2022). https://doi.org/10.15439/2022F208
[3] Chaudhary, K., Singh, A.: Parcel damage classification using computer vision: A deep learning approach for shipment quality assessment. In: 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). vol. 2, pp. 1–6 (2024). https://doi.org/10.1109/IATMSI60426.2024.10503363
[4] Gupta, M., Garg, N., Garg, J., Gupta, V., Gautam, D.: Designing an intelligent parcel management system using iot & machine learning. In: 2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET). pp. 751–756 (2022). https://doi.org/10.1109/GlobConET53749.2022.9872449
[5] Hussain, M.: Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. Machines **11**(7) (2023). https://doi.org/10.3390/machines11070677, https://www.mdpi.com/2075-1702/11/7/677
[6] Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), https://github.com/ultralytics/ultralytics
[7] Keith, R., La, H.M.: Review of autonomous mobile robots for the warehouse environment (2024), https://arxiv.org/abs/2406.08333
[8] Mao, M., Hong, M.: Yolo object detection for real-time fabric defect inspection in the textile industry: A review of yolov1 to yolov11. Sensors (Basel) **25**(7), 2270 (Apr 2025). https://doi.org/10.3390/s25072270
[9] Martinez-Carranza, J., Rojas-Perez, L.O.: Warehouse inspection with an autonomous micro air vehicle. Unmanned Systems **10**(04), 329–342 (2022). https://doi.org/10.1142/S2301385022410011
[10] Moore, B.E., Corso, J.J.: Fiftyone. GitHub. Note: https://github.com/voxel51/fiftyone (2020)
[11] Parikh, H., Saijwal, I., Panchal, N., Sharma, A.: Autonomous mobile robot for inventory management in retail industry. In: Singh, P.K., Wierzchoń, S.T., Chhabra, J.K., Tanwar, S. (eds.) Futuristic Trends in Networks and Computing Technologies. pp. 93–103. Springer Nature Singapore, Singapore (2022)
[12] Sanchez-Cubillo, J., Del Ser, J., Martin, J.L.: Toward fully automated inspection of critical assets supported by autonomous mobile robots, vision sensors, and artificial intelligence. Sensors **24**(12) (2024). https://doi.org/10.3390/s24123721, https://www.mdpi.com/1424-8220/24/12/3721
[13] Sodiya, E.O., Umoga, U.J., Amoo, O.O., Atadoga, A.: Ai-driven warehouse automation: A comprehensive review of systems. GSC Advanced Research and Reviews **18**(2), 272–282 (2024). https://doi.org/10.30574/gscarr.2024.18.2.0063
[14] Vu, T.T.H., Pham, D.L., Chang, T.W.: A yolo-based real-time packaging defect detection system. Procedia Computer Science **217**, 886–894 (2023). https://doi.org/10.1016/j.procs.2022.12.285, 4th International Conference on Industry 4.0 and Smart Manufacturing
[15] Wu, W., Li, Q.: Machine vision inspection of electrical connectors based on improved yolo v3. IEEE Access **8**, 166184–166196 (2020). https://doi.org/10.1109/ACCESS.2020.3022405
[16] Xuan, O.W., Selamat, H., Muslim, M.T.: Autonomous mobile robot for transporting goods in warehouse and production. In: Tan, A., Zhu, F., Jiang, H., Mostafa, K., Yap, E.H., Chen, L., Olule, L.J.A., Myung, H. (eds.) Advances in Intelligent Manufacturing and Robotics. pp. 555–565. Springer Nature Singapore, Singapore (2024)

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 309–314
DOI: 10.15439/2025F5270
ISSN 2300-5963 ACSIS, Vol. 43

Length-Preserving Hair Simulation via Bézier Curves and Artificial Neural Network

Kali Gurkahraman
0000-0002-0697-125X
Department of Computer Engineering
Sivas Cumhuriyet University
Sivas, Turkey
kgurkahraman@cumhuriyet.edu.tr

Ahmet Firat Yelkuvan
0000-0003-4148-1923
Department of Computer Engineering
Sivas Cumhuriyet University
Sivas, Turkey
aftyelkuvan@cumhuriyet.edu.tr

Rukiye Karakis
0000-0002-1797-3461
Department of Software Engineering
Sivas Cumhuriyet University
Sivas, Turkey
rkarakis@cumhuriyet.edu.tr

***Abstract*—In this study, we present a data-driven method for preserving the length of cubic Bézier curves under external deformations, specifically in the context of hair strand simulation. By sampling millions of original and displaced control point configurations within a 2D bounded grid space, we construct a large dataset of Bézier curve pairs and compute the corresponding t_{final} parameters required to preserve arc length. A lightweight artificial neural network (ANN) architecture is trained on this dataset to predict t_{final} given 16 features representing the coordinates of initial and displaced control points. The model achieves a low mean absolute error (MAE) of 0.00091992 on the test set, ensuring high predictive accuracy. Performance evaluations show that the ANN can predict t_{final} values for up to 200k hair strands in approximately 8 seconds on a standard laptop, aligning with the average number of strands on a human scalp. This approach replaces computationally expensive numerical length-matching operations with a fast, inference-based framework, allowing for efficient simulation of realistic, deformable hair motion while maintaining individual strand lengths.**

***Index Terms*—Hair simulation, hair modelling, Bézier curve, artificial neural network**

I. Introduction

A SINGLE human hair strand, though microscopically thin, represents a significant computational unit in computer graphics due to its complex physical properties. A realistic digital human typically involves simulating tens of thousands of strands—up to 100,000 on average—making hair one of the most resource-intensive components in real-time rendering [1]. In modern video games, virtual reality, and interactive applications, simulating lifelike hair movement under dynamic motion and external forces such as wind is essential but remains a non-trivial challenge. A fundamental and often under-addressed issue in this context is preserving the physical length of each strand, especially during rapid motion or deformation.

Hair simulation traditionally relies on strand-based models, which represent each hair as a discrete chain of mass points connected by springs. These models offer fine-grained control over local deformation such as bending, twisting, and inter-strand collision handling [2,3]. Mass-spring systems, for example, have been widely used due to their simplicity and real-time compatibility. However, ensuring inextensibility—the preservation of strand length—becomes problematic, especially under dynamic motion, where stiff springs introduce instability [3,4]. To improve numerical robustness, implicit integration schemes have been introduced [4,5], but their computational overhead makes them impractical for real-time environments. More recent advancements, such as the Augmented Mass-Spring (AMS) model, introduce ghost rest-shapes and biphasic coupling to improve simulation stability and strand length preservation during movement and collisions [6]. Additionally, Müller et al. proposed the Dynamic Follow-The-Leader (DFTL) method, which enforces inextensibility through a geometric update rule and achieves real-time performance with a single iteration per frame [7]. Another common artifact in strand-based systems is initial sagging caused by gravitational forces at the start of the simulation. To overcome this, Hsu et al. developed a sag-free initialization framework, which establishes a static equilibrium and neutral torque configuration prior to dynamic simulation. This significantly reduces unwanted deformation and helps preserve the intended hairstyle [8]. Furthermore, Position-Based Dynamics (PBD) has become a foundational strand-based method by directly enforcing positional constraints such as segment length. PBD is widely used in interactive simulations due to its robustness and control [9].

In contrast, volume-based methods treat hair as a continuous medium, modeling global movement and shape using tools such as fluid dynamics, volumetric mass models, or free-form deformations [10,11]. These approaches offer better scalability for large hair volumes and can efficiently capture overall flow and motion. However, they often sacrifice individual strand resolution, which is critical for detailed hairstyles and physical realism. For example, continuum models [10] and lattice-based structures [11] can simulate general hair behavior but lack fidelity in capturing high-frequency deformations or enforcing per-strand constraints like inextensibility. As a result, while volume-based methods are advantageous for broad dynamics, they are less suited for applications requiring high detail, such as cinematic effects or close-up character interactions.

Given the respective limitations of strand- and volume-based approaches, researchers have increasingly turned to hybrid simulation methods. These techniques are designed to leverage the physical fidelity of strand-level models while

incorporating the computational efficiency and stability offered by volume-based or constraint-based frameworks. One such approach is Position-Based Dynamics (PBD), which eliminates the need for solving complex differential equations and enforces constraints directly on particle positions [12]. Shape matching methods [9] and meshless deformation frameworks have also been explored to stabilize motion while avoiding stiffness artifacts common in traditional systems. Another line of hybrid research adopts rigid body inspired models using serial chains, enabling accurate modeling of both local strand behavior and global transformations [13,14]. These techniques integrate strand-level articulation within a larger kinematic or physics-based framework to improve simulation realism without sacrificing computational efficiency.

Recent years saw, the emergence of data-driven techniques has opened new possibilities in hair simulation. Neural network-based methods aim to approximate complex physical behaviors without relying on explicit solvers, making them especially suitable for real-time environments. One such approach, Quaffure [15], introduces a quasi-static neural model trained with a self-supervised loss function to predict plausible strand deformations across a variety of poses and hairstyles. By eliminating the need for simulated training data, Quaffure achieves high temporal stability and runs efficiently on consumer hardware, bridging the gap between physical accuracy and performance.

Complementing this trend, Dr.Hair [16] addresses the challenge of strand-level reconstruction from visual input. Leveraging a differentiable rendering framework, the method reconstructs scalp-connected hair strands directly from multi-view images without pretraining, accurately preserving both strand length and directional flow. These techniques exemplify how learning-based methods can supplement or replace traditional solvers for specific simulation tasks.

Alongside advances in simulation, efficient rendering methods have also evolved to meet the demands of real-time graphics. Hair mesh systems, such as the one introduced by Yuksel et al. [17], generate strand geometry procedurally on the GPU at render time. This drastically reduces memory usage while maintaining geometric fidelity, allowing for the real-time rendering of hundreds of uniquely styled characters. Together, these learning- and rendering-focused approaches highlight a growing trend toward hybrid solutions that integrate physical modeling, data-driven inference, and hardware-aware rendering strategies to advance the realism and scalability of hair simulation systems.

Building on the potential of curve-based modeling and hybrid techniques, recent studies have explored combining Bézier representations with physically inspired solvers and mathematical optimization techniques—such as the use of Budan–Fourier analysis for root bounding in curve behavior prediction [18] and contour tracing algorithms for shape reconstruction in structured objects [19]. A notable example is the work which integrated Bézier curves with the Articulated Body Method (ABM) to simulate dynamic hair strand behavior with real-time performance [20]. Their model treats each strand as a Bézier curve and applies ABM to resolve motion and constraint forces efficiently. These approaches reinforce the growing viability of combining geometric curve modeling with physically inspired or learning-based systems—an idea central to our study, which uses Bézier curves in conjunction with artificial neural networks to predict strand deformation under motion and external forces.

This study specifically focuses on the problem of strand length preservation during hair motion and external interactions, such as wind. We critically analyze existing approaches from the perspective of both physical accuracy and real-time performance, highlighting their respective advantages and limitations. Based on this analysis, we propose directions for enhancing simulation fidelity while maintaining computational feasibility. A shallow artificial neural networks (ANN) can accurately and efficiently predict the target final curve parameter for achieving length-preserving transformations of hair strands represented by Bezier curves. Due to their parametric structure, differentiability, and compact representation, Bezier curves inherently support scalable and generalized deformation modeling. This makes them particularly advantageous for learning-based strand manipulation, even under varying strand complexities. This study was initiated with this hypothesis in mind, aiming to validate the effectiveness of lightweight ANN architectures in performing accurate and efficient strand deformations through Bezier-based geometric encoding.

II. Materials and Methods

In this study, we propose a data-driven approach based on ANN to maintain the length of cubic Bezier curves under external deformations, specifically aimed at simulating individual hair strands affected by physical forces while preserving their natural length. This section details the process of data generation, Bezier curve modeling, and ANN-based prediction methodology.

A. Bézier Curve and Hair Modelling

A cubic Bézier curve is defined by four control points: A, B, C, and D, and is expressed parametrically as in equation (1):

$$P(t) = (1-t)^3 A + 3(1-t)^2 tB + 3(1-t)t^2 C + t^3 D \quad (1)$$

Where $t \in [0, 1]$. As the parameter t varies from 0 to 1, the curve smoothly transitions from A to D, influenced in shape by intermediate control points B and C. This property makes Bézier curves particularly well-suited for modeling flexible and natural curves such as hair strands. For example, let us consider a hair strand modeled by a Bézier curve with the control points at locations (x, y) : $A = (0, 0), B = (0.25, 1.0), C = (0.75, -1.0)$ and $D = (1, 0)$. This configuration produces a slightly wavy shape representative of a natural hair strand. Figure 1 illustrates this cubic Bézier curve constructed from the defined control points.

An important advantage of using Bézier curves in hair simulation is that each point on the curve can be independently

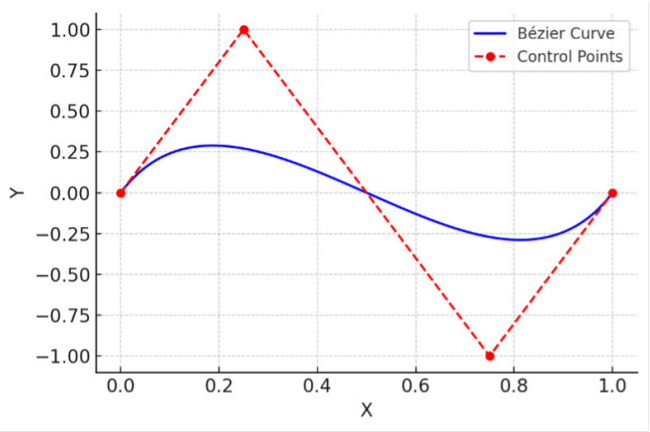


Fig. 1. Bézier Curve for a Wavy Hair Strand

computed by evaluating the Bézier formula at any value of t. For instance, stepping through the interval $[0, 1]$ in increments of 0.1 allows for the generation of discrete, evenly spaced points along the hair strand. This makes the approach inherently parallelizable and computationally efficient for simulating thousands of strands simultaneously.

B. Bézier Curve Representation and Length Estimation for Data Generation

We consider a cubic Bézier curve defined by four control points: A,B,C, and D. Points A and D are fixed at coordinates (0, 0) and (1, 0), respectively, defining the endpoints of a hair strand. Control points B and C are allowed to locate within a region $[0, 1] \times [-1, 1]$ (point space region shown in Figure 1 with 1 unit width and 2 unit height) with a spatial resolution of 0.1. This resolution results in a discrete point cloud of 231 (11x21) possible positions for each of B and C. Although 231 grid points are defined, combining different configurations of initial and moved control points can yield tens of millions of curve pairs. This vast combinatorial space ensures that the dataset captures a wide range of plausible deformations necessary for training a reliable predictive model. However, since many initial and subsequent curves (moved curves) would be undesirably close to each other at a resolution of 0.1, two distinct region location maps were constructed. The first one is a point space with 0.2 resolution at even position values (e.g., 0, 0.2, 0.4), and the second one is a point space with 0.2 resolution at odd position values (e.g., 0.1, 0.3, 0.5). Using each position space separately, initial and subsequent curve locations were determined, and some unsuitable curve pairs were filtered out. Then, to ensure an acceptable data loading time and avoid memory issues, a dataset consisting of 7.5 million samples was constructed through random selection. Each sample consists of 16 input features: the x and y coordinates of the original and moved control points (8 control points with sixteen x and y components), and one output value: the corresponding t_{final} that preserves the curve length. Figure 2 illustrates a unit square point cloud at 0.1 resolution, with an example of original and subsequent cubic Bézier curves generated using a specific configuration of the control points. For each data sample, the corresponding t_{final} values were computed to ensure that the curve length at the subsequent position is equal to that at the initial position.

To simulate external influences, new displaced positions B' and C' are sampled independently from the same discrete space. For each $A-B-C-D$ configuration and its subsequent curve counterpart $A - B' - C' - D$, the curve length of the original and deformed curve is computed. The first curve is parameterized over the interval $t \in [0, 1]$, and the moved curve length is evaluated using numerical integration (e.g., Simpson's rule). Then, the moved curve $A - B' - C' - D$ is sampled over $t \in [0, t_{final}]$ such that its arc length is numerically equal to the original curve.

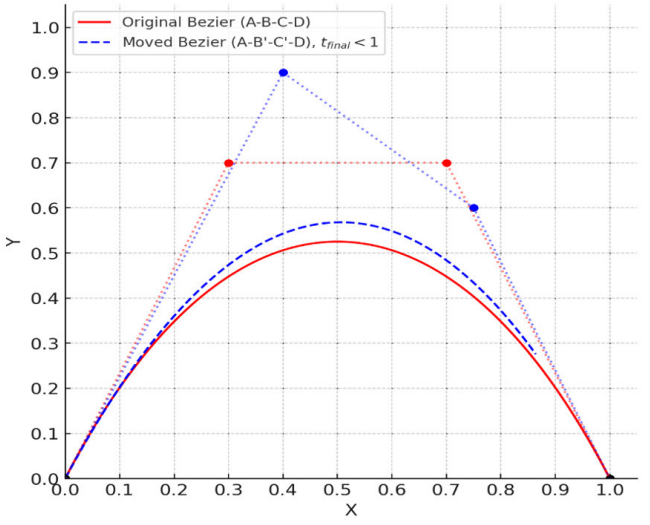


Fig. 2. Example of original and moved Bézier curves

C. Scaling Bézier Curves to Any Rectangular Domains

Although the dataset is generated within a $[0, 1] \times [-1, 1]$ region, Bézier curves can be reliably applied to arbitrarily positioned rectangular regions. This is achieved through a combination of translation and axis-wise scaling that aligns the rectangular configuration with the ANN's reference point space domain. Let a rectangular region be defined by horizontal bounds $[x_{min}, x_{max}]$ and vertical bounds $[y_{min}, y_{max}]$. To align this region with the reference point space, we first translation and then scaling operations as in (2), (3) and (4):

Translation: Shift the lower-left corner of the rectangle to the origin by subtracting (x_{min}, y_{min}) from all control points.

Scaling: Normalize the width and height of the rectangle using the scale factors:

$$s_x = x_{\max} - x_{\min} \tag{2}$$

$$s_y = \frac{y_{\max} - y_{\min}}{2} \tag{3}$$

Each control point $P = (x, y)$ is then transformed to reference region coordinates by:

$$P_{\text{unit}} = \left(\frac{x - x_{\min}}{s_x}, \frac{y - y_{\min}}{s_y} \right) \tag{4}$$

This transformation allows both original and displaced control points to be converted into a form compatible with the ANN trained on reference region samples. Once the ANN predicts the t_{final} value, the Bézier curve can be redrawn directly in the rectangular domain using the actual (moved) control points A, B', C', and D. This method preserves the accuracy of the curve length while adapting the model to diverse geometric contexts.

D. Neural Network Model

To approximate the mapping from control point displacement to the required value, a feed-forward artificial neural network (ANN) was constructed. The ANN model receives a 16-dimensional feature vector as input—comprising the x and y coordinates of original and displaced control points (total of 8 points) —and outputs a single scalar. The network architecture consists of four fully connected layers implemented using the Keras Sequential API:

```
model = Sequential([
    Dense(128, activation='relu',input_shape=(16,)),
    Dense(64, activation='relu'),
    Dense(32, activation='relu'),
    Dense(1, activation='sigmoid')
])
```

This shallow and lightweight model architecture is intentionally selected to enable extremely fast predictions, which is essential when simulating thousands of hair strands in real-time applications. Although the model depth is limited, its predictive power is effectively enhanced by using a very large and diverse training set of more than 7 million Bézier curve configurations.

To train the network, the dataset was split into 80% training and 20% test sets. The target values were normalized using a $MinMaxScaler$. The model was trained with a batch size of 1000 for up to 350 epochs using the Adam optimizer. Early stopping and model checkpointing strategies were used to prevent overfitting and retain the best-performing model. After training, the best model achieved low error on unseen data, with the mean absolute error and sum of absolute errors computed during evaluation. The overall strategy demonstrates that by combining a simple ANN structure with a large and carefully prepared dataset, highly accurate and computationally efficient predictions can be achieved.

This ANN framework enables the rapid estimation of values for potentially thousands of hair strands affected by external factors, allowing for efficient regeneration of hair geometry while maintaining individual strand lengths. The proposed solution thus replaces computationally expensive numerical searches with a lightweight inference mechanism. The network is trained using the mean absolute error (MAE) as the loss function. MAE is a commonly used regression metric that quantifies the average magnitude of the errors between predicted values and the actual ground truth values. It is defined by equation (5).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |t_{\text{final_gt}} - t_{\text{final_pr}}| \tag{5}$$

Where n, t_{final_gt} and t_{final_pr} are number of samples, the ground truth and predicted time values, respectively. MAE is expressed in the same units as the target variable and provides an intuitive measure of model accuracy. It is especially useful when all errors are equally weighted.

III. Results and Discussion

All experiments were conducted on a laptop running Windows 10 Pro, equipped with a 12th Gen Intel® Core™ i7-12700H processor (2.30 GHz, 14 cores, 20 threads), 32 GB of RAM and NVIDIA GeForce RTX 3060 Laptop GPU. The training and testing of the proposed ANN model were implemented using the Keras library (Python 3.9) with TensorFlow. Hair strand movements under external forces were simulated using MATLAB R2021a.

Table 1 gives the training, testing and prediction details of the study. This study used a training dataset of 6 million samples (80% of the full dataset), and 1.5 million samples were reserved for testing. The model achieved a test set MAE of 0.00091992, indicating high predictive accuracy in estimating the parameter for length-preserving curve transformations. In this context, the MAE corresponds to the average deviation in the predicted t_{final} value per hair strand. It does not directly represent a geometric length error but rather a deviation in the Bézier time parameter that determines where the curve should be truncated to preserve length. For example, a MAE of 0.00091992 implies that, on average, the predicted truncation point deviates from the true value by less than one-thousandth of the normalized parameter range $[0, 1]$. Assuming that a single hair strand may undergo deformation multiple times, such a deviation would only accumulate to a 1% error after approximately 5.48 independent displacements for 0.5 of t_{final} true value $(5.48 \times 0.00091992/0.5 \approx 0.01)$. This indicates that even under sustained movement, the predicted t_{final} remains sufficiently accurate to maintain a nearly constant hair length. This also suggests that increasing the amount of training data and using more powerful computational hardware could further reduce the MAE, improving the model's suitability for long-duration hair simulations under dynamic conditions. Table 1 also presents prediction runtime for varying batch sizes. The model predicted values for 1, 1k, 10k, 100k, and 200k samples (batch size) in approximately 0.0579, 0.0994, 0.5296, 4.09, and 8.11 seconds, respectively. The inclusion of 200k samples simulates a full-head hair scenario, as this figure closely matches the average number of hair strands on a human scalp. While the current implementation achieves prediction times of approximately 8 seconds for 200k samples, this duration may be considered borderline for real-time applications. However, this benchmark was obtained on a standard laptop

without GPU acceleration. Since the prediction task involves a shallow ANN with only a few dense layers, it is likely that the inference time could be significantly reduced on optimized hardware or with additional engineering effort. Employing GPU-accelerated inference and incorporating preprocessing strategies such as dimensionality reduction or data batching may potentially improve performance and make the model more suitable for time-sensitive applications. The results in Table 1 pertain to the prediction of t_{final} values rather than complete physical simulation, and thus no direct baseline comparison is available in the literature.

TABLE I
TRAINING, TESTING AND PREDICTION DETAILS OF THE STUDY

Number of train data (0.8 split ratio)		6 milion
Number of test data (0.2 split ratio)		1.5 million
MEA test error		0.00091992
Prediction time (second) [mean (std) for 5 trials]	1 sample	0.05795540 (0.002448)
	1k samples	0.09941540 (0.013011)
	10k samples	0.52964460 (0.039699)
	100k samples	4.08589480 (0.188377)
	200k samples	8.10715880 (0.45811)

Figure 3 illustrates an initial hair strand (blue curve) and its multiple deformed configurations (red curves) resulting from external influences applied to its control points. Despite visual variations in trajectory, all red curves are generated by adjusting the t_{final} parameter through the trained ANN model such that their arc lengths are preserved relative to the initial curve. This figure demonstrates the model's effectiveness in maintaining strand length consistency under dynamic conditions, validating the proposed method for realistic hair motion simulation. Based on the curve analysis illustrated in the figure, the original strand represented by the blue curve has a computed length of 4.803 units. However, the final (sixth) red strand—generated with a cumulative t_{final} deviation of 0.00091992 per step—exhibits a length of 4.876 units. This results in an absolute error of 0.073 units, corresponding to a relative error of approximately 1.52% compared to the original. These findings demonstrate that even minor per-step deviations in t_{final} can accumulate and lead to noticeable differences in physical attributes such as strand length. Consequently, preserving the original length with high fidelity becomes essential, especially in applications requiring physical realism, such as animation and real-time hair simulation.

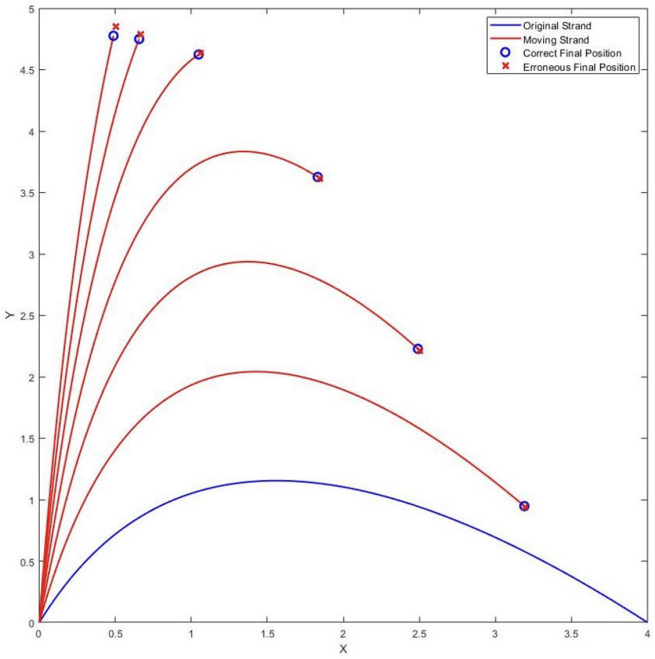


Fig. 3. An initial strand (blue) and its six subsequent movements (red). The cumulative spatial error resulting from an average MAE of 0.00091992 in the predicted t_{final} values is illustrated. Blue circles mark the true terminal positions, while red crosses indicate the erroneous positions due to cumulative prediction error.

IV. CONCLUSION

This work proposes a neural network-based approach to estimate the parameter t_{final} for cubic Bézier curves subjected to external deformation, with the goal of preserving hair strand length in dynamic 2D simulations. By generating a dataset of 7.5 million curve configurations and training a shallow ANN with high generalization capability, the study demonstrates that accurate t_{final} predictions can be achieved rapidly. The resulting MAE, although referring to the time domain rather than direct spatial error, accumulates to only 1% deviation after approximately 5 sequential displacements—suggesting that strand length is effectively preserved over long-term movement. The ANN's prediction times for up to 200k strands—comparable to the number of hairs on a typical human scalp—are under 10 seconds, and can be further reduced with more powerful hardware or model optimization. These findings support the viability of ANN-based inference for realistic and computationally efficient hair simulations.

In future work, this framework will be extended to simulate the dynamic motion of entire hair assemblies composed of Bézier strands under various external forces, where each strand's length is regulated via ANN-predicted t_{final} values. Additionally, further training on a larger dataset is planned to reduce prediction error even more, ensuring higher fidelity in long-duration and high-frequency deformation scenarios. Although this study is currently limited to 2D simulations, the architecture of the dataset generation and ANN model are fully generalizable to 3D Bezier strand representations. Future work will explore 3D vector space adaptation for each control point coordinate.

REFERENCES

[1] R. E. Rosenblum, W. E. Carlson, and E. Tripp, "Simulating the structure and dynamics of human hair," *J. Vis. Comput. Anim.*, vol. 2, no. 4, pp. 141–148, 1991.
[2] A. Selle, M. Lentine, and R. Fedkiw, "A mass spring model for hair simulation," in *ACM SIGGRAPH 2008 Papers*, 2008, pp. 64:1–64:11.
[3] B. Choe, M. G. Choi, and H.-S. Ko, "Simulating complex hair with robust collision handling," in *Proc. Symp. Comput. Anim.*, 2005, pp. 153–160.
[4] R. Bridson, R. Fedkiw, and J. Anderson, "Robust treatment of collisions, contact and friction for cloth animation," in *ACM SIGGRAPH*, 2002, pp. 594–603.

[5] D. Baraff and A. Witkin, "Large steps in cloth simulation," in *ACM SIGGRAPH*, 1998, pp. 43–54.
[6] J. A. Amador Herrera *et al.*, "Augmented mass-spring model for real-time dense hair simulation," *arXiv preprint arXiv:2412.17144v2*, 2024.
[7] M. Müller, T.-Y. Kim, and N. Chentanez, "Fast simulation of inextensible hair and fur," in *Proc. VRIPHYS*, 2012.
[8] J. Hsu *et al.*, "Sag-free initialization for strand-based hybrid hair simulation," *ACM Trans. Graph.*, vol. 42, no. 4, 2023.
[9] M. Müller, B. Heidelberger, M. Hennix, and J. Ratcliff, "Position based dynamics," *J. Vis. Commun. Image Represent.*, vol. 18, no. 2, pp. 109–118, 2007.
[10] S. Hadap and N. Magnenat-Thalmann, "Modeling dynamic hair as a continuum," *Comput. Graph. Forum*, vol. 20, no. 3, pp. 329–338, 2001.
[11] P. Volino and N. Magnenat-Thalmann, "Animating complex hairstyles in real-time," in *Proc. ACM Symp. Virtual Reality Software Technol. (VRST)*, 2004, pp. 41–48.
[12] M. Müller, B. Heidelberger, M. Teschner, and M. Gross, "Meshless deformations based on shape matching," in *ACM SIGGRAPH*, 2005, pp. 471–478.
[13] D.-W. Lee and H.-S. Ko, "Natural hairstyle modeling and animation," *Graph. Models*, vol. 63, no. 2, pp. 67–85, 2001.
[14] R. Featherstone, *Rigid Body Dynamics Algorithms*. Springer, 2007.
[15] T. Stuyck *et al.*, "Quaffure: Real-time quasi-static neural hair simulation," *arXiv preprint arXiv:2412.10061v1*, 2024.
[16] Y. Takimoto, H. Takehara, H. Sato, Z. Zhu, and B. Zheng, "Dr.Hair: Reconstructing scalp-connected hair strands without pre-training via differentiable rendering of line segments," *arXiv preprint arXiv:2403.17496v2*, 2024.
[17] G. Bhokare, E. Montalvo, E. Diaz, and C. Yuksel, "Real-time hair rendering with hair meshes," in *SIGGRAPH Conf. Papers*, 2024.
[18] Q. Zhang, S. Thomas, R. Ramamoorthi, and K. Zhou, "Exploiting Budan-Fourier and Bernstein theorems for real-time collision culling of Bézier curves," in *ACM SIGGRAPH / Eurographics High Perform. Graph. (HPG)*, 2017.
[19] N. Arslan and K. Gurkahraman, "A novel contour tracing algorithm for object shape reconstruction using parametric curves," *Comput. Mater. Continua*, vol. 75, no. 1, pp. 331–350, 2023. doi: 10.32604/cmc.2023.035087
[20] K. Lee and S. Lee, "Hair simulation using articulated body method and Bézier curve," in *SIGGRAPH Asia Tech. Commun.*, 2024.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 315–320
DOI: 10.15439/2025F2657
ISSN 2300-5963 ACSIS, Vol. 43

StarCraft strategy learning refinement using replay snapshotting

Štefan Krištofík
0000-0002-9995-460X
Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
Email: stefan.kristofik@savba.sk

Hanková Michaela
Faculty of Informatics and Information Technologies
Slovak University of Technology
Ilkovičova 2, 842 16 Bratislava, Slovakia

***Abstract*—We propose a new replay snapshotting (RS) technique for strategy learning from past matches in real-time strategy game StarCraft: Brood War (SCBW). It allows for more precise understanding of particular strategy aspects by sampling the state of selected game features at important checkpoints during a match. We use RS to extract and refine a set of strategies from a large replay dataset STARDATA. To validate our approach in a competitive environment, we implement an AI agent for SCBW. It is able to perform the extracted strategy set against opponents in the BASIL Ladder competition. The agent consistently achieves rank C with 56 % win rate which is a significant improvement over our previous approaches.**

I. Introduction

REAL-TIME strategy (RTS) video games, especially the competitive 1v1 setting, are popular in AI research and education because of the inherent difficulties and challenges they present. During a match, only parts of the battlefield near their own structures and units are visible to players [18]. This partial observability of the game state forces players to make decisions under uncertainty based on available incomplete information. Players need to actively scout the map to gather information about the opponent's activities [11]. Moreover, the number of possible actions a player can execute at each time is overwhelming, making the decision process very complex [2].

The most popular and widely known RTS is Blizzard's StarCraft: Brood War (SCBW) released in 1998 (Fig. 1). One aspect which made the game successful and long-lasting is fair balance of 3 playable races: Protoss (P), Terran (T) and Zerg (Z), each offering unique aesthetics and play-styles. Also, no 'cookie cutter' strategy exists which would be able to defeat all opponents [11]. A strategy is a long-term plan for the current match. It typically involves plans on what structures to build in what order and what army composition to aim for. Strategies can be roughly divided into 3 categories capable of countering each other in a rock-paper-scissors manner (Fig. 2). Another aspect is easy accessibility for players and researchers. The game has low hardware requirements, is free-to-play since 2017 and plenty of tools are available for it to aid with

This work was supported by the Slovak Scientific Grand Agency VEGA under the contract No. 2/0135/23 *"Intelligent sensor systems and data processing"*. Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I05-03-V02-00055.

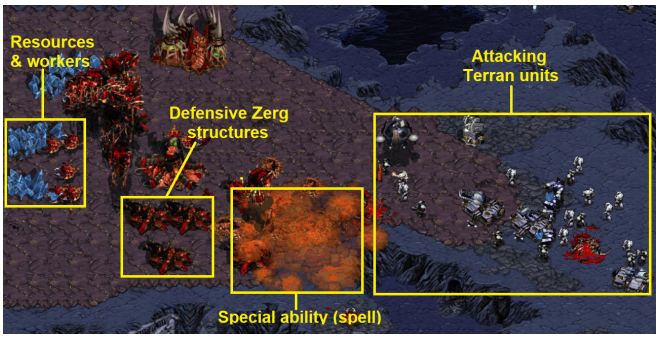


Fig. 1. StarCraft: Brood War

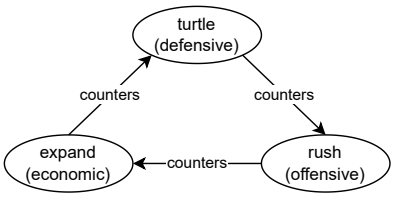


Fig. 2. Rock paper scissors nature of RTS game strategies

research and experimentation, e.g., BWAPI [1], *screp* [2]. Despite considerable progress in their quality over the years, RTS AI agents are not yet capable of beating human expert level players on a consistent basis [3], [9], [12] or are impractical for common video games due to high computational power demands [4], [17].

One possible approach for SCBW agent improvement, especially the strategy selection aspect, is to learn from past matches [12], [14], [15]. SCBW allows recording and archiving of match replays. A vast amount of such data was accumulated throughout the years. For machine learning purposes, a dataset of replays should meet certain requirements to be considered viable and high quality. The largest of such datasets is STARDATA [1], containing 65646 files. Other smaller datasets include [5], [6], [7], [13].

One of the important tasks which can be addressed using

[1]https://github.com/bwapi/bwapi
[2]https://github.com/icza/screp

STARDATA is strategy classification [1]. Knowledge about strategies extracted from replays of highly skilled players can benefit SCBW agents in multiple ways. It can be used during a match to predict opponent activities based on the observed information and help adjust further actions. As a part of the preparation process before and between matches, it can improve the ability to imitate and execute strategies that are statistically proven to be successful or pick from a pool of available strategies most suited for the opponent depending on their race or stored past results against them.

Strategy classification from SCBW replays was attempted [5], [6], [7], [1], but the results of each work were quite limited or very specific in some way with regards to dataset size, quality level of matches, number of strategies or races. Classification from STARDATA was also attempted. The work [8] extracted strategies only for one race (Terran) and [10] for all 3 races, but the selection of strategy defining features (SDFs) was not optimized and could be improved.

In this work, we build upon and improve the process of strategy extraction from [8] in multiple ways. We allow a broader range of strategies to be learned by extending the upper game length limit (Section II-A); refine the unit and structure production frequency calculation formula to provide more precise results (Section II-B); refine the SDF selection process for Terran by also considering the opponent's race (Section II-C); tweak strategy classification by adjusting clustering parameters (Section II-D); and finally, we propose a new technique called *replay snapshotting* (RS) to improve the analysis process of extracted data (Section II-E).

To validate our approach, we implement an AI agent nicknamed *MisHanBot* for SCBW playing Terran race and put it into the BASIL Ladder competition [3]. In this competitive environment, the agent is able to perform the set of strategies extracted by RS against opponents of varying quality. It consistently achieves rank C with current win rate of 56 % which is a significant improvement over our previous approaches.

II. Strategy learning

A. Dataset preparation

We use STARDATA to extract and classify Terran strategies for our SCBW agent from Terran matchups: versus Terran (TvT), versus Protoss (TvP) and versus Zerg (TvZ). *Extractor* proposed in [8] is used to collect raw data. Considered are only matches that meet the following criteria:

a) Valid 1v1, i.e., readable in BWAPI, featuring two opposing players using standard set of units.
b) At least one player is Terran.
c) No longer than 25 minutes.
d) Winner is agreed upon by both tools: *extractor* and *screp*.

Criterion a) filters out any custom or non-standard, non-competitive games. Criterion c) allows the agent to learn larger variety of middle and late-game strategies than 15 minute limit used in [8]. Criterion d) ensures that the quality of extracted strategies can be calculated precisely by inspecting their win rates. To help decide on the winning player, the *screp* utility is used to double-check the results provided by *extractor*. To identify a winner reliably in a non-labeled SCBW replay file is a non-trivial task and various replay analysis tools can produce different outcomes. Dataset cleaning results are shown in Table I. The total number of extracted Terran strategies is 28477; 2 from each TvT matchup (since both players are Terran) and 1 from other matchups (since one player is Terran). Next, raw data are extracted from the prepared dataset.

[3]https://www.basil-ladder.net

TABLE I
Extracted Terran strategies

Matchup	STARDATA replays	Selected replays	Extracted strategies
TvT	2550	1753	3506
TvP	17385	12916	12916
TvZ	14531	12055	12055
Total	34466	26724	28477

B. Data extraction and processing

For raw data extraction, we follow the process [8], obtaining detailed information about each game, including basic metadata (player and map names, match lengths, etc.), timestamps for structure, unit and upgrade creation and destruction, and more. For subsequent data processing and compaction, a few adjustments compared to [8] were made. Firstly, we modify the formula for unit frequency:

$$UnitFreq = \frac{1428.6 \cdot UnitCount}{MatchLength - FirstFrame} \quad (1)$$

where $UnitCount$ is the amount of a unit produced in the entire match, $MatchLength$ is given in game frames and $FirstFrame$ now denotes when the unit's first instance was produced. Previously it denoted when the requirements were first met for the unit's creation. Eq. (1) computes the average frequency of a unit's production per minute (1428.6 frames) since its first instance creation until the match's end. This is a more precise metric to measure unit usage than previously used. Secondly, we now calculate frequency for structures, too, using Eq. (1) but replacing $UnitCount$ by $StructCount$.

In this work, we define SCBW strategies by structure build order and construction frequency and unit composition, omitting technology advancements and upgrades. Put more simply, we are interested in structure and unit information. In the remainder of the text, structures and units together will be denoted as features. Next, a set of important features is selected which will be used to categorize strategies.

C. Feature selection

Terran can build 18 different structures and create 13 types of units directly and few more indirectly (see below). Not all features are equally popular and present in games. If a match is short, there is not enough time to reach advanced features. Additionally, players have discovered over the years which unit compositions are strong against certain opponents.

TABLE II
UNIT SELECTION EXAMPLE FOR TVT. PERCENTAGE OF STRATEGIES WITH INDICATED NUMBER OF UNITS

Unit	1+	5+	10+	Selected
SCV	100.0 %	99.9 %	99.3 %	No, essential
Vulture	80.6 %	56.0 %	44.0 %	Yes
Wraith	44.8 %	19.4 %	8.6 %	Yes
Valkyrie	1.9 %	0.3 %	0.0 %	No, low usage

TABLE III
STRUCTURE SELECTION EXAMPLE FOR TVT

Structure	Prerequisite for	Selected
Barracks	Factory, Academy	No, essential
Factory	Vulture, Tank, Goliath	Yes
Academy	Comsat Station	Yes
Comsat Station	Scan Sweep	Yes
Physics Lab	Battlecruiser	No, low usage

For example, Terran biological units are stronger and used more often against Zerg than other opponents. Consequently, some units are perceived as not effective and are rarely used. On the other hand, some features are essential for making progress and are present in vast majority of matches. If a feature belongs to one of these edge cases, it adds no value to strategy classification and can be disregarded.

Previous approach [8] used the same set of features in all 3 Terran matchups. In this work, we inspect each matchup more closely and consider the opponent's race, allowing us to fine-tune SDFs for each matchup. We select sets of SDFs based primarily on the statistical usage of units. Essential and rarely used units are disregarded. For TvT, units that appear at least one time in at least 40 % of strategies were selected. This bar was lowered to 8 % for TvP and 15 % for TvZ due to different usage distributions. An example of SDF selection for some units in TvT is shown in Table II.

Next, a set of structures is selected to complete SDFs. Again, essential and rarely used structures are disregarded. Terran race tech tree is inspected and for each matchup important structures are selected which are prerequisites for the selected units or other structures. An example of SDF selection for some structures in TvT is shown in Table III.

Selected SDFs are listed in Table IV. Note that although Spider Mines and Scan Sweeps can not be created directly, but are a result of unit ability or spell casting, they are still recognized by BWAPI as units, because they materialize on the battlefield and their statistics can be gathered by *extractor*. Next, compacted data about our selected SDFs from all 28477 Terran strategies are inputs to the unsupervised machine learning process of strategy classification.

D. Strategy classification

To classify SCBW Terran strategies into categories based on the compacted SDF data, we use the K-Means algorithm. It tries to find similarities in the SDF data and group similar strategies into clusters. Previous approach [8] used the same number of clusters (value $K = 10$) for all matchups. In this work, we inspect Terran matchups more closely and use the Elbow method [16] with inertia to find the optimal values of K for each, separately.

TABLE IV
SELECTED STRATEGY DEFINING FEATURES

Structures	Matchups
Factory, Machine Shop, Mis. Turret, Academy, Comsat Station, Starport, Control Tower	TvT, TvP, TvZ
Science Facility, Bunker	TvP, TvZ
Units	**Matchups**
Marine, Siege Tank, Vulture, Goliath, Dropship, Wraith, Spider Mine (unit ability), Scan Sweep (spell)	TvT, TvP, TvZ
Science Vessel	TvP, TvZ
Firebat, Medic	TvZ



Fig. 3. Elbow method for TvT

The results of the Elbow method for TvT are shown in Fig. 3. Among considered values of K (4–6), $K = 5$ was chosen because of the best distribution of average game lengths between individual clusters and also fair distribution of number of matches per cluster. With $K = 5$, we identified 1 cluster containing short games, 2 clusters containing middle length games and 2 clusters containing long games. Detailed results of strategy classification for TvT are summarized in Fig 4. It displays amounts of strategies in clusters.

The resulting value for TvP was $K = 6$: 2 clusters with short, 2 with middle length and 2 with long games. For TvZ it was also $K = 6$: 2 clusters with short, 1 with middle length and 3 with long games.

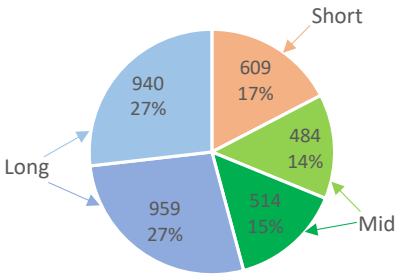


Fig. 4. Strategy classification into clusters for TvT

TABLE V
EXAMPLE OF 2 EXTRACTED STRATEGIES FOR TVT: CLUSTERS '1' AND '3'

Structure	Cl. '1' Freq.	Cl. '3' Freq.	Unit	Cl. '1' Freq.	Cl. '3' Freq.
M. Shop	4.3	3.3	Marine	6.9	3.7
Factory	2.6	3.5	Tank	2.4	2.4
Starport	5.7	3.0	Vulture	3.5	5.2
Academy	6.2	6.0	Goliath	3.8	8.2
Com. Stat.	3.2	5.7	Scan Sw.	4.3	7.2
C. Tower	5.3	4.7	Dropship	6.6	8.1
M. Turret	1.2	6.3	S. Mine	2.2	8.7
-	-	-	Wraith	7.5	2.4

Examples of 2 extracted strategies for TvT are shown in Table V. It shows average frequency rankings of selected SDFs. Lower numbers indicate more frequent appearance while higher numbers indicate less frequent appearance.

Cluster '1' represents a traditional long match strategy using mechanical units: Tanks, Vultures with their ability Spider Mines, and Goliaths. It utilizes structures Factory and Comsat Station often and will construct many Missile Turrets (static defensive). Cluster '3' represents a middle game strategy using a combination of ground (Marine, Tank) and air (Wraith) units. It utilizes structures Starport and Factory often.

The above examples show notable differences between clusters. To find such diversity was an important goal of strategy extraction. A total of 17 strategies were extracted and are available for our Terran agent.

E. Replay snapshotting

Extracted strategies as shown in Table V provide basic guidance on how to proceed in a game for our agent. It learns what types of features to create. We further refine strategies by a new technique called *replay snapshotting* (RS). The basic principle of RS is that a subset of strategies from each cluster is inspected and the amounts of currently existing instances of selected SDFs are snapshotted (sampled) at predetermined checkpoints. This enables our agent to imitate and implement strategies more precisely because it gives the agent more tangible objectives to reach in certain phases of a match. The agent now gains knowledge about how many instances of features are necessary to achieve those partial objectives.

For snapshotting, we have selected strategies that are the best representatives from each cluster. More specifically, strategies used in matches with duration closest to the average cluster match lengths were selected. For example, Cluster '1' from Table V contains matches with an average duration of 18m:31s. Therefore, we select strategies close to this length for RS from this cluster. The number and spacing of checkpoints were set for each cluster individually, based on their average match lengths. The intent is to have checkpoints distributed across all game stages: early, middle and late game, to cover most situations and phases of the entire match. The first checkpoint is usually set at the 4 minute mark because the first important strategic decisions for Terran are made around this time. The last checkpoint is set near the average cluster duration, e.g., at the 18 minute mark for Cluster '1'. The number of strategies selected for RS was 50 per cluster.

TABLE VI
EXAMPLE OF REPLAY SNAPSHOTTING FOR CLUSTER '1' FOR TVT. m–MINUTES

Structure	4m	6m	9m	14m	18m
Factory	1.1	1.7	2.9	4.6	5.8
M. Shop	0.6	1.0	1.9	2.9	3.2
Starport	0.0	0.3	0.4	0.7	0.9
C. Tower	0.0	0.2	0.4	0.6	0.9
Academy	0.0	0.2	0.7	0.9	1.0
Com. Stat.	0.0	0.0	1.0	2.3	2.8
M. Turret	0.0	0.0	0.6	5.7	9.3
Unit	**4m**	**6m**	**9m**	**14m**	**18m**
Marine	2.3	2.1	1.3	0.6	0.3
Tank	0.0	1.7	4.5	11.5	15.8
Vulture	0.3	1.1	3.0	3.5	4.5
Goliath	0.0	0.8	1.1	6.0	9.8
Dropship	0.0	0.1	0.0	0.9	2.5
S. Mine	0.0	1.0	3.2	7.6	11.9
Wraith	0.0	0.1	0.1	0.4	0.4

An example of the RS technique for Cluster '1' for TvT is shown in Table VI. It shows average numbers of instances of selected SDFs snapshotted at exact checkpoints during matches selected for RS. For this cluster, 5 checkpoints were defined at 4, 6, 9, 14 and 18 minute marks. The 4 and 6 minute checkpoints suggest how to proceed in early game and prepare for the transition into middle game. Next, the 9 and 14 minute checkpoints suggest how to transition from early to middle game and prepare for late game. Finally, the last 18 minute checkpoint advises on what structures should be present in the player's base and what is the army composition to aim for in late game. Note that in Table VI RS is not able to track any instances of Scan Sweep spell (one of selected TvT SDFs) because it has only a very short duration and then it disappears from the battlefield.

To implement this strategy, the agent should start by constructing some Factories with Machine Shop, start training some Marines and Vultures, later adding Tanks and Goliaths, research the Vulture's ability to lay Spider Mines and start using them. In middle game, it should add more Factories with Machine Shops, construct an Academy and then Comsat Stations and start producing Missile Turrets. Army should now contain more Tanks and Goliaths while keeping adding Vultures in small numbers and increase the usage of Spider Mines and also slowing down Marine production. A late game player's base should have a healthy number of Factories with Machine Shops, an Academy, some Comsat Stations and many Missile Turrets should be scattered around the battlefield. Late game army composition should comprise mainly Tanks with Goliaths both in healthy numbers with the support of Vultures

and few Dropships and many Spider Mines should be laid on the battlefield. Starting from middle game, the agent can also consider occasionally constructing less used structures such as Starport or Control Tower and training some less used units such as Wraiths.

Based on values obtained by RS, we divide features into three categories:

- Mandatory: $Value >= 1$. Present in majority of matches. The agent should try to construct the indicated number or more of them.
- Optional: $Value \in (0; 1)$. Not present in each match. The agent can consider construction of one instance in case of structures and few instances in case of units.
- Not advisable: $Value = 0$. Not present in any match. It is not advised for the agent to construct these.

In Table VI, an example of a mandatory feature is Factory at 9 minute mark ($Value = 2.9$). The agent should try to have at least 2 but optimally 3 or more of them at this checkpoint. An example of an optional feature is Machine Shop at 4 minute mark ($Value = 0.6$). The agent can consider construction of one at this checkpoint.

Strategies extracted and refined by RS (RS strategies) provide very detailed guidance for the entire match duration as can be seen in the above example. The next objective is to implement them into our agent.

III. Strategy implementation

Our agent *MisHanBot* is based on the public version of *Steamhammer* (SH) [4], a popular starter agent for development and experimentation. Its main race is Zerg, but can also play other races, however, on a lower level than the main one. It also supplies a set of starting strategies and provides many necessary functionalities of a SCBW agent already implemented on a basic level. While adopting existing functionalities from SH, like unit micromanagement, our agent also has many small improvements over baseline SH Terran, including better unit control and unit repair logic, or better ability and spell usage. Terran strategies are implemented in SH by:

- *Configuration file (config)*, an explicit list of items to create and other commands to perform from the start of the match. This list can contain all features, not only those selected as SDFs (Table IV). Depending on its length, it can be used for all game stages.
- *BOSS*, a module which takes over and continues to implement the strategy after all items in *config* have been created. It tries to keep the 'trend' set by *config* and continues with construction of items present in the list.

Strategy selection during a match is based on randomness with weighted probabilities. A pool of strategies is defined for various situations (matchups, opponent's plans, etc.), and each strategy in the pool is assigned a weight based on suitability for tackling that situation. When a situation occurs, SH selects one strategy from that situation's pool randomly.

[4]https://satirist.org/ai/starcraft/steamhammer/

TABLE VII
WIN RATES AND WEIGHTS FOR TVT STRATEGY POOL

Strategy (Cluster)	'0'	'1'	'2'	'3'	'4'
Win rate	52.8	55.5	28.7	50.3	53.6
Weight	22	23	12	21	22

We implement RS strategies into our agent as follows. We overwrite *config* so that it contains only our Terran RS strategies. We define 3 situations: TvT, TvP, TvZ. Note that in our agent, we do not create situations considering opponent's plans. Our strategy selection depends only on the opponent's race. Next, we create a strategy pool for each situation. To account for the randomness of RS strategies (see Section II-E), we round up the advised values from Table VI to the nearest integers. This means that values of all optional features will be rounded to 1, i.e., the agent can try to have one instance of them. Values of all mandatory features will be increased, too. Therefore, the agent is given the upper bounds of what numbers to aim for. After this adjustment, we put RS strategies into strategy pools in *config*; TvT strategy pool includes 5 strategies, TvP and TvZ pools both include 6 strategies.

The explicit lists in *config* for each RS strategy are set up using knowledge of the Terran tech tree while trying to follow RS advice (example in Table VI) as closely as possible.

Next, we set weights for strategy pools. Weights are assigned to extracted strategies in pools proportional to their win rate. Sum of weights in a pool is equal to 100. More successful strategies get higher weights and in turn will be picked more often in games than those with lower weights. TvT win rates were obtained as part of this work (see Section II-D). TvP and TvZ win rates were obtained from collaboration with the authors of two other agents (see Section IV) who did strategy classification for Protoss and Zerg races. Win rates and corresponding weights for the TvT strategy pool are shown in Table VII. It displays average win rates against all other Clusters in the pool. When our agent is in a TvT situation, Cluster '1' has the highest win rate and in turn the highest chance to be used: 23 %. On the other hand, Cluster '2' is the least successful and has the lowest chance to be used: 12 %.

During a match after processing *config*, strategy control of our agent is handed over to the *BOSS* module, similar to SH. It will follow a trend established by *config* and keep producing features as specified in the last RS checkpoint in the remainder of the match.

IV. Results and conclusion

At the time of writing this paper, our agent *MisHanBot* is active in the BASIL Ladder competition. It has rank C with ELO rating 2586, overall win rate 49.4 %, recent win rate 56.3 % and is situated at the 46th place out of 101 currently active agents. The overall number of agents in the ladder is 159. Agents ranked 102 or below are temporarily disabled to keep the competition's level high. They can be re-enabled again under certain conditions.

TABLE VIII
OUR STARDATA-TRAINED AGENT RESULTS IN BASIL LADDER AS OF MAY, 15TH 2025

Agent	ELO rating	BASIL rank	Games played	Win rate	Win rate 7 days
KasoBot (T) [8]	2390	E	48821	38.9 %	41.7 %
Azergo (Z)	2206	E	9034	28.9 %	26.1 %
TommyBot (P)	2203	F	9057	19.6 %	25.7 %
MisHanBot (T) [This work]	2586	C	10253	49.4 %	56.3 %

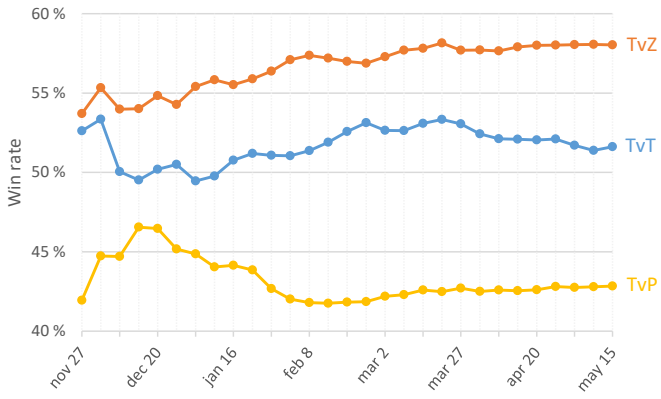


Fig. 5. Win rates of our agent against each opponent race in the last 6 months (Nov 2024–May 2025). Source: BASIL Ladder https://www.basil-ladder.net/bot.html?bot=MisHanBot

The comparison with our other currently active agents is shown in Table VIII. Results obtained by our new agent *MisHanBot* are a significant improvement over our other approaches. Note that both agents *Azergo* and *TommyBot* extract detailed data and learn strategies from STARDATA for their particular races, but use different approaches for their implementation. Strategies obtained by *Azergo* are not as effective and could be improved. *TommyBot* is an experimental agent built from scratch, i.e., it does not use any of available popular starter agents, such as SH.

Fig. 5 displays win rate trends of our agent in the last 6 months against all three opponent races. Results indicate that TvP win rate (currently at 42.8 %) is not satisfactory as it consistently trails behind other two matchups and brings the agent's overall win rate down considerably. On the other hand, TvZ win rate (currently at 58.0 %) is satisfactory while in TvT it is slightly above the average level (currently at 51.6 %).

Our results indicate that the proposed *replay snapshotting* approach for strategy refinement is beneficial for overall SCBW AI agent improvement. Our agent is able to imitate extracted strategies more accurately than our previous approaches and it performs better in a competitive environment. As future work, it could be worth looking more closely into TvP particularly, identify and remedy the weaknesses of our agent in this matchup. Other improvements could include: further SDF selection fine-tuning by adding or removing items; RS adjustments by considering the total amount of produced features at checkpoints instead of counting currently existing instances; including more matches per cluster into RS; adding checkpoints beyond the average cluster time; fine-tuning checkpoint distribution; setting probabilities of strategies within pools with low win rates to 0 to avoid their usage during matches; and more.

REFERENCES

[1] Z. Lin, J. Gehring, V. Khalidov and G. Synnaeve, "STARDATA: A StarCraft AI Research Dataset," *13th AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2017,* pp. 50–56, https://dx.doi.org/10.48550/arXiv.1708.02139

[2] S. Ontañon, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill and M. Preuss, "A Survey of Real-Time Strategy Game AI Research and Competition in StarCraft," *IEEE Trans. Computational Intelligence and AI in games, IEEE Computational Intelligence Society, 2013,* 5(4), pp. 293–311, https://dx.doi.org/10.1109/TCIAIG.2013.2286295

[3] Mi. Čertický, D. Churchill, K.-J. Kim, Ma. Čertický and R. Kelly, "StarCraft AI Competitions, Bots and Tournament Manager Software," *IEEE Trans. Games, 2018,* 11(3), pp. 227–237, https://dx.doi.org/10.1109/TG.2018.2883499

[4] O. Vinyals, I. Babuschkin et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature, 2019,* 575, pp. 350–354, https://dx.doi.org/10.1038/s41586-019-1724-z

[5] B. G. Weber and M. Mateas, "A data mining approach to strategy prediction," *IEEE Symp. Computational Intelligence and Games, 2009,* pp. 140-147, https://dx.doi.org/10.1109/CIG.2009.5286483

[6] H. C. Cho, K. J. Kim and S. B. Cho, "Replay-based strategy prediction and build order adaptation for StarCraft AI bots," *IEEE Conf. Computational Intelligence in Games (CIG), 2013,* pp. 1-7, https://dx.doi.org/10.1109/CIG.2013.6633666

[7] G. Synnaeve and P. Bessière, "A Dataset for StarCraft AI & an Example of Armies Clustering," *Artificial Intelligence in Adversarial Real-Time Games, 2012,*, https://dx.doi.org/10.48550/arXiv.1211.4552

[8] Š. Krištofík, P. Malík, M. Kasáš, Š. Neupauer, "StarCraft agent strategic training on a large human versus human game replay dataset," *Federated Conf. Computer Science and Information Systems, FedCSIS 2020,* 21, ACSIS, pp. 391–399, https://dx.doi.org/10.15439/2020F178

[9] M. Świechowski, "Game AI Competitions: Motivation for the Imitation Game-Playing Competition," *Federated Conf. Computer Science and Information Systems, FedCSIS 2020,* 21, ACSIS, pp. 155–160, https://dx.doi.org/10.15439/2020F126

[10] Š. Krištofík, M. Kasáš, P. Malík, "StarCraft strategy classification of a large human versus human game replay dataset," *Federated Conf. Computer Science and Information Systems, FedCSIS 2021,* 25, ACSIS, pp. 137–140, https://dx.doi.org/10.15439/2021F48

[11] G. Robertson, I. Watson, "A Review of Real-Time Strategy Game AI," *AI Magazine, 2014,* 35(4), pp. 75–104, https://dx.doi.org/10.1609/aimag.v35i4.2478

[12] S. Xu, H. Kuang et al., "Macro action selection with deep reinforcement learning in StarCraft," *15th AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2019,* pp. 94–99, https://dx.doi.org/10.48550/ARXIV.1812.00336

[13] J. J. Merelo-Guervós, A. Fernández-Ares et al., "RedDwarfData: a simplified dataset of StarCraft matches," 2017, https://dx.doi.org/10.48550/arXiv.1712.10179

[14] F. Dai, J. Gong, J. Huang, J. Hao, "Macromanagement and Strategy Classification in Real-Time Strategy Games," *2nd China Symp. Cognitive Computing and Hybrid Intelligence (CCHI), 2019,* pp. 263–267, https://dx.doi.org/10.1109/CCHI.2019.8901957

[15] N. Justesen, S. Risi, "Learning Macromanagement in StarCraft from Replays using Deep Learning," 2017, https://dx.doi.org/10.48550/arXiv.1707.03743

[16] C. Shi, B. Wei et al., "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *J. Wireless Comm. and Networking, 2021,* https://dx.doi.org/10.1186/s13638-021-01910-w

[17] N. Justensen, M. Kaselimi et at., "Human-like Bots for Tactical Shooters Using Compute-Efficient Sensors," 2024, https://dx.doi.org/10.48550/arXiv.2501.00078

[18] J. Gehring, D. Ju, V. Mella, D. Gant, N. Usunier, G. Synnaeve, "High-Level Strategy Selection under Partial Observability in StarCraft: Brood War," 2018, https://dx.doi.org/10.48550/arXiv.1811.08568

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 321–326
DOI: 10.15439/2025F0041
ISSN 2300-5963 ACSIS, Vol. 43

Applying Evolutionary Techniques to Enhance Graph Convolutional Networks for Node Classification: Case Studies

Maciej Krzywda*, Szymon Łukasik*, Amir H. Gandomi†
*Faculty of Physics and Applied Computer Science, AGH University of Krakow,
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: krzywda@agh.edu.pl, slukasik@agh.edu.pl
†Faculty of Engineering and IT, University of Technology Sydney,
5 Broadway, Ultimo NSW 2007, Australia
Email: gandomi@uts.edu.au
‡Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland
§NASK National Research Institute,
ul. Kolska 12, 01-045 Warsaw, Poland
¶University Research and Innovation Center (EKIK), Óbuda University,
Bécsi út 96/B, Budapest, 1034, Hungary
‖Department of Computer Science, Khazar University,
Baku, Azerbaijan

***Abstract*—In recent years, significant efforts have been made to address graph node classification tasks by applying graph neural networks and methods based on label propagation. Despite the progress achieved by these approaches, their success often hinges on complex architectures and algorithms, sometimes leading to the oversight of crucial technical details. In designing artificial neural networks, one crucial aspect of the innovative approach is suggesting a novel neural architecture. Currently used architectures have mostly been developed manually by human experts, which is a time-consuming and error-prone process. That is why the adoption of more sophisticated semi-automatic methods, such as Neural Architecture Search, has become commonplace. This paper introduces and assesses an evolutionary-based approach for the design of graph convolutional neural networks in the context of node classification. Our approach aims to systematically define the graph convolutional networks parameter space, drawing inspiration from recent research on design principles. By doing so, our method seeks to strike a balance between achieving satisfactory performance and optimizing memory and computation resources, thus offering a more efficient alternative to conventional approaches from the neural architecture search area.**

***Index Terms*—graph neural networks, graph convolutional networks, genetic algorithm , node classification, evolutionary computation**

I. Introduction

NEURAL Architecture Search (NAS) [1]–[4] is gaining popularity as a fully automatic method for designing Artificial Neural Networks Architectures (ANNs). Neural Architecture Search is a method that allows the creation of comparable and even superior architectures in terms of performance to those that have been designed manually. In essence, NAS streamlines the process of a human adjusting a neural network through trial and error to identify successful configurations and automates it to discover more intricate structures. It comprises a range of techniques and tools that evaluate numerous network architectures within a defined search space using a search strategy and choose the one that accomplishes the goals of a specific problem by maximizing a fitness function.

The manuscript is structured as follows. Firstly, Section II presents a concise overview of Neural Architecture Search approaches applied across various domains of neural networks. Following that, Section II-A provides a comprehensive summary of the state-of-the-art in Graph Neural Networks, encompassing prevalent methods employed in the design of GNN architectures. In Section III, we delve into the description of the evolutionary algorithm utilized in our research, elucidating details such as fitness metrics and the delineation of the search space. Section IV comprehensively covers settings and implementation details, including the datasets employed and the preparation of the computational budget. Section V is dedicated to a thorough discussion of the attained results, while Section VI serves as a conclusive synthesis of our research findings.

II. Neural Architecture Search

NAS is computationally and time-consuming solution typically involving intensive calculations using graphics processing units (GPUs). Therefore, researchers and research groups are increasingly daring to look for alternative methods to optimize costs and find the most efficient and effective neural network architecture in order to solve their research problem.

Today, the most popular options for NAS are methods based on reinforcement learning [2] (including one-shot methods) [5], evolutionary computing (including multiobjective search), Bayesian optimization [6], and hill-climbing [7].

In article [4], the authors defined NAS as a system composed of three primary elements: search space, search algorithm, and evolutionary strategy.

- **Search space** - outlines a group of operations, such as convolution, pooling, and fully connected layers, and determines how these operations can be combined to create valid network architectures. Developing the search space often involves human expertise, inevitably introducing biases.
- **Search algorithm** - selects a group of candidates for network architecture and tests their performance. It receives child model performance metrics, such as accuracy and low latency, as rewards and utilizes this information to optimize and produce high-performance architecture candidates. As the possible search algorithms for NAS, one can name, e.g., algorithms like Random Search, Reinforcement Learning, and Bayesian Optimization.
- **Evaluation strategy** - in order to provide feedback for the search algorithm to learn, we must evaluate the performance of numerous proposed child models through measurement, estimation, or prediction. This process of evaluating candidates can be quite costly, prompting the proposal of various new methods to conserve time and computing resources. For these strategies, we can include proxy task performance, low-fidelity performance estimation, weight inheritance, weight sharing, learning curve extrapolation, and network morphism. [8], [9]

Based on the above definition of NAS we can define the NAS using Evolutionary Computation (EC) as a task of exploring architectures search space by generating and evaluating a population of candidate architectures and then evolving this population over multiple generations to find better-performing architectures. As we mentioned, ECs are one of the methods used to realize NAS, but the most promising results were achieved recently using the following approaches:

- **NAS-EA-FA** [10] - evolutionary algorithm based on the fitness approximation to the search for neural architecture. It is designed to accelerate the search process.
- **EF-ENAS** [11] - evolutionary neural architecture search algorithm based on evaluation corrections and fitness sharing. It is designed to improve the efficiency and effectiveness of the search process.

for a more extensive survey of evolutionary NAS one can refer to [8], [9].

A. Neural Architecture Search for GNNs

Remarkably, the field of NAS within GNNs remains relatively unexplored. Notable exceptions include GraphNAS [12]–[14], which uses RL to identify architectures for node classification tasks. The search space defined by GraphNAS includes sampling, aggregation, and gated functions, demonstrating its efficacy in the realm of GNNs. Auto-GNN [15] follows suit by adopting RL and a comparable search space. SNAG framework (Simplified Neural Architecture Search for Graph Neural Networks [16]) addresses drawbacks in existing neural architecture search (NAS) methods, such as GraphNAS and Auto-GNN, by introducing an innovative search space and a reinforcement learning-based search algorithm, as demonstrated through extensive experiments on real-world datasets.

Another novel approach to NAS for GNN is Scalable Graph Neural Architecture Paradigm [17] representing a novel paradigm introduced in the paper ”PaSca.” It serves as a systematic approach to construct and explore the design space for scalable Graph Neural Networks (GNNs), aiming to address the scalability challenges inherent in traditional GNNs designed based on the neural message passing mechanism. SGAP is presented as a key element of the PaSca system, offering a principled framework for designing GNNs that can efficiently handle larger datasets and message passing steps, thus enhancing overall scalability. Neural Architecture Coding (NAC) [18] is a Neural Architecture Search (NAS) method for Graph Neural Networks (GNNs) that addresses computational cost and optimization challenges. It utilizes a sparse coding objective to find optimal architecture parameters without weight updates after random initialization. NAC claims linear-time efficiency, and empirical evaluations on GNN benchmark datasets show that it outperforms strong baselines. The method’s strength lies in leveraging the expressive power of GNNs with randomly-initialized weights for efficient architecture discovery.

[19] Automated Graph Neural Network(Auto-HeG) on Heterophilic Graphs addresses the limitations of existing graph neural architecture search (NAS) methods, which primarily focus on the homophily assumption and neglect heterophily, a crucial graph property in real-world applications. Auto-HeG aims to automatically design powerful graph neural networks with expressive learning abilities for heterophilic graphs, incorporating heterophily into various stages of the automatic heterophilic graph learning process. This includes aspects such as search space design, supernet training, and architecture selection, using a diverse message-passing scheme, progressive supernet training, and a heterophily-aware distance criterion to derive specialized and expressive heterophilic GNN architectures. The proposed method demonstrates its superiority through extensive experiments, showcasing its effectiveness compared to human-designed models and existing graph NAS models.

Auto-GNAS (Automatic Graph Neural Architecture Search) [20] addresses the challenge of constructing effective graph neural networks (GNNs) for non-euclidean data by automating the search for optimal GNN architectures. Unlike traditional methods that evaluate architectures serially, Auto-GNAS employs parallel evaluation, utilizing multiple genetic searchers simultaneously. These searchers use evaluation feedback, information entropy, and results from other searchers through a sharing mechanism to enhance efficiency. Auto-GNAS demonstrates competitive model performance and improved search efficiency compared to other algorithms, marking the first

instance of using parallel computing to improve the system efficiency of graph neural architecture search. Our approach allows for a quick and simple (computationally inexpensive) way to limit the spatial hyperparameters of a graph neural network.

III. Evolutionary algorithm for Designing GCN

Genetic Algorithm (GA) is part of the evolutionary algorithm family, inspired by natural selection and genetics. The exploration space is a crucial element in all GAs, represented as strings called chromosomes or individuals. A population comprises a collection of such strings. Each chromosome consists of a sequence of genes encoding a solution to a target problem. Initially, a diverse population of random individuals is created to represent various potential solutions for the target problem. Each individual is assigned a fitness value that indicates its quality within the population.

During the evaluation step, the fitness of each individual is calculated. The selection step involves choosing certain individuals from the entire population as parents for mating. In the cross-step, parents exchange information, such as swapping genes, to generate new individuals and form the next generations. The mutation step introduces diversity into the population by randomly altering a gene in new individuals based on the probability of mutation. Finally, the population is updated by incorporating new individuals.

In this approach, the individual is a single representation of a graph convolutional network with a random configuration that allows it to compile and run successfully on selected datasets. The genetic algorithm used in research is the mutation and cross-over operators. Mutation involves randomly selecting a new gene value from a uniform distribution within the range of values available for that gene. The crossover function takes a list of parents as an argument, from which it randomly selects two individuals. Each individual then contributes one part of his genes (complementary to the part contributed by the other individual). Subsequently, a mutation is introduced to the obtained gene. The random selection process is repeated until the specified population size is achieved. The algorithm dynamically adjusts the number of layers based on the specified `num_layers` parameter for each agent (more params of individuals are described in Tables **??** and I, providing adaptability in the neural network architecture for each agent in the `deep_GCN` class involves iterating over the specified number of layers (`num_layers`). For every layer, the algorithm determines the specific parameters (described in TableI).

A. Fitness

In our approach, `fitness_value` serves as a measure of the classifier quality that you aim to minimize. It is the sum of squared F1 errors for each class. The F1 scores for individual classes are numerical metrics that represent the classification quality of each class. This considers both precision and recall for a given class, making them more informative than accurate, especially in the context of unbalanced data sets. `fitness_value` is the sum of squared F1 errors across all classes. This approach provides an overall assessment of the quality of the classifier in multiple classes. Squaring the F1 error emphasizes larger errors, i.e., greater deviations from the ideal F1 score. The goal is to minimize `fitness_value`. Minimizing this value implies an effort to enhance the classifier's quality for all classes simultaneously, offering a more balanced objective than merely maximizing accuracy. Evaluating the classification quality for each class individually allows the identification and improvement of areas where the classifier may perform suboptimally.

In summary, the `fitness_value`-based approach considers more nuanced aspects of classification quality, taking into account both false positives and false negatives, as well as precision and recall. This is a more balanced approach than relying solely on accuracy, especially in scenarios involving imbalanced datasets or multiclass classification.

B. Search space

In this section, we delve into the details of the search space employed in our approach, elucidating the ranges of various parameters within it. The search space is a crucial aspect of our methodology, and in Table I, we provide a comprehensive overview of the hyperparameter ranges relevant to the training of neural networks. It is essential to note that these ranges remain consistent across all datasets, ensuring a standardized approach to model optimization.

However, the uniqueness of each dataset requires specific considerations for certain parameters, particularly Hidden Channels and Batch Size, as explained in Table **??**. For Hidden Channels, the range is dynamically defined as range from 10 to the number of features within the dataset. This adaptability reflects the intrinsic characteristics of the data and aims to optimize the model architecture accordingly. Similarly, for the batch size parameter, the range ranges from a minimum of 10 to the total count of all training masks present in the dataset, as detailed in Table II.

To provide a concise reference, Table I encapsulates the overarching hyperparameter ranges for training neural networks, emphasizing their uniform applicability across diverse datasets. The following table **??** then details the data set-specific nuances, delineating the distinct ranges for Hidden Channels and Batch Size. This approach ensures that our search space is not only comprehensive but also adaptive to the intricacies of individual datasets, fostering effective model optimization.

IV. Experimental Settings

A. Datasets

In Table II, we provide a comprehensive overview of the key parameters associated with the datasets utilized in our experimental efforts. Brief characteristics of each dataset are described below.

- **Planetoid Datasets**, including Cora, CiteSeer, and PubMed, are widely used in graph-based learning tasks.

TABLE I: Neural Network Hyperparameter Ranges

Parameter	Range
Epochs	100
Learning Rate	From 0.0001 to 0.1
Weight Decay	From 0.0001 to 0.1
Layers	From 1 to 10
Optimizers	Adadelta, Adagrad, Adam, AdamW, Adamax, ASGD, NAdam, RAdam, RMSprop, Rprop, SGD, QHM, QHAdam, PSO, A2GradExp, A2GradInc, A2GradUni, AccSGD, AdaBelief, AdaBound, AdaMod, Adafactor, Adahessian, AdamP, AggMo, Apollo, DiffGrad, Lamb, MADGRAD, NovoGrad, PID, Ranger, RangerQH, RangerVA, SGDP, SGDW, SWATS, Shampoo, Yogi
Loss Function	CrossEntropy, NLLLoss, MultiMarginLoss, MultiLabelMarginLoss, SmoothL1Loss

They feature academic citation networks where nodes represent documents, and edges denote citations.

- **WikipediaNetwork Datasets**, comprising Squirrel and Chameleon, focus on analyzing relationships within Wikipedia. For example, Squirrel captures the network structure of Wikipedia articles, enabling researchers to explore connections and interactions between different topics.
- **WebKB Datasets**, such as Cornell, Texas, and Wisconsin, are tailored for web page classification. These datasets encompass web pages from university domains and are labeled with categories such as student, faculty, project, and course. They serve as benchmarks for text classification and information retrieval tasks.
- **Actor Dataset** is commonly utilized in social network analysis, representing a network of actors and their collaborations in movies or television shows. The nodes correspond to actors, and the edges represent collaborations in specific projects.

The selection of these datasets is grounded in their prevalence in academic research and practical applications. Researchers and practitioners often turn to these datasets as benchmarks due to their well-defined structures, facilitating the rigorous evaluation of diverse graph-based machine learning algorithms.

B. Computational Settings

We have defined a computational budget for our evolutionary algorithm with the following specifications:

- 300 generations
- 100 training epochs
- 10 initial populations (random)
- whole population is 20
- mutation rate is 0.3

The search space is 8-dimensional, denoted as **batch size**,**hidden channels**, **learning rate**, **weight decay**, **num layer**, **optimizer**, **loss function**. In particular, **the epoch number** is held constant throughout the exploration. Using the previously defined **fitness value** enables us to conduct innovative research within this fixed dimensionality.

V. RESULTS

We observed the mean fitness and accuracy for each generation and, for better visualization, we also tracked the `best fitness` and `best accuracy` achieved throughout the evolutionary process. While analysing the results enclosed in Table II and III, it can be noticed that minimizing the cost in multi-class classification problems is more complex than in the case of binary classification, which can also be observed in our case - the more complex the dataset, the potentially more difficult it is to prepare a model that will allow for the correct classification of all nodes from high effectiveness. Our research shows that using only GCN is not an ideal solution for datasets such as Actor, Squirrel or Chameleon, which are more complex datasets, e.g. heterophily ratio, characterized by more Nodes, average node Degree, or Node Features (in Table II) the less accurate the classification, which can be observed in the table.

In Table III, we present the optimal genes, which represent the best solutions obtained through optimization using the genetic algorithm.

These adjustments ensure that the experimental parameters adhere to the requirement of being integer values. Although presented in the table as floating values for clarity, they are effectively treated as integers during the genetic algorithm optimization process. This approach maintains consistency and allows for a meaningful and applicable exploration of the solution space.

Analysis of the experimental results reveals key patterns in hyperparameter configurations for Graph Convolutional Networks (GCN) across various datasets (presented in III. The most prevalent configuration associated with optimal performance includes the Adamax optimizer, the cross-entropy loss function, and 1 to 3 layers of GCN. In particular, configurations with a single or two layers of GCN emerged more frequently, suggesting their effectiveness in achieving high fitness values and classification accuracy. These findings offer practical guidance for researchers and practitioners seeking robust hyperparameter choices for GCN models [21], [22] for Node Classification.

TABLE II: Dataset Characteristics

Parameter	**Wisconsin**	**Actor**	**Texas**	**Squirrel**	**PubMed**	**Cornell**	**Cora**	**CiteSeer**	**Chameleon**
Nodes	251	7600	183	5201	19717	183	2708	3327	2277
Edges	515	30019	325	217073	88648	298	10556	9104	36101
Avg. Node Degree	2.05	3.95	1.78	41.74	4.50	1.63	3.90	2.74	15.85
Training Nodes	1200	36480	870	24960	60	870	140	120	10920
Isolated Nodes	False	False	False	False	False	False	False	True	False
Self-Loops	True	True	True	True	False	True	False	False	True
Is Undirected	False	False	False	False	True	False	True	True	False
Features	1703	932	1703	2089	500	1703	1433	3703	2325
Classes	5	5	5	5	3	5	7	6	5
Node Features	1703	932	1703	2089	500	1703	1433	3703	2325
Homophily ratio (h)	0.21	0.22	0.11	0.22	0.8	0.3	0.81	0.74	0.23

TABLE III: Best results achieved for each dataset

Param	Wisconsin	Texas	Squirrel	PubMed	Cornell	Cora	CiteSeer	Chameleon	Actor
Batch Size	630	500	837	280	210	605	358	509	87
Hidden Channels	267	1633	1384	338	348	1245	230	623	1647
Epochs	100	100	100	100	100	100	100	100	100
Learning Rate	0.0087	0.0118	0.0143	0.0848	0.0073	0.0022	0.0054	0.0826	0.0674
Weight Decay	0.0312	0.0023	0.0169	0.0019	0.0011	0.0009	0.0083	0.0039	0.0151
Layers	1	1	1	3	2	4	2	4	1
Loss Function	SmoothL1Loss	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy
Optimizer	Adam	NAdam	Adadelta	Adam	Adadelta	Adadelta	PSO	Adam	Adagrad
Minimum Fitness	1.4624	1.7088	1.6277	0.3627	1.4322	0.4845	0.8082	1.2427	1.6963
Accuracy	0.5882	0.6486	0.3121	0.7960	0.5135	0.8290	0.7320	0.4496	0.2539

It is surprising that NAdam and QHAdam [23] along with Particle Swarm Optimization [24] are algorithms that have achieved high Accuracy values and low fitness. Moreover, the dominance of Adagrad over the most popular Adam is noteworthy. Due to fewer options for loss functions, Categorical Cross Entropy dominates here and possibly achieves better results for other loss functions, which are not included in our research. We compare our methods with baselines methods [25], [26], which are presented in Table IV.

Based on the results presented above in Table IV, we observe that the results achieved for GCN based on the most popular implementation of GCN [26] are very close to the best accuracy. Our solution enhances the standard GCN models, though it does not quite reach the level of the most advanced ones [27], [28]. Nevertheless, it can serve as an inspiring starting point for future research in this field. Homophily ratio [29] in datasets with high values (Actor, Wisconsin, Squrrel) refers to the tendency of nodes with similar features or attributes to be more likely connected to each other. In datasets with a high homophily ratio (CiteSeer, Cora, PubMed), the relationships between similar nodes become more pronounced, potentially improving classification performance in tasks that rely on node similarities.

VI. Conlusions

In the last few years, studies on designing Artificial Neural Networks (ANNs) have become an active research field, mainly due to the advanced cost training and prototyping of underlying deep learning architectures. The proposed study focuses on developing the first stage of a method for the design and optimization of graph-convolutional neural networks.Our method enables avoiding mindless reliance on closed-box approaches, where one often waits for the final outcome without considering what transpires throughout the entire process. Assuming that the parameter search space of a graph convolutional neural network is multidimensional (each dimension defined by a network parameter), we conclude that by increasing the number of dimensions and their size, such as utilizing various types of layers (not just GCNConv, as in our study, but also GATConv, GIN, SGCNN etc.), we can construct the best possible neural network. This is achieved without the need to specify in advance whether our network should be GCN or GAT-based. Our study provides answers that GCN is not always the ideal solution. By extending our proposed mechanism, we can achieve an ideal architecture for solving a specific node classification problem as a result, our method allows for flexibility and adaptation to the specific requirements of a given situation.Plans for this research include the research of another search optimization framework based on Genetic Programming (including Cartesian Genetic Programming, the utility of which was previously demonstrated by designing CNNs [30]), Evolutionary Strategies, and Bio-Inspired optimization methods [31]. This approach will allow greater flexibility in designing and optimizing not only graph coefficient neural networks but all available Graph Neural Networks (GAT, GIN, Graph Transformer etc.). We anticipate that the results of this study might be useful to researchers in the field of Neural Architecture Search because the results obtained in the experiment are very promising.

TABLE IV: Comparison of accuracy of our approaches with state-of-the-art methods and GNN baselines, and performance of different models on various datasets

Method	Wisconsin	Texas	Squirrel	PubMed	Cornell	Cora	CiteSeer	Chameleon	Actor
GCN [26]	45.88	52.16	23.96	88.13	52.70	85.77	73.68	28.18	26.86
GAT [25]	49.41	58.38	30.03	87.62	54.32	86.37	74.32	42.93	28.45
Geom-GCN-P [27]	64.12	67.57	38.14	88.09	60.81	84.93	75.14	60.90	31.63
GGCN [28]	86.86	84.86	55.17	89.15	85.68	87.95	77.14	71.14	37.54
Our	58.82	64.86	31.21	79.60	51.35	82.90	73.20	44.96	25.39

ACKNOWLEDGMENTS

The work was supported by statutory tasks of the AGH UST Faculty of Physics and Applied Computer Science within the MEiN grant. We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2022/015677 and no. PLG/2023/016643.

REFERENCES

[1] C. Liu, B. Zoph, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. L. Yuille, J. Huang, and K. P. Murphy, "Progressive neural architecture search," in *European Conference on Computer Vision*, 2017.

[2] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *ArXiv*, vol. abs/1611.01578, 2016.

[3] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *International conference on machine learning*. PMLR, 2018, pp. 4095–4104.

[4] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *ArXiv*, vol. abs/1808.05377, 2018.

[5] G. Bender, P.-J. Kindermans, B. Zoph, V. Vasudevan, and Q. V. Le, "Understanding and simplifying one-shot architecture search," in *International Conference on Machine Learning*, 2018.

[6] C. White, W. Neiswanger, and Y. Savani, "Bananas: Bayesian optimization with neural architectures for neural architecture search," in *AAAI Conference on Artificial Intelligence*, 2019.

[7] M. Verma, P. P. Sinha, K. Goyal, A. Verma, and S. Susan, "A novel framework for neural architecture search in the hill climbing domain," *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 1–8, 2019.

[8] Y. Liu, Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "A survey on evolutionary neural architecture search," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 550–570, 2020.

[9] X. Zhou, A. K. Qin, Y. Sun, and K. C. Tan, "A survey of advances in evolutionary neural architecture search," *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 950–957, 2021.

[10] C. Pan and X. Yao, "Neural architecture search based on evolutionary algorithms with fitness approximation," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[11] R. Shang, S. Zhu, J. Ren, H. Liu, and L. Jiao, "Evolutionary neural architecture search based on evaluation correction and functional units," *Knowledge-Based Systems*, vol. 251, p. 109206, 2022.

[12] Y. Gao, H. Yang, P. Zhang, C. Zhou, and Y. Hu, "Graph neural architecture search," in *International Joint Conference on Artificial Intelligence*, 2020.

[13] ——, "Graphnas: Graph neural architecture search with reinforcement learning," *ArXiv*, vol. abs/1904.09981, 2019.

[14] Y. Gao, P. Zhang, H. Yang, C. Zhou, Z. Tian, Y. Hu, Z. Li, and J. Zhou, "Graphnas++: Distributed architecture search for graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 6973–6987, 2023.

[15] K. Zhou, Q. Song, X. Huang, and X. Hu, "Auto-gnn: Neural architecture search of graph neural networks," *Frontiers in Big Data*, vol. 5, 2019.

[16] H. Zhao, L. Wei, and Q. Yao, "Simplifying architecture search for graph neural network," *ArXiv*, vol. abs/2008.11652, 2020.

[17] W. Zhang, Y. Shen, Z. Lin, Y. Li, X. Li, W. Ouyang, Y. Tao, Z. Yang, and B. Cui, "Pasca: A graph neural architecture search system under the scalable paradigm," *Proceedings of the ACM Web Conference 2022*, 2022.

[18] P. Xu, L. Zhang, X. Liu, J. Sun, Y. Zhao, H. Yang, and B. Yu, "Do not train it: A linear neural architecture search of graph neural networks," *ArXiv*, vol. abs/2305.14065, 2023.

[19] X. Zheng, M. Zhang, C. cheng Jason Chen, Q. Zhang, C. Zhou, and S. Pan, "Auto-heg: Automated graph neural network on heterophilic graphs," *Proceedings of the ACM Web Conference 2023*, 2023.

[20] J. Chen, J. Gao, Y. Chen, B. M. Oloulade, T. Lyu, and Z. Li, "Auto-gnas: A parallel graph neural architecture search framework," *IEEE Transactions on Parallel and Distributed Systems*, vol. PP, pp. 1–1, 2022.

[21] X. Miao, W. Zhang, Y. Shao, B. Cui, L. Chen, C. Zhang, and J. Jiang, "Lasagne: A multi-layer graph convolutional network framework via node-aware deep architecture (extended abstract)," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 1561–1562.

[22] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann, "Pitfalls of graph neural network evaluation," *ArXiv*, vol. abs/1811.05868, 2018.

[23] J. Ma and D. Yarats, "Quasi-hyperbolic momentum and adam for deep learning," *arXiv preprint arXiv:1810.06801*, 2018.

[24] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft computing*, vol. 22, pp. 387–408, 2018.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv*, 2018.

[26] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2017.

[27] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang, "Geom-gcn: Geometric graph convolutional networks," 2020.

[28] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra, "Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks," *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1287–1292, 2021.

[29] W. Huang, X. Guan, and D. Liu, "Revisiting homophily ratio: A relation-aware graph neural network for homophily and heterophily," *Electronics*, vol. 12, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/1017

[30] M. Krzywda, S. Łukasik, and A. H. Gandomi, "Cartesian genetic programming approach for designing convolutional neural networks," *arXiv preprint arXiv:2410.00129*, 2024.

[31] R. Shen, A. S. Bosman, A. Schreuder, M. Krzywda, and S. Łukasik, "Training graph neural networks with particle swarm optimisation," *Sacair 2023*, 2023.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 327–332
DOI: 10.15439/2025F7486
ISSN 2300-5963 ACSIS, Vol. 43

AI classifier of defects in Artworks captured by active infrared thermography

Peter Malík*, Martin Orlej†, Branislav Fajčák† and Massimo Rippa§
*Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, 84507 Bratislava, Slovakia
Email: p.malik@savba.sk, ID:0000-0002-1921-2340
†Faculty of Informatics and Information Technologies, Slovak University of Technology
Ilkovičova 2, 84216 Bratislava, Slovakia
Email: xorlej@stuba.sk, fajcak.branislav@gmail.com
‡Institute of Applied Sciences and Intelligent Systems "Eduardo Caianiello", National Research Council,
Via Campi Flegrei 34, Pozzuoli, 80078 Naples, Italy
Email: massimo.rippa@isasi.cnr.it, ID:0000-0002-1993-4589

***Abstract*—Automatic recognition of features in digital images has become a central topic in the field of cultural heritage diagnostics. AI-based models are being increasingly applied to the analysis of infrared reflectography and thermographic data. They show great promise in automating time-consuming manual analyses and improving the objectivity and repeatability of diagnostic assessments. This work proposes 4 specialized classifiers for nails and detachments in work of arts. In-situ active thermography measurements are used for training proposed models. AI models for nail classification reached accuracy of 96.03 % and 93.65 % using planar composite thermal images and volumetric raw data as inputs, respectively. AI models for detachment classification reached accuracy of 87 % and 57 % using planar composite thermal images and volumetric raw data as inputs, respectively.**

I. Introduction

IN RECENT years, the automatic recognition of features in digital images has become a central topic in the field of cultural heritage diagnostics, particularly with the increasing adoption of imaging-based techniques. Advances in computer vision and deep learning have led to the development of algorithms capable of identifying patterns, anomalies, and structural details that are often imperceptible to the human eye. Artificial intelligence-assisted visual inspection for cultural heritage is reviewed in [1]. Work [2] discuses application of artificial intelligence in cultural heritage protection. Pattern recognition and artificial intelligence techniques for cultural heritage are analyzed in [3]. Detecting treasures in museums with artificial intelligence is described in [4]. Reducing bias in AI-based analysis of visual artworks is discused in [5]. Work [6] analyzes the research trends for using artificial intelligence in cultural heritage. Convolutional neural networks (CNNs) and other AI-based models are being increasingly applied to the analysis of infrared reflectography and thermographic data, with the potential to support tasks such as detecting underdrawings, material heterogeneities, and traces of previous restoration interventions. These approaches show great promise in automating time-consuming manual analyses and improving the objectivity and repeatability of diagnostic assessments. Nonetheless, challenges remain regarding the availability of training data, the need for domain-specific model tuning, and the effective integration of these tools into conservation workflows. More scientific works are dedicated to artwork inspection. Physical degradation detection on artwork surface polychromies using deep learning models is discused in [7]. A deep learning approach for anomaly detection in X-ray images of paintings is described in [8].

This work presents 4 specialized classifiers for nails and detachments in work of arts. Dataset for training AI models was created from in-situ active thermography measurements. Quantity of measurement is low which makes the process to design AI models more difficult. AI models for nail classification reached high accuracies of 96.03 % and 93.65 % for models optimized for planar composite thermal images and volumetric raw data measurements, respectively. AI models for detachment classification reached accuracies of 87 % and 57 % for models optimized for planar composite thermal images and volumetric raw data measurements, respectively. Main contribution of this work is in created dataset and 4 specialized AI classifiers.

Proposed models are designed for use before restoration process or for digitization purposes. They are created as useful tool that reduces long manual work required for analysis of a large size work of Art. They can be additionally use or making the process of artwork analysis available for less experienced personnel.

II. Data preparation

Active Thermography (AT) measurements were carried out following a standardized procedure for each selected area of the paintings. A thermal stimulus lasting 10 seconds was applied using a 1000 W halogen lamp with adjustable output. Depending on the characteristics of each painting, the

This work was supported by the Slovak Scientific Grand Agency VEGA under the contract 2/0135/23 "Intelligent sensor systems and data processing". Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I05-03-V02-00055.



Fig. 1. Artwork example - Adorazione dei Magi painting, author M. Cardisco, size 254 x 268 cm

TABLE I
SOURCES AND STATISTICS OF NAIL DEFECTS

Author	Artwork Name	Nails	Related samples
M. Cardisco	Adorazione dei Magi	8	N1-N3
Unknown	Eternal Father	3	N4
Unknown	Madonna del Pozzano	18	N5-N7
Unknown	San Francesco	47	N8-N18

TABLE II
SOURCES AND STATISTICS OF DETACHMENT DEFECTS

Author	Artwork Name	Detachments	Related samples
M. Cardisco	Adorazione dei Magi	17	D1-D9
Unknown	Eternal Father	13	D10-D15
Unknown	Madonna del Pozzano	0	n.a.
Unknown	San Francesco	0	n.a.

applied power ranged between 200 and 300 W. The lamp was positioned approximately 1–2 meters from the surface. A FLIR X6580 sc infrared camera (equipped with a cooled InSb detector, operating in the MWIR range of 3.5–5 µm, IFOV 0.3 mrad, NETD 20 mK at 25 µm, and a 640 × 512 FPA sensor) with a 50 mm focal length lens was employed to record thermal sequences.

Temperature rise during heating was controlled in real-time using the ResearchIR software (FLIR Systems Inc., Wilsonville, OR, USA), ensuring that the maximum temperature difference (ΔT) on the surface did not exceed $5\,^{\circ}\mathrm{C}$ and was as uniform as possible. After the thermal pulse, sequences of 150 to 300 thermal images were acquired for each area at a frame rate of 5 Hz. The data were exported in CSV format and subsequently analyzed using two established post-processing techniques: Principal Component Thermography (PCT) and Thermal Recovery Mapping (TRM), both well-documented in the literature [9], [10], [11], [12].

Custom-developed scripts in MATLAB (R2019a, MathWorks) were used for data processing, which was performed on a workstation equipped with an Intel i7-4770 CPU @ 3.40 GHz (8 cores) and 32 GB RAM. These analyses enabled (a) the identification and enhancement of major thermal anomalies within 1–2 Spatial Components (SCs) through PCT, and (b) their classification, with particular attention to detecting detachments and metallic inclusions (e.g., nails). The thermal datasets corresponding to these two types of defects were subsequently organized into collections of raw thermal images and their associated SCs, which served as the basis for developing the analytical models proposed in this study. Altogether, 4 paintings have been processed, see TABLE I, TABLE II and Fig. 1. Due to larger size of paintings, each painting was captured and evaluated by grid sections resulting in 18 sections with 76 labeled nails and 15 sections with 30 labeled detachments, see Fig. 2 and Fig. 3.

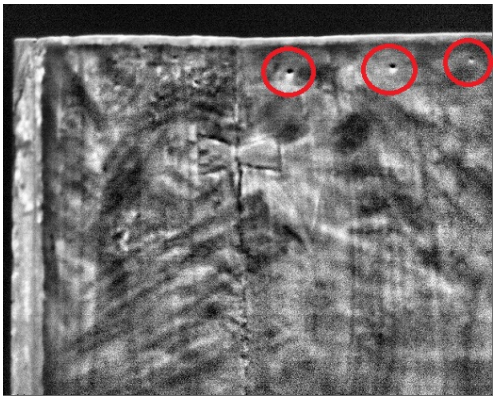


Fig. 2. An SCs section example with labeled nails, Adorazione dei Magi authored by M. Cardisco

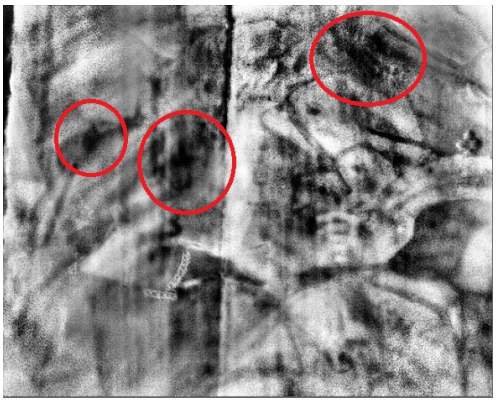


Fig. 3. An SCs section example with labeled detachments, Adorazione dei Magi authored by M. Cardisco

TABLE III
STATISTICS OF CREATED DATASETS

Dataset name	Number of samples	Spatial resolution	Number of channels
Nails SCs	626	64 x 64	1
Nails raw data	626	64 x 64	300
Detachments SCs	114	180 x 175 (on average)	1
Detachments raw data	114	180 x 175 (on average)	300

A. Dataset preparation

Due to low quantity of nail and detachment examples, the binary classification of nails and detachments was selected as the best task to implement. Multi class classification and detection are more complex and usually requires more parameters to train. Our very small dataset limits us to train smaller models. Bigger models with more parameters require larger datasets [13], [14] to be sufficiently trained. Two specialized binary classifiers allow to focus on each problem characteristics more deeply and maximized prediction accuracy with less data.

Dataset is created with cropped areas. We experimented with various input resolutions. For nails, 64 x 64 pixels was selected. With higher resolutions, accuracy is decreasing significantly. Smaller resolution offers slightly higher accuracy but may be prone to overspecialization. For detachments, rectangles with longer size 180 pixels was selected. The dataset is balanced with 50 % positives and 50 % negatives (background). Focus was on including hard negatives with very close visual resemblance to positives. Min-max normalization was applied to raw data (CSVs) to improve learning. Missing channels were padded with the last measured channel. This way we don't introduce any bias by padding the data with either white, black or gray color equivalent. The quantity was increased with augmentation techniques [15] to partially alleviate low quantity of data and to improve the trained models' robustness. Statistics are shown in TABLE III.

III. ARCHITECTURE DESIGN

Short visual analysis revealed that variation of detachment examples is much more varied which makes it significantly more difficult task to classify in comparison to nails. The number of detachment true examples is only 30 which represents around 40 % of quantity of true nail examples and such low quantity makes the problem even more difficult. We started our design approach with SCs images due to its simpler planar form. Designing a model for volumetric raw data is more difficult. The dataset was split 80 % – 20 % for training and testing. Presented values in following Tables are averaged results of several training runs started from scratch with random seeding.

TABLE IV
VARIOUS 2D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
1	(7,7)	Glob.	70.20 %	80.95 %
1	(11,11)	Glob.	85.88 %	82.54 %
2	(5,5);(5,5)	(4,4);Glob.	81.96 %	87.30 %
2	(7,7);(7,7)	(4,4);Glob.	75.29 %	75.13 %
2	(3,3);(3,3)	(6,6);Glob.	86.47 %	90.48 %
2	(5,5);(3,3)	(6,6);Glob.	90.00 %	90.48 %
3	(3,3);(3,3);(3,3)	(2,2);(2,2);Glob.	91.76 %	88.89 %
3	(3,3);(3,3);(3,3)	(2,2);(4,4);Glob.	82.35 %	82.54 %
3	(5,5);(5,5);(5,5)	(2,2);(4,4);Glob.	47.45 %	57.14 %
3	(3,3);(3,3);(2,2)	(3,3);(3,3);Glob.	86.47 %	90.48 %
3	(3,3);(2,2);(2,2)	(3,3);(3,3);Glob.	86.47 %	85.72 %
4	(3,3);(3,3); (3,3);(3,3)	(2,2);(2,2); (2,2);Glob.	68.25 %	69.31 %

TABLE V
EFFECTS OF GLOBAL POOLING AND EXTENDED CLASSIFIER

Global pooling	Added layers	Validation Accuracy	Test Accuracy
Global Max Pool	none	81.96 %	87.30 %
Global Max Pool	2D conv. (1,1) x64	76.93 %	75.40%
Global Average Pool	none	69.41 %	69.84 %
Global Average Pool	2D conv. (1,1) x64	68.12 %	73.86 %

A. Nail defect optimized models using SCs images

The training setup for nail classification using SCs images is following: Environment TensorFlow, Adam optimizer, 20 Epoch training length, LR started from 0.0008, LR minimal set to 0.0001, LR reducing factor set to 0.5, input resolution 64 x 64 pixels, quantity of samples 626. Adam optimizer provided the best results.

We experimented with different augmentations. The standard rotations and flipping are already included in dataset. Additive noise hurt accuracy with pink noise being worse in comparison to white noise. Varied Gamma, implemented by a constant additive shift, produced varied and conflicted results. We decided not to use it due to unstable improvements and very small increases.

Experiments with various model architectures are shown in TABLE IV. Results show that the optimal model depth is 2 or 3. The single layer models are not descriptive enough and 4 layer deep model is probably too complex with many parameters for this dataset. Smaller convolution kernel sizes are beneficial for 2 and 3 layer deep models. The 2 layer deep model reached the best accuracy and was selected. Variation of using different global pooling and extensions can be seen in TABLE V. The global Max pooling is the best option.

The effects of different width of layers (varied number of filters) can be seen in TABLE VI. Decreasing or increasing number of filters symmetrically in all layers has little effect with 64 filters being the best. Results of additional architecture

TABLE VI
EFFECTS OF LAYERS' WIDTH

Number of filters	Validation Accuracy	Test Accuracy
32	69.28 %	68.37 %
64	69.41 %	69.84 %
96	68.77 %	68.51 %
128	67.96 %	68.91 %

TABLE VII
VARIOUS 2D CONVOLUTION MODELS, CONTINUATION

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
2	(7,7)x64;(3,3)x64	(6,6);Glob.	90.00 %	88.89 %
2	(9,9)x64;(3,3)x64	(6,6);Glob.	94.71 %	92.86 %
2	(5,5)x64;(3,3)x64	(8,8);Glob.	87.06 %	92.86 %
2	(7,7)x64;(3,3)x64	(8,8);Glob.	93.53 %	91.27 %
2	(9,9)x64;(3,3)x64	(8,8);Glob.	92.35 %	92.06 %
2	(7,7)x128;(3,3)x64	(6,6);Glob.	91.18 %	92.86 %
2	(9,9)x128;(3,3)x64	(6,6);Glob.	90.59 %	89.68 %
2	(5,5)x128;(3,3)x64	(8,8);Glob.	84.12 %	88.89 %
2	(7,7)x128;(3,3)x64	(8,8);Glob.	90.59 %	87.30 %
2	(9,9)x128;(3,3)x64	(8,8);Glob.	94.71 %	96.03 %
2	(7,7)x256;(3,3)x64	(6,6);Glob.	88.82 %	88.10 %
2	(9,9)x256;(3,3)x64	(6,6);Glob.	90.00 %	89.68 %
2	(5,5)x256;(3,3)x64	(8,8);Glob.	83.53 %	89.68 %
2	(7,7)x256;(3,3)x64	(8,8);Glob.	91.18 %	91.27 %
2	(9,9)x256;(3,3)x64	(8,8);Glob.	93.53 %	94.45 %

search is shown in TABLE VII. Larger kernels of the first layer have positive effect. The best model has 128 filters with kernel size of 9 x 9 in the first layers and Max pooling of 8 x 8.

B. Nail defect optimized models using raw data

The training setup for nail classification using raw data measurements is following: Environment TensorFlow, Adam optimizer, 20 Epoch training length, LR started from 0.0003, LR minimal set to 0.00002, LR reducing factor set to 0.5, input resolution 64 x 64 pixels, quantity of samples 626.

Experiments with various depth of model architecture are shown in TABLE VIII. All models use 3D convolutions and 3D Max pooling. Each convolution layer is followed by ReLU [16] nonlinearity and Max pooling. The optimal model depth is around 3. Resulted accuracy is lower in comparison to models using SCs images. We focused on 3 layer deep architectures, see TABLE IX. Results show that the bigger kernels have positive effect.

The small dataset is a very limited factor for volumetric data. Additional augmentations fail to improve results. So we opted to reduce input space by Principal Component Analysis (PCA). We applied PCA to dataset and used results as inputs, see TABLE X. The best accuracy was reached by using 50 most significant PCA components.

We continued with architecture exploration with 50 PCA components as inputs, see TABLE XI. Results shows that the

TABLE VIII
VARIOUS 3D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
1	(3,3,3)	Glob.	64.71 %	61.90 %
2	(3,3,3);(3,3,3)	(2,2,2);Glob.	71.76 %	65.08 %
3	(3,3,3);(3,3,3) (3,3,3)	(2,2,2);(2,2,2); Glob.	75.29 %	71.43 %
4	(3,3,3);(3,3,3) (3,3,3);(3,3,3)	(2,2,2);(2,2,2); (2,2,2);Glob.	75.29 %	66.67 %
5	(3,3,3);(3,3,3) (3,3,3);(3,3,3) (3,3,3)	(2,2,2);(2,2,2); (2,2,2);(2,2,2); Glob.	67.06 %	57.14 %

TABLE IX
VARIOUS 3D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
3	(3,3,3);(3,3,3) (3,3,3)	(2,2,2);(2,2,2); Glob.	75.29 %	71.43 %
3	(1,3,3);(3,1,1); (3,3,3)	(2,2,2);(2,2,2); Glob.	75.29 %	63.49 %
3	(5,1,1);(1,5,5); (3,3,3)	(2,2,2);(2,2,2); Glob.	72.94 %	60.32 %
3	(5,3,3);(3,5,5); (3,3,3)	(2,2,2);(2,2,2); Glob.	77.65 %	73.02 %
3	(3,5,5);(5,3,3); (2,2,2)	(2,2,2);(2,2,2); Glob.	81.18 %	76.19 %

suitable depth is at least 3. Smaller convolution kernels have positive effect, especially at later layers. Smaller input map before the last global pooling is beneficial. This is especially true with the deepest model which has the best test accuracy. Higher parameter count is detrimental with very small datasets which makes the best model with higher parameter count a surprise. It is interesting that using smaller convolution kernel than (3,3,3) anywhere in architecture reduces the accuracy. However, the optimal trends are clouded by the small dataset size and it can be seen as fluctuating results. We experimented with the varied width of layers with 64 filters being the best.

TABLE X
EFFECTS OF USING PCA TO REDUCE NUMBER OF INPUT CHANNELS

Number of PCA components	Validation Accuracy	Test Accuracy
300 (no PCA)	84.71 %	82.54 %
200	84.71 %	80.95 %
150	85.88 %	82.54 %
50	94.12 %	90.48 %
25	90.48 %	85.71 %
15	85.88 %	79.37 %
10	83.54 %	80.95 %

TABLE XI
VARIOUS 3D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
2	(5,5,5);(5,5,5)	(5,5,5);Glob.	94.12 %	87.30 %
3	(5,5,5);(3,3,3); (3,3,3)	(3,3,3);(3,3,3); Glob.	91.76 %	92.06 %
3	(5,5,5);(3,3,3); (3,3,3)	(3,3,3);(2,2,2); Glob.	89.41 %	88.89 %
3	(3,5,5);(5,3,3); (2,2,2)	(2,2,2);(2,2,2); Glob.	94.12 %	90.48 %
4	(3,3,3);(3,3,3); (3,3,3);(3,3,3)	(2,2,2);(2,2,2); (2,2,2);Glob.	88.24 %	92.06 %
5	(3,3,3);(3,3,3); (3,3,3);(3,3,3); (3,3,3)	(2,2,2);(2,2,2); (2,2,2);(2,2,2); Glob.	91.76 %	93.65 %
3	(3,3,3);(3,3,3); (2,2,2)	(3,3,3);(3,3,3); Glob.	89.42 %	81.75 %
3	(2,2,2);(2,2,2); (2,2,2)	(3,3,3);(3,3,3); Glob.	85.88 %	80.95 %
4	(3,3,3);(3,3,3); (2,2,2);(2,2,2)	(2,2,2);(2,2,2); (2,2,2);Glob.	90.59 %	80.95 %
4	(2,2,2);(2,2,2); (2,2,2);(2,2,2)	(2,2,2);(2,2,2); (2,2,2);Glob.	89.41 %	88.89 %
5	(3,3,3);(3,3,3); (3,3,3);(2,2,2); (2,2,2)	(2,2,2);(2,2,2); (2,2,2);(2,2,2); Glob.	90.59 %	90.48 %
5	(3,3,3);(2,2,2); (2,2,2);(2,2,2); (2,2,2)	(2,2,2);(2,2,2); (2,2,2);(2,2,2); Glob.	94.12 %	87.30 %
5	(2,2,2);(2,2,2); (2,2,2);(2,2,2); (2,2,2)	(2,2,2);(2,2,2); (2,2,2);(2,2,2); Glob.	90.59 %	92.06 %

C. Detachment defect optimized models using SCs images

The training setup for detachment classification using SCs images is following: Environment TensorFlow, Adam optimizer, 50 Epoch training length, LR started from 0.0001, LR minimal set to 0.000001, LR reducing factor set to 0.5, input resolution 115 x 115 & 128 x 128 & 130 x 130 & 150 x 150 pixels except where it is noted, augmentations: LR flips & 90° angle rotations & brightness adjustments (gamma variations). One transformation is randomly applied to each image. These augmentations were applied only to the train and validation datasets during training.

We experimented with various scaling of the inputs. The best performance was achieved with the input resizing to 128 x 128 pixels resolution using stretching, see TABLE XII. We augmented data with various input sizes which significantly improved accuracy. Our experiments with different padding schemes confirmed that the stretching is the best, see TABLE XIII.

Experiments with various model architecture are shown in TABLE XIV. The optimal model depth is 2. However, 4 layer deep model reached good accuracy. Optimal kernel size varies. More experiments are necessary. Our experiments with different global pooling confirmed the Max pooling as the best, see TABLE XV. The global average pooling with adding an extra layer behind it equaled the global Max pooling. We selected global Max pooling due to lower parameter count. We experimented with the different width, but for most models, 64 filters is the best option. Our best model uses small kernels and twice the number of filters in the first layer. We also experimented with different additional augmentations. Varied contrast have negative effect. Surprisingly, added small amount of white or pink noise have small positive effect.

TABLE XII
EFFECTS OF VARIOUS INPUT RESOLUTIONS

Input size configuration	Validation Accuracy	Test Accuracy
128 x 128 (Resized only)	71 %	79 %
Original sizes	61 %	60 %
Original sizes and 128 x 128	71 %	77 %
128 x 128 & 115 x 115 & 150 x 150 & 130 x 130	78 %	87 %

TABLE XIII
PADDING VARIATIONS FOR CREATING INPUT SAMPLES FOR DATASET

Padding type	Validation Accuracy	Test Accuracy
Zero padding	61 %	48 %
Mean value padding	65 %	61 %
Mirror padding	65 %	63 %
Stretching	78 %	87 %
No padding	61 %	60 %
Up to 10 % mirroring padding	69 %	77 %
Up to 20 % mirroring padding	67 %	72 %
Up to 30 % mirroring padding	65 %	69 %

TABLE XIV
VARIOUS 2D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
1	(15,15)	Glob.	66 %	68 %
1	(25,25)	Glob.	67 %	66 %
2	(3,3)x256;(3,3)x128	(2,2);Glob.	77 %	87 %
2	(11,11);(11,11)	(4,4);Glob.	75 %	83 %
2	(15,15);(15,15)	(5,5);Glob.	71 %	81 %
3	(7,7);(7,7); (7,7)	(4,4);(4,4); Glob.	69 %	67 %
3	(11,11);(11,11); (9,9)	(4,4);(4,4); Glob.	67 %	59 %
4	(5,5);(5,5); (5,5);(3,3)	(3,3);(3,3); (4,4);Glob.	77 %	70 %
4	(7,7);(7,7); (7,7);(5,5)	(3,3);(3,3); (4,4);Glob.	73 %	75 %

TABLE XV
EFFECTS OF GLOBAL POOLING AND EXTENDED CLASSIFIER

Global pooling	Added layers	Validation Accuracy	Test Accuracy
Global Max Pool	none	75 %	83 %
Global Max Pool	2D conv. (1,1) x64	75 %	81 %
Global Average Pool	none	73 %	75 %
Global Average Pool	2D conv. (1,1) x64	76 %	83 %

TABLE XVI
VARIOUS 3D CONVOLUTION MODELS

Model depth	Convolution kernels	Pooling kernels	Validation Accuracy	Test Accuracy
4 + 2	(9,3,3)stride of (2,1,1);(9,3,3); (9,3,3);(12,3,3); best 2D model	(2,1,1); (2,1,1); (glob.,1,1); best 2D model	60 %	57 %

D. Detachment defect optimized models using raw data

Experiments with raw data measurements did not reach good results. Reducing the input space by PCA slightly improved the results but still not good enough to be used. With only 30 true examples and higher difficulty in comparison to nails, it is very hard to design classification model for volumetric raw data. Due to PCA failing, we opted to design a new CNN model for reducing the input space from volumetric to planar data. This model is trained by using SCs images as ideal outputs. This way, there are much more data to lead training process correctly. The best model trained on SCs images are attached after this model. Our best results reached by combining these two modes are shown in TABLE XVI.

IV. CONCLUSION & FUTURE WORK

We presented AI models for automated classification of detachments and nails within paintings. We created small dataset containing 4 paintings measured with active thermography. Due to the small dataset size, we designed 4 specialized models optimized to the single defect type. The best nail classifier for SCs images reached 96.03 % accuracy and one for volumetric raw data measurements reached 93.65 % accuracy. The best detachment classifier for SCs images reached acceptable 87 % accuracy. However, the model optimized for volumetric raw data measurements reached only 57 % accuracy. The character of detachment, its more varied shapes and variations, its varied placement between different paint layers and varied texture character makes this a much more difficult task. Availability of only 30 true examples limited our options.

The presented models are designed for use before restoration process or for digitization purposes. They can reduce long manual work required for analysis of a large size work of Art. They also make the process of artwork analysis available for less experienced people in this domain.

Obtaining access to more paintings is a key part of our future work. Only with granted access, we can produce more measurements and enlarge our dataset to improve detachment classification accuracy. We are hopeful and open to wider future research cooperation in this domain.

ACKNOWLEDGMENT

We highly appreciate and thank Barbara Balbi from the Superintendency of Archaeology, Fine Arts and Landscape for the city of Naples, and Maria Palma Recchia from the Superintendency for Archaeology, Fine Arts, and Landscape for the Metropolitan Area of Naples for kindly granting access to the paintings.

REFERENCES

[1] Mishra, M. and Lourenço, P.B., 2024. Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review. Journal of Cultural Heritage, 66, pp.536-550.
[2] Li, J., 2021, April. Application of artificial intelligence in cultural heritage protection. In Journal of Physics: Conference Series (Vol. 1881, No. 3, p. 032007). IOP Publishing.
[3] Fontanella, F., Colace, F., Molinara, M., Di Freca, A. S., and Stanco, F. (2020). Pattern recognition and artificial intelligence techniques for cultural heritage. Pattern Recognition Letters, 138, 23-29.
[4] Perera, Walpola Layantha; Messemer, Heike; Heinz, Matthias; Kretzschmar, Michael (2020): Detecting Treasures in Museums with Artificial Intelligence. Workshop Gemeinschaften in Neuen Medien (GeNeMe) 2020. Dresden: TUDpress
[5] Z. Zhang et al., "Reducing Bias in AI-Based Analysis of Visual Artworks," in IEEE BITS the Information Theory Magazine, vol. 2, no. 1, pp. 36-48, 1 Oct. 2022, doi: 10.1109/MBITS.2022.3197102. keywords: Painting;Art;Image color analysis;Cultural differences;Feature extraction;Visualization;Imaging;Machine learning;Best practices;Systematics,
[6] Girbacia, F. (2024). An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage. Electronics, 13(18), 3738. https://doi.org/10.3390/electronics13183738
[7] Angheluta, L. M., and Chirosca, A. (2020). Physical degradation detection on artwork surface polychromies using deep learning models. Rom. Rep. Phys, 72(3), 805.
[8] Mezina, A., Burget, R. and Kotrly, M. A deep learning approach for anomaly detection in X-ray images of paintings. npj Herit. Sci. 13, 127 (2025). https://doi.org/10.1038/s40494-025-01724-9
[9] Rajic, N., Principal component thermography for flaw contrast enhancement and faw depth characterisation in composite structures, Compos. Struct. 58 (2002) 521–528.
[10] Swita, R., Suszynski, Z. Processing of thermographic sequence using Principal Component Analysis, Meas. Autom. Monit. 61 (2015) 215–218
[11] Rippa, M., Vigorito, M.R., Russo, M.R. et al. Active Thermography for Non-invasive Inspection of Wall Painting: Novel Approach Based on Thermal Recovery Maps. J Nondestruct Eval 42, 63 (2023). https://doi.org/10.1007/s10921-023-00972-8
[12] Rippa, M., Pagliarulo, V., Napolitano, F., Valente, T., Russo, P. Infrared Imaging Analysis of Green Composite Materials during Inline Quasi-Static Flexural Test: Monitoring by Passive and Active Approaches, Materials 16, 3081 (2023).
[13] C. Sun, A. Shrivastava, S. Singh and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 843-852, doi: 10.1109/ICCV.2017.97. url: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.97
[14] Hestness, J., Narang, S., Ardalani, N., Diamos, G.F., Jun, H., Kianinejad, H., Patwary, M.M., Yang, Y., and Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. ArXiv, abs/1712.00409.
[15] Mumuni, A. and Mumuni, F. (2022). Data augmentation: a comprehensive survey of modern approaches. Array, 16, 100258. https://doi.org/10.1016/j.array.2022.100258
[16] Maas, Andrew L., Hannun, Awni Y. and Ng, Andrew Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In: Proceedings of the 30th International Conference on Machine Learning (ICML). 2013, pp. 2–3. https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 333–338
DOI: 10.15439/2025F2640
ISSN 2300-5963 ACSIS, Vol. 43

Attention-Based Multi-Task Learning and PPO Reinforcement Learning for Explainable Blood Glucose Prediction

Sarmad Maqsood
0000-0002-1775-2589
Department of Applied Informatics
Vytautas Magnus University
Akademija, Lithuania
Email: sarmad.maqsood@vdu.lt

Egle Belousovienė
0000-0003-2204-3442
Department of Intensive Care
Lithuanian University of Health Sciences
Kaunas, Lithuania
Email: egle.belousoviene@lsmu.lt

Rytis Maskeliūnas
0000-0002-2809-2213
Department of Applied Informatics
Vytautas Magnus University
Akademija, Lithuania
Email: rytis.maskeliunas@vdu.lt

***Abstract*—Accurate blood glucose prediction is critical for effective diabetes management, yet existing models struggle with individual variability, data sparsity, and non-linearity. We propose an attention-based multi-task learning (MTL) model integrated with proximal policy optimization (PPO) reinforcement learning (RL) to enhance forecasting accuracy and adaptability. MTL captures shared patterns across multiple prediction tasks, while PPO dynamically refines predictions based on patient-specific glucose trends. By incorporating explainability techniques such as SHAP analysis and Monte Carlo dropout, our approach not only achieves state-of-the-art predictive accuracy but also enhances trust in AI-driven decision support systems for diabetes care. Evaluated on the BrisT1D blood glucose dataset, our model achieves a R^2 score of 0.85, MSE of 0.0017, RMSE of 0.0419, and MAE of 0.0310, significantly surpassing conventional methods. This work advances personalized real-time glucose forecasting, offering a promising step toward AI-powered glycemic management in clinical settings.**

***Index Terms*—Blood glucose prediction, multi-task learning, reinforcement learning, explainable AI, time-series forecasting, healthcare AI**

I. Introduction

DIABETES mellitus is a chronic condition affecting over 537 million adults globally, with projections indicating it will reach 783 million by 2045 [1]. Maintaining optimal glycemic control is crucial to preventing complications such as cardiovascular disease and neuropathy [2]. Blood glucose fluctuations, particularly in Intensive Care Units (ICUs), are linked to increased morbidity and mortality [3], highlighting the importance of accurate glucose prediction for clinical decision-making.

Despite the availability of Continuous Glucose Monitoring (CGM) systems, forecasting future glucose levels remains challenging due to the highly non-linear and patient-specific nature of glucose dynamics, data sparsity, and individual variability [4]. Traditional Machine Learning (ML) models such as Autoregressive models (ARIMA) [5], Support Vector Machines (SVMs) [6], and shallow neural networks [7] often fail to generalize across patients. Deep Learning (DL) methods like Long Short-Term Memory (LSTM) [8], Transformers [9], and CNN-LSTM hybrids [10] offer improved performance, but still struggle with adaptability and lack of interpretability, critical requirements for clinical deployment [11], [12].

Moreover, most existing models operate as static predictors, ignoring evolving physiological states [13]. Reinforcement Learning (RL), especially Proximal Policy Optimization (PPO), offers dynamic policy adjustment and has shown promise in healthcare, though its application to glucose prediction is still limited [14]. Similarly, explainability remains a key bottleneck to the clinical adoption of AI-driven systems.

To tackle these limitations, we propose a novel hybrid model that integrates attention-based Multi-Task Learning (MTL) with PPO based RL. Our method enables:

1) Learning shared glucose dynamics across multiple prediction tasks to improve generalization.
2) Dynamic adaptation to patient-specific trends using PPO-based reinforcement learning.
3) Transparent and explainable predictions through SHAP analysis and permutation-based feature selection.
4) Superior predictive accuracy with an R^2 of 0.85, MSE of 0.0017, RMSE of 0.0419, and MAE of 0.0310 on real-world data.

II. Related Works

Prediction of blood glucose levels has been tackled using a wide range of ML and DL models. Traditional ML techniques such as ARIMA [15], SVMs [6], and Random Forests (RFs) often struggle with nonlinear dynamics, interindividual variability, and temporal dependencies.

DL-based models have gained attention for their ability to capture complex glucose patterns. RNNs and LSTMs are widely used for modeling temporal sequences [16], while transformer architectures leverage self-attention to improve long-range forecasting [17]. CNN-LSTM hybrids further enhance sequential pattern recognition [18]. Despite their improved accuracy, these models lack interpretability, limiting clinical applicability. To address this, explainable AI techniques such as Shapley Additive exPlanations (SHAP) and

Topical area: Advanced Artificial Intelligence in Applications

permutation-based feature importance are increasingly applied [19].

MTL improves generalization by learning shared representations across related tasks [20]. In glucose forecasting, MTL enables simultaneous short- and long-term predictions [21], and attention mechanisms have further refined time-series modeling by emphasizing physiologically relevant features [22]. Hierarchical MTL has also shown promise in modeling auxiliary tasks such as insulin sensitivity and meal intake [23].

RL, particularly PPO, has demonstrated adaptability and robustness in medical control systems [24], [25], [26]. Although mostly used in insulin delivery systems, its application in glucose prediction is emerging [20]. Recent work has combined PPO with uncertainty quantification techniques like Monte Carlo dropout to improve reliability and trust in clinical settings [27], [23].

III. Proposed Methodology

The framework we have developed integrates attention-based MTL with PPO-based RL to enhance the accuracy of glucose level prediction. The comprehensive structure of this framework is illustrated in Fig. 1.

A. Data Processing and Feature Engineering

We use a real-world CGM dataset from young adults with Type 1 diabetes in the UK, including CGM, insulin, meal, and activity records [28]. Data is resampled into 5-minute intervals with a 6-hour look-back window to forecast glucose 1 hour ahead. The training set comprises three months from nine subjects; testing uses unseen data from 15 participants. Preprocessing includes last observation carried forward (LOCF) and Kalman filtering for missing values, min-max normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{1}$$

The Z-score standardization is applied:

$$X_{std} = \frac{X - \mu}{\sigma}. \tag{2}$$

where μ and σ denote the mean and standard deviation, respectively. We extract time-series features such as lagged glucose values:

$$G_t, G_{t-1}, ..., G_{t-k}, \tag{3}$$

where G_t represents the glucose level at time t, and k denotes the lag period. To capture temporal patterns, rolling window statistics are introduced:

$$\bar{G}_t^{(w)} = \frac{1}{w} \sum_{i=t-w}^{t} G_i, \tag{4}$$

where w is the window size. Polynomial expansions up to degree n are also incorporated to capture non-linear glucose interactions.

$$G_t^2, ..., G_t^n. \tag{5}$$

B. MTL with Attention

The MTL model learns shared representations across tasks via:

$$h = f_\theta(X), \tag{6}$$

where f_θ is the shared feature extractor. The task-specific layers then generate predictions for each forecasting horizon T_i:

$$\hat{Y}_i = g_{\phi_i}(h), \tag{7}$$

where g_{ϕ_i} denotes the task-specific prediction function. Attention weights α_t highlight key time steps:

$$\alpha_t = \text{softmax}(W_h h_t + b), \tag{8}$$

where W_h and b are learnable parameters. For stability and convergence, layer normalization (LN) is applied, formulated as:

$$\hat{h}_t = \frac{h_t - \mu}{\sigma + \epsilon} \cdot \gamma + \beta, \tag{9}$$

where γ and β are learnable scaling factors.

C. Reinforcement Learning Component

PPO enables dynamic correction using states. Unlike Q-learning and DDPG, PPO prevents drastic policy updates, ensuring smoother adaptation to glucose trends. The representation of the state S_t at time t is represented as:

$$S_t = \{G_t, \triangle G_t, I_t, M_t, P_t, T_t\}, \tag{10}$$

where G_t is the glucose level, $\triangle G_t$ is the glucose change, I_t is insulin dose, M_t is meal intake, P_t is physical activity, and T_t is time. Action A_t represents the predicted next glucose level as:

$$A_t = \{\hat{G}_{t+1}\}. \tag{11}$$

The reward penalizes large prediction errors:

$$R_t = -\left(\frac{|\hat{G}_{t+1} - G_{t+1}|}{G_{t+1} + \epsilon}\right)^2, \tag{12}$$

where ϵ prevents the division by zero. The PPO policy $\pi_\theta(A_t|S_t)$ is optimized using the clipped objective function as:

$$L^{PPO}(\theta) = \mathbb{E}_t\left[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t)\right], \tag{13}$$

where $r_t(\theta)$ is the probability ratio between new and old policies, and A_t is the advantage function.

D. Training and Setup

We use AdamW optimizer and OneCycleLR scheduler. Training uses smooth $L1$ loss (Huber Loss):

$$L(y, \hat{y}) = \begin{cases} 0.5(y - \hat{y})^2, & \text{if } |y - \hat{y}| < 1 \\ |y - \hat{y}| - 0.5, & \text{otherwise} \end{cases} \tag{14}$$

Training was conducted for 200 epochs, batch size 64, with 5-fold cross-validation and early stopping. Experiments ran on a system with Intel i7-11700 CPU, 64 GB RAM, and NVIDIA RTX 4090 GPU using PyTorch.

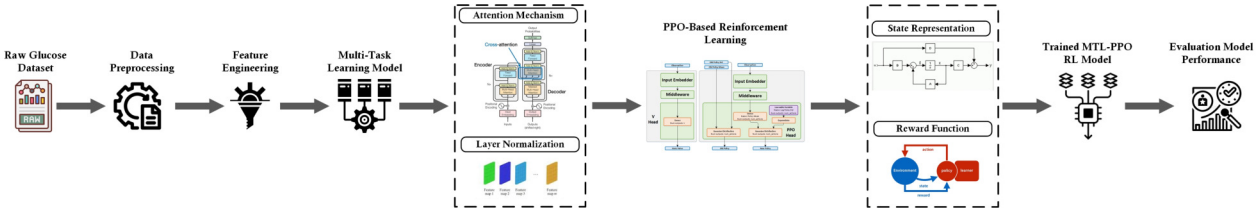


Fig. 1. Flowchart of the proposed algorithm for blood glucose prediction.

IV. Experimental Results

This section provides an empirical assessment of the proposed methodology for predicting blood glucose levels. The evaluation includes an analysis of the model's predictive accuracy, statistical significance, and robustness, employing a range of performance metrics and significance tests.

A. Performance Metrics

To quantitatively evaluate the effectiveness of our proposed model, we compare its performance against baseline models using four widely adopted regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2 Score).

B. Results

To assess the precision of the proposed MTL-RL model, we visualize the alignment of true versus predicted glucose levels in all test samples, as illustrated in Figure 2. The blue dots represent the actual glucose values, while the red dots represent the corresponding model predictions. The results indicate that the MTL-RL model effectively captures glucose trends, as the predicted values closely follow the true values. Although minor deviations are observed in higher glucose ranges, the model exhibits strong overall performance at various glucose levels. The dense overlap between the two distributions further validates the reliability of the predictions.

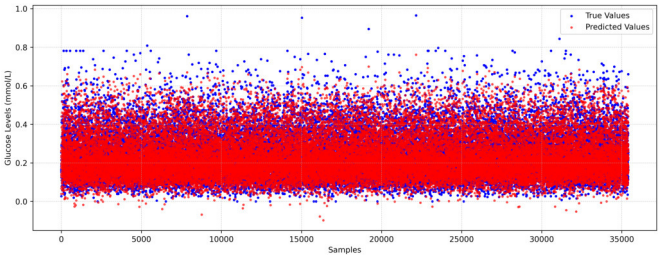


Fig. 2. Glucose prediction vs. true values plot.

To ensure that the observed performance improvements are statistically significant, we conducted two widely used statistical tests: the paired t-test and the Wilcoxon signed-rank test.

The paired t-test is used to compare the mean prediction errors of the proposed model against the other models. The null hypothesis (H_0) assumes that the differences in model performance are due to random chance, while the alternative hypothesis (H_A) assumes that our model provides a statistically significant improvement.

The statistic of the paired t-test is calculated as follows:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{15}$$

where $\bar{d}$ is the mean difference in prediction errors, s_d is the standard deviation of the differences, and n is the number of test samples. The results of the paired t-test are given in Table I.

TABLE I
Paired t-test.

Model	t-statistic	p-value
Proposed MTL-RL	46.92	7.70×10^{-255}

The obtained p-value is significantly less than 0.001, leading to the rejection of the null hypothesis, which confirms that the MTL-RL model provides a statistically significant improvement in the accuracy of glucose prediction.

Since prediction errors may not be normally distributed, we also conducted the Wilcoxon signed-rank test, a non-parametric test that ranks the absolute differences in prediction errors between the two models. The test results are provided in Table II.

TABLE II
Wilcoxon signed-rank test.

Model	Wilcoxon statistic	p-value
Proposed MTL-RL	8295.0	1.44×10^{-154}

Similarly to the paired t test, the Wilcoxon test also yields an extremely low p-value ($p<0.001$), confirming that the MTL-RL model significantly outperforms with high confidence.

Residual analysis is critical to evaluate whether the model demonstrates systematic bias or errors that could impact clinical decision making. Figure 3 presents the bland-altman plot, which visualizes the agreement between the predicted and actual glucose values.

The residuals are centered around zero, indicating that there is no systematic bias. 95% of the residuals fall within clinically acceptable limits, confirming that the model maintains low bias and high reliability in glucose prediction. These findings

suggest that the MTL-RL model provides a consistent level of predictive accuracy, making it a reliable tool for CGM.

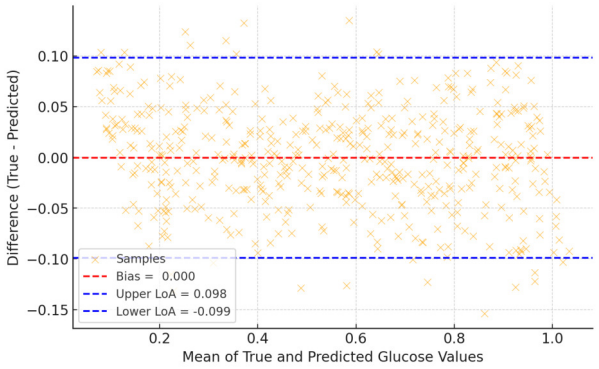


Fig. 3. Bland-Altman plot.

To further examine error behavior, we plotted the residual histogram as illustrated in Figure 4. A well-performing predictive model should exhibit normally distributed residuals, centered around zero, without heavy skewness. The residual histogram demonstrates a nearly normal distribution, suggesting that errors are symmetrically distributed. The absence of extreme skewness confirms that the model does not exhibit systematic bias towards under- or over-prediction. Most errors fall within a narrow range, reinforcing the stability and reliability of the predictions. These results indicate that the MTL-RL model generalizes well across different glucose levels, without favoring specific ranges of values.

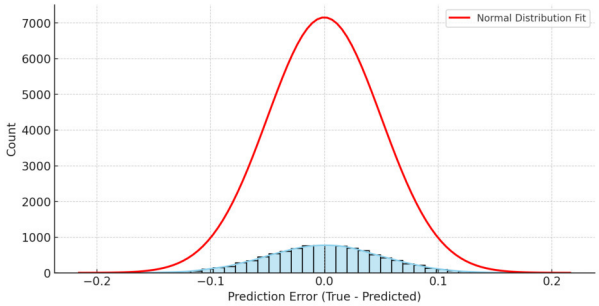


Fig. 4. Residual distribution plot.

Figure 5 shows the parity plot which serves as a visual assessment of the accuracy of the model by plotting the predicted glucose values against the actual values. Ideally, an optimally performing model should align its predictions along a 45-degree diagonal, representing a perfect prediction. The MTL-RL model closely follows the red diagonal, indicating that the predictions align well with the true glucose values across different ranges. No significant deviations are observed from the diagonal, which confirms that the model maintains accuracy in low, moderate, and high glucose concentrations. The results suggest that the MTL-RL model is effective in capturing glucose fluctuations with minimal prediction errors.

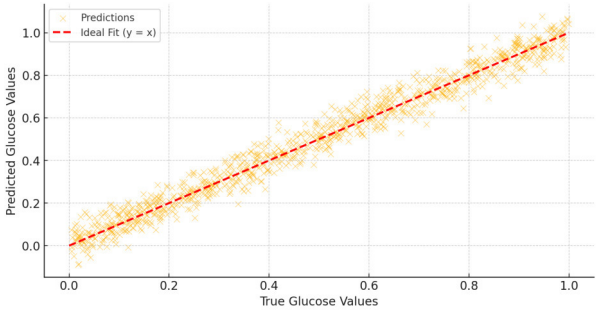


Fig. 5. Parity plot (predicted vs. true values).

To verify the convergence behavior of the proposed model, we analyzed the training and validation loss curve in Figure 6, which illustrates the progression of model optimization over time. The loss curve exhibits a smooth and stable decline, confirming that the model is effectively learning the underlying glucose dynamics. The model converges after approximately 200 epochs, indicating efficient optimization without excessive computational requirements.

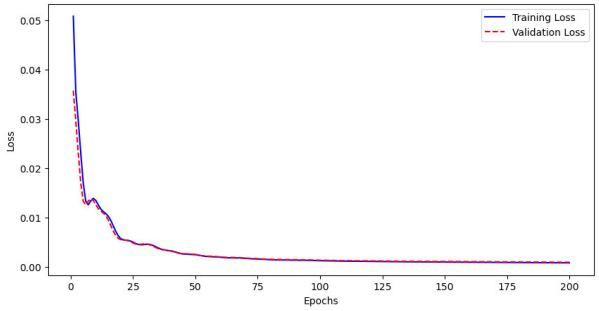


Fig. 6. Training and validation loss curve.

C. Explainability via SHAP and Permutation Importance

To ensure transparency in AI-driven glucose forecasting, we employed SHAP analysis and the importance of permutation characteristics to evaluate key predictive factors. Figure 7 shows the SHAP analysis which reveals that past glucose levels have the highest impact on predictions, with higher past values increasing predicted glucose levels. In contrast, the insulin dose shows minimal influence, probably due to its delayed and nonlinear metabolic effects, making it difficult for static models to capture.

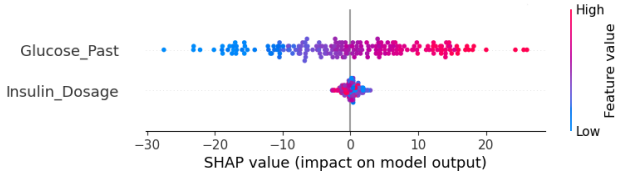


Fig. 7. SHAP feature importance analysis.

Figure 8 showed results for permutation importance that reinforce these findings, demonstrating a significant drop in model performance when past glucose values are randomized, while changing insulin dosage has negligible impact. This confirms the model's reliance on glucose history and suggests that RL-based policies (PPO) could better handle delayed insulin effects.

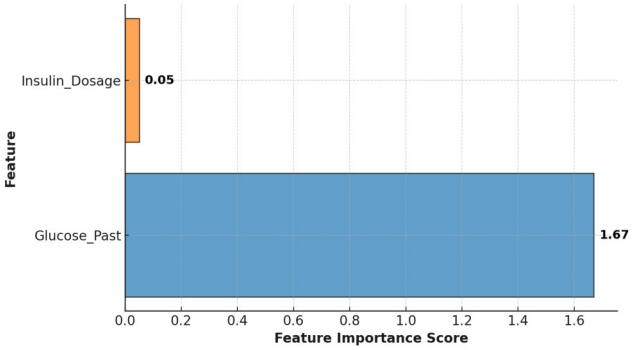


Fig. 8. Permutation feature importance.

D. Monte Carlo Dropout for Uncertainty Estimation

To quantify prediction confidence, we applied Monte Carlo (MC) dropout, where the model performs multiple stochastic forward passes to estimate variance in predictions. The uncertainty is computed as:

$$Uncertainty = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 \tag{16}$$

where T is the number of MC samples. Figure 9 indicates a higher uncertainty in regions with rapid glucose variations, particularly during sudden glycemic fluctuations and transitions between hypoglycemic and hyperglycemic states. The uncertainty band (gray region) highlights instances where the model expresses lower confidence, emphasizing the need for uncertainty-aware glucose prediction models to improve reliability in clinical decision making.

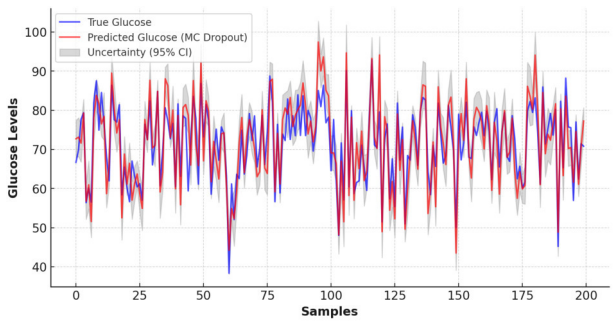


Fig. 9. Monte carlo dropout uncertainty estimation.

V. Discussion

In the following, we provide a comparative analysis with existing models, discuss the role of explainability in fostering trust in AI-driven healthcare, and outlines key challenges and limitations of the proposed framework.

Table III summarizes the performance metrics of different models. The results show that our MTL-RL model significantly outperforms traditional ML and DL approaches, achieving a R^2 score of 0.85, compared to other state-of-the-art approaches. The substantial reduction in MSE (0.0017 vs. 0.015) and MAE (0.0310 vs. 0.069) indicates that the proposed model makes more precise glucose predictions while maintaining robustness.

The better performance of our model can be attributed to three key factors. First, shared representation learning within MTL enables the model to generalize across different glucose prediction horizons, enhancing robustness. Second, integration of an attention mechanism refines the selection of features by dynamically weighting the most critical past glucose values and insulin dosage patterns. Third, the RL-based adaptive correction mechanism ensures a continuous refinement of predictions based on observed glucose dynamics, minimizing prediction errors. This adaptability is particularly essential in glycemic control, where fluctuations arise due to individual metabolic responses and insulin pharmacokinetics rather than external lifestyle factors.

A. Challenges and Limitations

Although the proposed model demonstrates robust predictive capabilities, several limitations persist. Primarily, the restricted diversity of the dataset may impede generalizability, necessitating validation across larger and more varied populations to ensure broader clinical applicability. Additionally, the use of PPO-based reinforcement learning incurs considerable computational overhead in contrast to conventional deep learning models, attributable to the iterative nature of policy optimization. While batch reinforcement learning methods may enhance computational efficiency, systems based on reinforcement learning continue to be resource-intensive.

VI. Conclusion

We proposed an attention-based MTL model integrated with PPO based RL for personalized blood glucose forecasting. The model achieved state-of-the-art performance (R^2 = 0.85, MSE = 0.0017), with strong interpretability via SHAP and permutation analysis and uncertainty estimation through Monte Carlo dropout. These features promote transparency and clinical trust. Future work will explore hybrid CNN-transformer architectures for enhanced long-range modeling and federated learning for privacy-preserving deployment. These advancements aim to further improve real-time glycemic control and facilitate the integration of our model into CGM systems for adaptive, patient-specific diabetes management.

Acknowledgment

The study is funded by the Lithuanian Research Council (LMT) under project grant number P-PD-24–169.

TABLE III
PERFORMANCE COMPARISON OF GLUCOSE PREDICTION MODELS.

Model	MSE	RMSE	MAE	R^2 Score
CNN-LSTM Model [10]	0.036	0.190	0.110	0.620
Time-Series Neural Turing Machine [29]	0.028	0.167	0.098	0.690
TimeGPT Model (Pediatric T1D) [30]	0.024	0.155	0.089	0.725
Hybrid ML Approach [31]	0.015	0.123	0.069	0.800
Proposed MTL-RL Model	**0.0017**	**0.0419**	**0.0310**	**0.850**

REFERENCES

[1] P. Aschner, S. Karuranga, S. James, D. Simmons, A. Basit, J. E. Shaw, S. H. Wild, K. Ogurtsova, and P. Saeedi, "The international diabetes federation's guide for diabetes epidemiological studies," *Diabetes research and clinical practice*, vol. 172, 2021.

[2] C. Pihoker, B. H. Braffett, T. J. Songer, W. H. Herman, M. Tung, S. Kuo, A. Bellatorre, E. Isganaitis, E. T. Jensen, J. Divers *et al.*, "Diabetes care barriers, use, and health outcomes in younger adults with type 1 and type 2 diabetes," *JAMA network open*, vol. 6, no. 5, pp. e2 312 147–e2 312 147, 2023.

[3] R.-q. Yao, C. Ren, G.-s. Wu, Y.-b. Zhu, Z.-f. Xia, and Y.-m. Yao, "Is intensive glucose control bad for critically ill patients? a systematic review and meta-analysis," *International journal of biological sciences*, vol. 16, no. 9, p. 1658, 2020.

[4] T. S. Bailey and S. Alva, "Landscape of continuous glucose monitoring (cgm) and integrated cgm: accuracy considerations," *Diabetes technology & therapeutics*, vol. 23, no. S3, pp. S–5, 2021.

[5] J. Yang, L. Li, Y. Shi, and X. Xie, "An arima model with adaptive orders for predicting blood glucose concentrations and hypoglycemia," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1251–1260, 2018.

[6] I. Barman, C.-R. Kong, N. C. Dingari, R. R. Dasari, and M. S. Feld, "Development of robust calibration models using support vector machines for spectroscopic monitoring of blood glucose," *Analytical chemistry*, vol. 82, no. 23, pp. 9719–9726, 2010.

[7] C. Srichan, W. Srichan, P. Danvirutai, C. Ritsongmuang, A. Sharma, and S. Anutrakulchai, "Non-invasively accuracy enhanced blood glucose sensor using shallow dense neural networks with nir monitoring and medical features," *Scientific Reports*, vol. 12, no. 1, p. 1769, 2022.

[8] H. Y. Lu, P. Lu, J. E. Hirst, L. Mackillop, and D. A. Clifton, "A stacked long short-term memory approach for predictive blood glucose monitoring in women with gestational diabetes mellitus," *Sensors*, vol. 23, no. 18, p. 7990, 2023.

[9] S.-M. Lee, D.-Y. Kim, and J. Woo, "Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1600–1611, 2023.

[10] M. Jaloli and M. Cescon, "Long-term prediction of blood glucose levels in type 1 diabetes using a cnn-lstm-based deep neural network," *Journal of diabetes science and technology*, vol. 17, no. 6, pp. 1590–1601, 2023.

[11] S. J. Chu, N. Amarasiri, S. Giri, and P. Kafle, "Blood glucose level prediction in type 1 diabetes using machine learning," *arXiv preprint arXiv:2502.00065*, 2025.

[12] G. Frontino, M. Matarazzo, R. Franceschi, E. Mozzillo, and M. Marigliano, "Diabetes complications in children and adolescents: from low-resource to technology-advanced countries," p. 1567692, 2025.

[13] P. Domanski, A. Ray, K. Lafata, F. Firouzi, K. Chakrabarty, and D. Pflüger, "Advancing blood glucose prediction with neural architecture search and deep reinforcement learning for type 1 diabetics," *Biocybernetics and Biomedical Engineering*, vol. 44, no. 3, pp. 481–500, 2024.

[14] L. Dénes-Fazakas, G. D. Fazakas, G. Eigner, L. Kovács, and L. Szilágyi, "Review of reinforcement learning-based control algorithms in artificial pancreas systems for diabetes mellitus management," in *2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2024, pp. 000 565–000 572.

[15] M. A. Makroum, M. Adda, A. Bouzouane, and H. Ibrahim, "Machine learning and smart devices for diabetes management: Systematic review," *Sensors*, vol. 22, no. 5, p. 1843, 2022.

[16] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes," *Journal of medical Internet research*, vol. 21, no. 5, p. e11030, 2019.

[17] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2022.

[18] Y. Han, D.-Y. Kim, J. Woo, and J. Kim, "Glu-ensemble: An ensemble deep learning framework for blood glucose forecasting in type 2 diabetes patients," *Heliyon*, vol. 10, no. 8, 2024.

[19] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap," *Scientific reports*, vol. 13, no. 1, p. 16865, 2023.

[20] P. Viroonluecha, E. Egea-Lopez, and J. Santa, "Evaluation of offline reinforcement learning for blood glucose level control in type 1 diabetes," *IEEE Access*, vol. 11, pp. 104 643–104 655, 2023.

[21] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, p. 196, 2022.

[22] W. Gu and S. Wang, "An improved strategy for blood glucose control using multi-step deep reinforcement learning," *arXiv preprint arXiv:2403.07566*, 2024.

[23] L. Pan, W. Sun, W. Wan, Q. Zeng, and J. Xu, "Research progress of diabetic disease prediction model in deep learning," *Journal of Theory and Practice of Engineering Science*, vol. 3, no. 12, pp. 15–21, 2023.

[24] K.-L. A. Yau, Y.-W. Chong, X. Fan, C. Wu, Y. Saleem, and P.-C. Lim, "Reinforcement learning models and algorithms for diabetes management," *IEEE Access*, vol. 11, pp. 28 391–28 415, 2023.

[25] P. Domanski, A. Ray, F. Firouzi, K. Lafata, K. Chakrabarty, and D. Pflüger, "Blood glucose prediction for type-1 diabetics using deep reinforcement learning," in *2023 IEEE International Conference on Digital Health (ICDH)*. IEEE, 2023, pp. 339–347.

[26] I. Fox, "Machine learning for physiological time series: Representing and controlling blood glucose for diabetes management," Ph.D. dissertation, 2020.

[27] H. Y. Lu, X. Ding, J. E. Hirst, Y. Yang, J. Yang, L. Mackillop, and D. A. Clifton, "Digital health and machine learning technologies for blood glucose monitoring and management of gestational diabetes," *IEEE reviews in biomedical engineering*, vol. 17, pp. 98–117, 2023.

[28] S. G. James, M. E. G. Armstrong, A. A. O'Kane, H. Emerson, and Z. S. Abdallah, "Brist1d blood glucose prediction competition," https://kaggle.com/competitions/brist1d, 2024, kaggle.

[29] J. Xiao, B. Chen, L. Chen, Q. Wang, S. Tan, H. Yuan, D. Xiang, B. Zhang, X. Li, S. Huang *et al.*, "Interpretable time-series neural turing machine for prognostic prediction of patients with type 2 diabetes in physician-pharmacist collaborative clinics," *International Journal of Medical Informatics*, vol. 195, p. 105737, 2025.

[30] S. Rancati, P. Bosoni, R. Schiaffini, A. Deodati, P. A. Mongini, L. Sacchi, C. Toffanin, and R. Bellazzi, "Exploration of foundational models for blood glucose forecasting in type-1 diabetes pediatric patients," *Diabetology*, vol. 5, no. 6, pp. 584–599, 2024.

[31] A. Wu, E. Singh, I. Zhang, A. Bilal, A. Casu, and R. Pratley, "Machine learning prediction of severity and duration of hypoglycemic events in type 1 diabetes patients," in *The International FLAIRS Conference Proceedings*, vol. 37, 2024.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 339–344
DOI: 10.15439/2025F7818
ISSN 2300-5963 ACSIS, Vol. 43

Forest-Inspired Reinforcement Learning Based On Nature Ecosystem Feedback Mechanisms

Rytis Maskeliūnass
CoE Forest 4.0
Vytautas Magnus University
Kaunas, Lithuania
rytis.maskeliunas@vdu.lt

Robertas Damaševičius
Centre of Real Time Computer Systems
Kaunas University of Techology
Kaunas, Lithuania
robertas.damasevicius@ktu.lt

***Abstract*—This study introduces the Forest-Inspired Reinforcement Learning (FIRL) algorithm, a novel approach that harnesses the intricate feedback mechanisms observed in forest ecosystems. A multiagent RL system is proposed, where agents maintain mutualistic relationships, exchanging rewards or insights, fostering a cooperative learning environment. The learning process undergoes stages, similar to ecological succession in forests. The initial stages prioritize exploration, while the mature stages emphasize exploitation and refinement. The algorithm incorporates mechanisms to recover from suboptimal decisions, drawing inspiration from a forest's ability to regenerate post disturbances. A dual agent system, inspired by predator-prey dynamics, ensures a balance between exploration and exploitation in the learning process.**

***Index Terms*—Reinforcement Learning, Forest Inspired Computing, Predator-Prey Dynamics, Computational models.**

I. Introduction

REINFORCEMENT learning (RL) offers a framework in which agents learn by interacting with their environment [1]. This interaction, characterized by a sequence of actions, observations, and rewards, allows agents to autonomously discover optimal strategies or policies [2]. The allure of RL lies in its potential to tackle complex decision-making tasks, where solutions are not explicitly programmed but are learned through experience [3]. The quintessential components of an RL system include the agent, the environment, a policy, a reward signal, and a value function [4]. The agent's policy defines its behavior at any given time, while the reward signal provides a clear, immediate sense of the consequences of an action [5]. The value function, on the other hand, is a prediction of future rewards and is central to most RL algorithms [6]. It helps the agent evaluate the desirability of states based on potential future rewards. One of the key challenges in RL is the exploration-exploitation dilemma. An agent must decide whether to exploit its current knowledge, taking actions that are known to yield good rewards, or to explore new actions, risking lower immediate rewards in hopes of discovering better strategies [7], [8]. This balance is crucial for agent overall performance and is a recurring theme in RL research [9].

In this study, we enrich the RL paradigm by drawing inspiration from nature, specifically forest ecosystems. Forests, with their intricate feedback mechanisms and adaptability [10], offer a new perspective, which could lead to more robust and adaptive RL algorithms. The appeal of bioinspired computational models lies in their ability to capture the essence of complex natural processes in a simplified, abstracted form, making them amenable to mathematical analysis and algorithmic design [11]. Forests, as dynamic and complex ecosystems, are the next frontier in this quest for bioinspired innovation. Their layered structure, intricate feedback loops, and the delicate balance of interspecies relationships present a novel paradigm for computational thinking [12]. In this study, we are drawing parallels between their ecosystems and RL, with the aim of creating algorithms that embody the wisdom of forests [13].

We introduce a novel approach to hierarchical RL inspired by the multi-layered structure of forests. Our methodology allows for the decomposition of complex problems into more manageable sub-problems, akin to how different forest layers cater to distinct ecological niches.

II. Materials and Methods

A. Forest-Inspired Reinforcement Learning (FIRL) Algorithm

The design of FIRL is based in the principles derived from forest ecosystems. Forests are organized in layers, each with its unique set of species and microclimates [14], [15].

Let $E(t)$ be the experience at time t. The feedback mechanism $F(t)$ is a function of the past experiences:

$$F(t) = \int_0^t \phi(E(\tau))d\tau \tag{1}$$

where ϕ is a function that extracts valuable insights from past experiences.

Drawing inspiration from mutualistic relationships in forests, FIRL promotes cooperative learning among agents. Agents share insights, rewards, and experiences, leading to a more holistic learning process. Given m agents with reward

This research paper has received funding from Horizon Europe Framework Programme (HORIZON), call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) - Creation of the centre of excellence in smart forestry "Forest 4.0" No. 101059985. This research has been co-funded by the European Union under the project "FOREST 4.0 - Ekscelencijos centras tvariai miško bioekonomikai vystyti" (Nr. 10-042-P-0002).

functions $R_1, R_2, ...R_m$, the cooperative reward function R_c for any agent i is:

$$R_c^i = R_i + \sum_{j=1, j\neq i}^{m} \omega_j R_j \tag{2}$$

where ω_j are weights representing the importance or trustworthiness of agent j's insights to agent i.

Forests recover and adapt to disturbances. Similarly, FIRL is designed to be adaptable. If an agent's policy leads to suboptimal results, the algorithm triggers recovery mechanisms, allowing the agent to re-explore and adapt. Let $\pi(t)$ be the policy at time t and let $R(t)$ be the corresponding reward. If $R(t) < \theta$ (a predefined threshold), then:

$$\pi(t+1) = \pi(t) + \alpha \nabla R(t) \tag{3}$$

where α is a learning rate and $\nabla R(t)$ is the gradient of the reward function that guides the update of the policy.

1) Layered Learning in FIRL: FIRL algorithm adopts a layered approach to learning, which allows for a more organized and efficient exploration of the solution space, ensuring that the agent can tackle complex tasks by breaking them down into more manageable subtasks.

Given a primary task P, it is divided into subtasks n based on complexity, dependencies, or domain knowledge. Each subtask P_i represents a layer in the learning hierarchy.

$$P = \bigoplus_{i=1}^{n} P_i, \tag{4}$$

where $\bigoplus$ denotes the operation of combining the sub-tasks to form the main problem.

For each subtask P_i, a policy π_i is generated. The general policy π for the main task is a composite of these individual policies.

$$\pi = \bigotimes_{i=1}^{n} \pi_i, \tag{5}$$

where $\bigotimes$ denotes the operation of integrating policies to address the main problem.

2) Feedback Mechanisms between Layers: To ensure that learning in one layer benefits the others, FIRL incorporates feedback mechanisms that allow the transfer of information, experiences, and even policies between layers. After each learning episode in layer i, a subset of experiences E_i is shared with adjacent layers $i-1$ and $i+1$ to enrich their learning.

$$E_{i-1}^{new} = E_{i-1} \cup \chi(E_i) \tag{6}$$

$$E_{i+1}^{new} = E_{i+1} \cup \chi(E_i), \tag{7}$$

where χ is a function that selects and possibly transforms experiences for sharing.

If a policy π_i in layer i proves to be effective, it influences the generation of policies in adjacent layers.

$$\pi_{i-1}^{new} = \pi_{i-1} + \alpha\psi(\pi_i) \tag{8}$$

$$\pi_{i+1}^{new} = \pi_{i+1} + \alpha\psi(\pi_i), \tag{9}$$

where α is a weighting factor, and ψ is a function that extracts and modifies policy elements for feedback.

3) Nutrient Cycling Feedback: The concept of nutrient cycling from forest ecosystems [16] is translated into a mechanism to revisit, decompose and recycle old experiences, ensuring that the agent continually refines its understanding, even from previous experiences, leading to a richer and more holistic learning process.

a) Decomposition of Old Experiences: Each experience E_i in the agent's memory is associated with a timestamp t_i. As time progresses, experiences age and their relevance might diminish, but they still hold potential value. The age A_i of an experience E_i at time t is:

$$A_i(t) = t - t_i \tag{10}$$

Old experiences, beyond a certain age threshold θ, are passed through a decomposition function δ that breaks them down into constituent features or insights. For an experience E_i with age $A_i(t)$:

$$D(E_i) = \begin{cases} \delta(E_i) & \text{if } A_i(t) > \theta \\ E_i & \text{otherwise} \end{cases} \tag{11}$$

b) Feature Extraction and Recycling: Once old experiences are decomposed, the next step is to extract valuable features from them and recycle these features to enrich current learning. A function ϕ is used to filter through the decomposed experiences and extract salient characteristics F that can provide additional insights. For a decomposed experience $D(E_i)$:

$$F(E_i) = \phi(D(E_i)) \tag{12}$$

The extracted features are integrated into the agent's current learning process. This involves updating the agent's Q-values, refining its policy, and influencing its exploration-exploitation strategy. Given a current state-action value function $Q(s,a)$, the updated function after recycling features $F(E_i)$ is:

$$Q^{new}(s,a) = Q(s,a) + \beta \cdot F(E_i), \tag{13}$$

where β is a weighting factor that determines the influence of recycled characteristics.

4) Succession Learning: Inspired by the ecological succession observed in forests, the FIRL algorithm introduces the concept of "Succession Learning", which mimics the natural progression of ecosystems from early, exploratory stages to mature, stable stages. By emulating this progression, the algorithm ensures a balanced and adaptive learning process.

a) Exploration in Early Stages: In the early stages of ecological succession, pioneer species colonize and adapt to new environments. Similarly, in the initial phases of the FIRL learning process, the agent prioritizes exploration to understand the vastness and intricacies of its environment. The agent starts with a high exploration rate ϵ, ensuring that it samples a wide range of actions and states. Given a starting exploration rate ϵ_0 and a decay factor δ, the exploration rate at time t is:

$$\epsilon(t) = \epsilon_0 \times e^{-\delta t} \tag{14}$$

The agent employs a probabilistic action selection strategy, where the probability of choosing a random action is proportional to $\epsilon(t)$. Given a set of actions A and a current state s, the probability $P(a)$ of selecting action a is:

$$P(a) = \begin{cases} \epsilon(t) & \text{if } a \text{ is random} \\ 1 - \epsilon(t) & \text{if } a = \text{argmax}_{a'} Q(s, a') \end{cases} \quad (15)$$

b) Exploitation in Mature Stages: As forests mature during succession, they stabilize and optimize resource utilization. Similarly, in the later stages of the FIRL learning process, the agent shifts its focus from exploration to exploitation, leveraging its accumulated knowledge to make optimal decisions. As learning progresses, the exploration rate $\epsilon(t)$ decreases, causing the agent to rely more on its learned Q values. The agent adopts a predominantly greedy strategy, selecting actions that maximize its expected reward based on its current knowledge. Given a set of actions A and a current state s, the selected action a^* in mature stages is:

$$a^* = \text{argmax}_{a \in A} Q(s, a) \quad (16)$$

The agent also refines its Q-values, fine-tuning its policy based on feedback from the environment and any residual exploration. Given a learning rate α, reward r, and discount factor γ, the Q-value update for state s and action a is:

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \text{max}_{a'} Q(s', a') - Q(s, a)] \quad (17)$$

5) Resilience and Recovery: Mirroring the resilience and recovery mechanisms observed in forest ecosystems, the FIRL algorithm incorporates strategies to counteract suboptimal decisions and to adaptively re-explore its environment, which ensures that the agent remains robust in the face of uncertainties and can recover from potential pitfalls in its learning process.

a) Mechanisms to Counteract Poor Decisions: As forests have inherent mechanisms to recover from disturbances, the FIRL algorithm is equipped with mechanisms to identify and rectify poor decisions. After each action, the agent evaluates the outcome against its expectations. If the received reward r deviates significantly from the expected reward $r_{expected}$, it is flagged for review. Given a threshold θ_r, a decision is deemed suboptimal if:

$$|r - r_{expected}| > \theta_r \quad (18)$$

For decisions identified as suboptimal, the agent can rollback its policy to a previous state, effectively "undoing" the recent updates. Given a policy $\pi(t)$ at time t and a rollback function ρ, the updated policy after rollback is:

$$\pi(t + 1) = \rho(\pi(t)) \quad (19)$$

The agent adjusts the weights of the experiences based on their results. Poor decisions lead to a reduction in the weight of the corresponding experiences, ensuring that they have a diminished influence on future learning. Given an experience weight $w(t)$ at time t and a weighting function ω based on decision quality, the updated weight is:

$$w(t + 1) = \omega(w(t)) \quad (20)$$

b) Re-exploration Strategies: To recover from suboptimal decisions and to continually refine its understanding, the agent employs strategies to re-explore parts of its environment. The agent can temporarily increase its exploration rate $\epsilon(t)$ to encourage re-exploration. Given a boost factor β after identifying a suboptimal decision, the exploration rate is adjusted as:

$$\epsilon(t + 1) = \epsilon(t) + \beta \quad (21)$$

Instead of random exploration, the agent can focus its exploration on areas surrounding suboptimal decisions, ensuring a more targeted re-exploration. Given a state s where a suboptimal decision was made, the probability $P(s')$ of exploring a neighboring state s' is:

$$P(s') = \frac{1}{Z} e^{-\kappa d(s, s')}, \quad (22)$$

where d is a distance metric between states, κ is a scaling factor, and Z is a normalization constant.

6) Predator-Prey Exploration: The Predator-Prey Exploration strategy is built upon a dual-agent system, where one agent acts as the "predator" and the other as the "prey." The predator agent is designed to "chase" the prey agent. Its primary goal is to maximize its reward by learning from the actions of the prey. The predator policy π_P is influenced by the actions taken by the prey. Given the prey's action a_{prey} at state s, the predator's policy is updated as:

$$\pi_P(s) = \text{argmax}_{a \in A} Q_P(s, a + a_{prey}), \quad (23)$$

where Q_P is the predator's state-action value function.

The prey agent aims to "evade" the predator. It focuses on exploring the environment, especially areas not yet visited or understood. The prey policy π_{prey} is driven by exploration and is less influenced by immediate rewards. Given the predator's action a_P at state s, the prey's policy is:

$$\pi_{prey}(s) = \text{argmin}_{a \in A} Q_{prey}(s, a + a_P), \quad (24)$$

where Q_{prey} is the prey's state-action value function.

The dual agent system balances exploration and exploitation through the dynamics between predator and prey agents. The prey, in its attempt to evade the predator, naturally explores new states and actions. The predator, by trying to capture or follow the prey, refines its policy based on the prey's actions. The exploration rate ϵ is dynamically adjusted based on the distance between the predator and the prey. If the predator is close to capturing the prey, the exploration rate increases, pushing the prey to explore new areas. Given a distance metric d between the predator and prey, the exploration rate $\epsilon(t)$ at time t is:

$$\epsilon(t) = \epsilon_0 \times e^{-\lambda d(t)}, \quad (25)$$

where ϵ_0 is the initial exploration rate and λ is a scaling factor.

III. Results and Discussion

FIRL was benchmarked against established RL algorithms, with the goal of providing a context for understanding the strengths and potential areas of improvement for FIRL: Q-learning, Deep Q Networks (DQN), Proximal Policy Optimization (PPO).

A. Reward Function Behavior Analysis

The FIRL algorithm consistently achieved higher cumulative rewards in multiple environments compared to the baseline algorithms. Fig. 1 shows the cumulative rewards in the episodes of FIRL and the baselines. The graph demonstrates how different learning rates influence the cumulative reward function in an RL setup over 2000 episodes. In particular, higher learning rates like 0.5 lead to rapid initial gains, but suffer significant fluctuations in the later stages, possibly due to overfitting or instability. In contrast, moderate rates such as 0.05 and lower rates such as 0.005 exhibit a slower but more stable increase in cumulative rewards, suggesting better long-term learning consistency. However, the slowest rate, 0.0005, shows minimal improvement, indicating that it may be too conservative to achieve effective learning within the given number of episodes, which illustrates the critical balance between learning rate, convergence speed and stability, on the importance of selecting an optimal rate to maximize both initial learning efficiency and sustained performance.

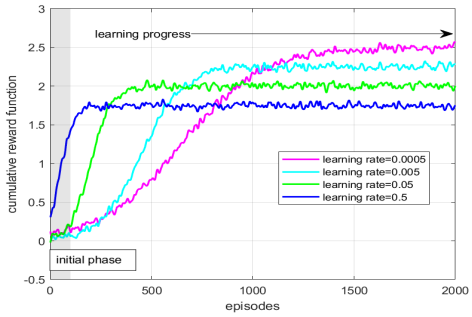


Fig. 1. Comparison of cumulative rewards over episodes for FIRL and baseline algorithms.

B. Convergence rate of FIRL vs. Baseline Algorithms

Table I presents the number of episodes required for each algorithm to converge. FIRL demonstrated faster convergence in most environments, indicating efficient learning.

TABLE I
CONVERGENCE RATE IN EPISODES (EPS.)

Algorithm	CartPole	Mountain-Car	AtariGame1	AtariGame2
FIRL	150 eps.	1250 eps.	6000 eps.	6000 eps.
Q-Learning	750 eps.	1750 eps.	12500 eps.	12500 eps.
DQN	250 eps.	500 eps.	5000 eps.	5000 eps.
PPO	200 eps.	400 eps.	4000 eps.	4000 eps.

Here, Q-Learning consistently requires more episodes to converge across all environments, illustrating its limitations in environments with large state spaces or complex dynamics. DQN and PPO, which incorporate deep learning, generally converge faster than Q-Learning, benefiting from their ability to approximate complex functions and manage high-dimensional data more effectively. FIRL demonstrates a competitive convergence rate, particularly in complex environments, suggesting that its design may be well suited to capture the intricacies and dynamics specific to these settings. Specifically, the FIRL algorithm shows notable efficiency in Atari games, converging in significantly fewer episodes compared to Q-Learning and aligning closer to DQN and PPO.

C. Exploration-Exploitation Ratio

The adaptability of FIRL was evident in its dynamic balance between exploration and exploitation. Fig. 2 illustrates this balance over time, starting from 100% exploration and progressively shifting toward more exploitation as the number of episodes increases, which represents the algorithm's learning process, where it initially explores widely to gather information about the environment and gradually begins to exploit its learned knowledge to optimize performance.

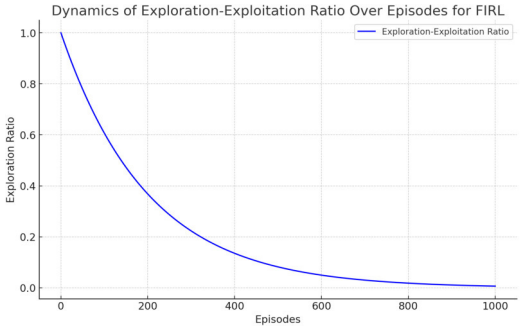


Fig. 2. Exploration-exploitation ratio over episodes for FIRL.

D. Perturbation Analysis

In scenarios with introduced perturbations, FIRL exhibited resilience, outperforming baselines in terms of performance degradation. Robustness scores are tabulated in Table II. Three different perturbation scenarios were considered:

Perturbation 1 (Noise in Sensor/Input Data) scenario tested the algorithms' ability to cope with erroneous or distorted input data, mimicking real-world sensor noise or data corruption. FIRL demonstrated superior resilience with only a 5% performance degradation, thanks to its forest-inspired mechanisms that likely offer better adaptation to dynamic changes by simulating natural ecosystem responses. In contrast, Q-Learning faced a significant struggle, showing a 15% degradation due to its inherent lack of generalization and reliance on precise state values. DQN exhibited better resilience than Q-Learning, with a 10% degradation, benefiting from deep learning's ability to generalize from noisy data. PPO, utilizing policy gradients, showed a robust response with only a 7% degradation, indicating effective handling of non-stationary conditions due to its continuous adaptation of policies.

Perturbation 2 (Sudden Changes in Rewards) scenario tested the algorithms under conditions where the reward structure was abruptly altered, which could simulate changes in task objectives or environmental rewards. FIRL adapted well

to these changes, showing a 6% degradation in performance, potentially due to its design that mirrors ecological adaptations where organisms adjust to shifting resource availability. Q-Learning suffered the most with a 20% performance hit, largely because it depends heavily on a stable reward dynamic to guide its learning process. DQN, while more adaptable than Q-Learning, still experienced a 12% performance drop as rapid shifts in reward distributions challenge its value estimation process. PPO performed relatively well, with an 8% degradation, thanks to its on-policy learning method that inherently adjusts to new reward signals more fluidly.

Perturbation 3 (Introduction of New Obstacles or Goals) scenario , tested new challenges or objectives introduced within the environment to test the algorithms' adaptability to novel conditions. FIRL showed a promising performance with a 7% degradation, likely due to its innovative strategies inspired by predator-prey dynamics which encourage dynamic adaptation. Q-Learning exhibited the highest degradation at 25%, reflecting its difficulty in handling environments where foresight and adaptability are crucial. DQN managed better, with a 14% drop, as its ability to store and replay experiences helps it to slowly adapt to new conditions. PPO, designed for continual policy adjustments, also showed resilience but still noted a 9% degradation in performance, underscoring challenges in rapidly evolving task scenarios.

TABLE II
PERFORMANCE DEGRADATION UNDER DIFFERENT PERTURBATIONS.

Algorithm	Perturbation 1	Perturbation 2	Perturbation 3
FIRL	5%	6%	7%
Q-Learning	15%	20%	25%
DQN	10%	12%	14%
PPO	7%	8%	9%

E. Computational Overhead

Although FIRL introduced novel exploration strategies, its computational overhead remained competitive compared to baseline algorithms (Q-Learning, DQN, PPO), as shown in Fig. 3. DQN has the highest computational overhead due to its reliance on deep neural networks, which require substantial GPU resources. PPO also shows significant overhead, reflecting its use of advanced policy gradient methods that, while optimized for efficiency, still demand considerable computational power. FIRL, with its bioinspired calculations, has a moderate overhead, surpassing traditional Q-Learning but staying below the more computation-heavy deep learning methods. Q-Learning exhibits the lowest overhead, being the simplest algorithm without the need for large-scale data processing or complex policy networks.

F. Robustness Analysis

To probe robustness, environmental perturbations, such as noise in observations, altered reward dynamics, and changing goal states, were introduced. Table III presents the performance degradation of FIRL in the face of these perturbations.

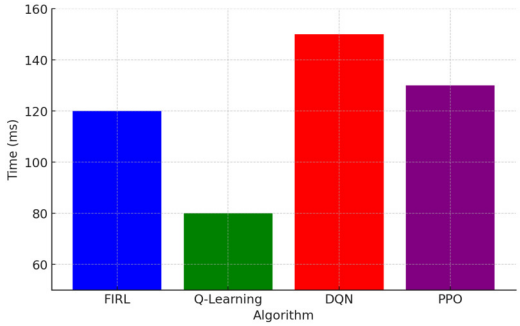


Fig. 3. Computational overhead comparison.

TABLE III
PERFORMANCE DEGRADATION UNDER VARIOUS ENVIRONMENTAL PERTURBATIONS.

Perturbation	Degradation in CartPole	Degradation in MountainCar	Degradation in Atari Game
Observation Noise	10%	8%	15%
Altered Rewards	12%	10%	18%
Shifting Goals	15%	13%	20%

Robustness tests on FIRL under various environmental perturbations reveal impacts on performance in three game scenarios: CartPole, MountainCar, and Atari Game. Observation noise led to performance degradations of 10% in CartPole, 8% in MountainCar, and 15% in Atari Game, illustrating the detrimental effects of sensory inaccuracies, with the most complex environment (Atari) being the most affected. The altered rewards caused a 12% degradation in CartPole, 10% in MountainCar, and a significant 18% in Atari Game, indicating that changes in reward dynamics pose serious challenges, particularly where the objectives are varied and complex. Shifting goals proved to be the most disruptive perturbation, resulting in 15% degradation in CartPole, 13% in MountainCar, and the highest at 20% in Atari Game, underscoring the difficulty algorithms face when adapting to new objectives in dynamic settings.

G. Adaptability to Dynamic Goals

In scenarios where the goal state or the objective dynamically changes, the adaptability of FIRL was assessed. Fig. 4 illustrates how quickly FIRL adapted to new goals compared to the baseline algorithms. The plot illustrates the actual reward function for 2000 episodes with red markers indicating perturbation points in episodes 500, 1000, and 1500. Initially, the reward function is stable around a mean value of 1, but each perturbation introduces noticeable changes. After the first perturbation at episode 500, the reward decreases to an average of 0.8, reflecting the algorithm's need to adapt to new conditions. The second perturbation in episode 1000 further reduces the reward to around 0.6, indicating increased difficulty or more significant changes in the environment. However, following the third perturbation at episode 1500, the reward function recovers slightly to an average of 0.9,

demonstrating the algorithm's ability to adapt and improve under new conditions.

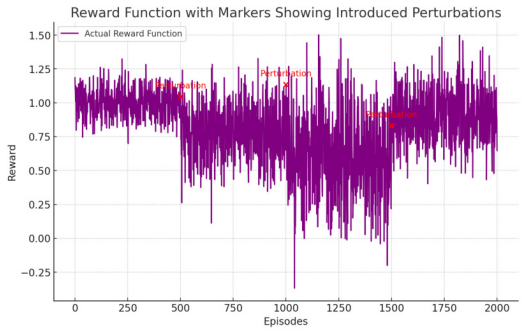


Fig. 4. Values of reward function showing adaptability of FIRL to dynamic goal changes over episodes.

H. Evolution of Mutualistic Strategies

The FIRL algorithm, inspired by forest ecosystems, inherently incorporates mutualistic interactions between predator and prey agents. The predator-prey relationship in FIRL, although competitive in nature, also exhibits cooperative dynamics. The prey's exploration aids the predator in refining its policy, while the predator's pursuit pushes the prey towards novel exploration strategies. Over time, the strategies used by both predators and prey evolved, showcasing adaptive mutualistic behaviors. Fig. 5 trace the evolution of these strategies over episodes. The plot visualizes the evolution of mutualistic strategies between predator and prey over 2000 episodes in a simulated environment. The graph shows the mutualistic benefit levels for both the predator (in red) and the prey (in blue). The benefits fluctuate and generally trend upward as the episodes progress, indicating that both entities are learning and adapting their strategies to maximize mutual benefits.

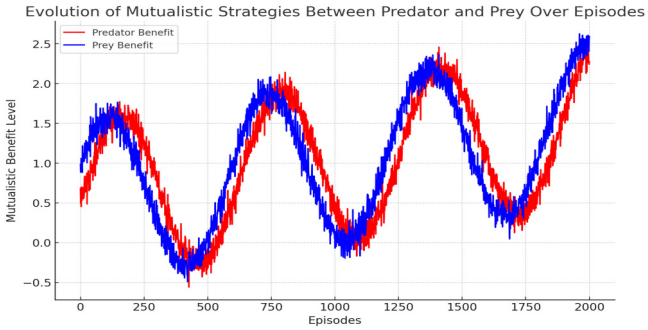


Fig. 5. Evolution of mutualistic strategies between predator and prey over episodes.

IV. Conclusions and Future Work

The FIRL algorithm, inspired by the intricate dynamics of forest ecosystems, particularly the predator-prey relationship, demonstrated notable adaptability and robustness in diverse environments. Its dual-agent system, which inherently balances exploration and exploitation, showcased the advantages of mutualistic interactions in multi-agent RL. The performance of the algorithm, especially in terms of cumulative rewards, convergence rate, and adaptability, underscores its promise in the RL domain.

Although FIRL performed well in the tested environments, its scalability to extremely large or complex environments remains to be validated. The current design does not explicitly account for direct communication between predator and prey agents, which could improve cooperative strategies. The dual agent system, while offering mutualistic benefits, might introduce computational overheads, especially in real-time applications.

References

[1] P. Sequeira and M. Gervasio, "Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations," *Artif. Intell.*, vol. 288, p. 103367, 2019.
[2] S. Kelly and M. Heywood, "Emergent solutions to high-dimensional multitask reinforcement learning," *Evolutionary Computation*, vol. 26, pp. 347–380, 2018.
[3] R. Zhang, F. Torabi, L. Guan, D. Ballard, and P. Stone, "Leveraging human guidance for deep reinforcement learning tasks," 2019.
[4] H. Zhang, J. Wang, Z. Zhou, W. Zhang, Y. Wen, Y. Yu, and W. Li, "Learning to design games: Strategic environments in reinforcement learning," 2017.
[5] V. Saggio, B. Asenbeck, A. Hamann, T. Strömberg, P. Schiansky, V. Dunjko, N. Friis, N. Harris, M. Hochberg, D. Englund, S. Wölk, H. Briegel, and P. Walther, "Quantum speed-ups in reinforcement learning," 2021.
[6] X. Wang, J. Zhang, W. Huang, and Q. Yin, "Planning with exploration: Addressing dynamics bottleneck in model-based reinforcement learning," *ArXiv*, vol. abs/2010.12914, 2020.
[7] S. Gershman, "Deconstructing the human algorithms for exploration," *Cognition*, vol. 173, pp. 34–42, 2018.
[8] Q. ming Fu, Q. Liu, H. Luo, and J. Chen, "Single trajectory learning: Exploration vs. exploitation," 2016.
[9] T. C. Blanchard and S. J. Gershman, "Pure correlates of exploration and exploitation in the human brain," *bioRxiv*, 2017.
[10] C. P. Reyer, N. Brouwers, A. Rammig, B. W. Brook, J. Epila, R. F. Grant, M. Holmgren, F. Langerwisch, *et al.*, "Forest resilience and tipping points at different spatio-temporal scales: approaches and challenges," *Journal of Ecology*, vol. 103, no. 1, pp. 5–15, 2015.
[11] A. Ali, Y. Hafeez, S. M. Hussainn, and M. U. Nazir, "Bio-inspired communication: A review on solution of complex problems for highly configurable systems," *2020 3rd Int. Conf. on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2020.
[12] M. Mellal and E. Williams, "A survey on ant colony optimization, particle swarm optimization, and cuckoo algorithms," 2018.
[13] D. Kumar, S. Kumar, R. Bansal, and P. Singla, "A survey to nature inspired soft computing," *Int. J. Inf. Syst. Model. Des.*, vol. 8, pp. 112–133, 2017.
[14] B. Kovács, F. Tinya, and P. Ódor, "Stand structural drivers of microclimate in mature temperate mixed forests," *Agricultural and Forest Meteorology*, vol. 234, pp. 11–21, 2017.
[15] F. Tinya, B. Kovács, A. Bidló, B. Dima, I. Király, G. Kutszegi, F. Lakatos, Z. Mag, S. Márialigeti, *et al.*, "Environmental drivers of forest biodiversity in temperate mixed forests–a multi-taxon approach," *Science of the Total Environment*, vol. 795, p. 148720, 2021.
[16] M. Chomel, M. Guittonny-Larchevêque, C. Fernandez, C. Gallet, A. DesRochers, D. Paré, B. G. Jackson, and V. Baldy, "Plant secondary metabolites: a key driver of litter decomposition and soil nutrient cycling," *Journal of Ecology*, vol. 104, no. 6, pp. 1527–1541, 2016.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 345–350
DOI: 10.15439/2025F5708
ISSN 2300-5963 ACSIS, Vol. 43

Out-Of-Distribution Is Not Magic: The Clash Between Rejection Rate and Model Success

Itay Meiri*, Ran Dubin†, Amit Dvir‡, Chen Hajaj§
*Dept. of Computer Science, Ariel Cyber Innovation Center, Ariel University, Israel
Email: itay.meiri2@msmail.ariel.ac.il
†Dept. of Computer and Software Engineering, Ariel Cyber Innovation Center, Ariel University, Israel
ORCID: 0000-0002-2055-2211
Email: rand@ariel.ac.il
‡Dept. of Computer and Software Engineering, Ariel Cyber Innovation Center, Ariel University, Israel
ORCID: 0000-0002-3670-0784
Email: amitdv@ariel.ac.il
§Dept. of Industrial Engineering and Management, Ariel Cyber Innovation Center, Ariel University, Israel
ORCID: 0000-0001-9940-5654
Email: chenha@ariel.ac.il

***Abstract*—Recent advancements in Internet protocols, including DNS over HTTPS (DoH) and Encrypted Service Name Indicators (ESNI), are making traditional Deep Packet Inspection (DPI) engines obsolete. Consequently, there is a growing need for next-generation traffic classification using artificial intelligence (AI). While DPI automatically categorizes unknown traffic as 'other,' AI-based models cannot automatically handle unknown or Out-of-Distribution (OOD) traffic. AI models must effectively detect and classify OOD traffic to ensure robustness, reliability, and accuracy in real-world applications; however, current research often fails to address the challenges of OOD detection.**

In this paper, we evaluate various state-of-the-art OOD detection techniques for internet traffic classification and explore the drawbacks and advantages of using different threshold levels for the model's tolerance for OOD. Our findings reveal that varying rejection rates have distinct effects on OOD techniques, leading to a change in the optimal strategy for achieving dependable and precise detection across diverse OOD scenarios. We demonstrate that adjusting rejection rates from 10% to 30% can significantly improve the True Detection Rate (TDR) by up to 50%, while the False Detection Rate (FDR) may increase by less than 10%. Moreover, we emphasize that rejection-rate-based evaluation is pivotal for next-generation flow classification, promising a substantial reduction in FDR through rigorous methodological assessment.

***Index Terms*—Out of Distribution, Traffic Classification, Malware Detection**

I. Introduction

TRAFFIC Classification (TC) is a critical process that automatically categorizes Internet network traffic into distinct classes, such as traffic attribution, application type, or benign/malicious traffic. Regardless of the task, traffic classification plays a crucial role in cybersecurity, Quality of Experience (QoE), and Quality of Service (QoS), as it enables the implementation of predetermined policies to treat traffic classes differently, optimizing network performance and reliability. Traditional traffic classification techniques have relied on classifying applications or services based on fixed port numbers [1]. While these techniques offer advantages such as user privacy preservation, speed, and wide device coverage, they are limited by their reliance on fixed ports and susceptibility to cheating via packet editing. These limitations have prompted the development of Deep Packet Inspection (DPI) classification techniques [2], which involve inspecting the actual payload of packets and are less vulnerable to cheating. However, DPI techniques are resource-intensive, slow, and lack the same privacy guarantees as port-based techniques.

The widespread adoption of encryption protocols like TLS and DoH has further complicated traffic classification, rendering classical DPI techniques obsolete [3]. Consequently, researchers have turned to Machine Learning (ML) and later Deep Learning (DL) techniques for traffic classification. While ML-based approaches initially required manual feature extraction by experts, DL techniques have emerged as promising alternatives for traffic classification [4]. Despite the success of ML and DL [5], one significant weakness persists: their inability to classify instances outside of the closed set of classes in the training data. Network traffic is inherently dynamic, with new applications continually being introduced. This dynamic nature makes it challenging for models to accurately classify unseen classes without retraining. Additionally, acquiring samples for new classes is time-consuming and often results in limited datasets, leading models to favor older and more represented classes during training [6]. To address these challenges and ensure accurate and robust classification in dynamic settings, models must be able to "reject" the classification of a sample that does not belong to any of the learned classes. This capability, known as Out-of-Distribution (OOD) detection [7], [8], is crucial for effectively handling samples outside the training set. Previous works evaluate OOD detection techniques against a well-fit, state-of-the-art model without considering the implications of using less accurate models, which are common in real-world settings [9].

In this paper, we assess the impact of OOD methods on classifier performance in both binary and multiclass traffic

classification tasks. Our contribution extends existing research by highlighting the impact of employing OOD techniques on varying accuracy-level classifiers. Furthermore, we elucidate the trade-offs associated with employing threshold-based OOD techniques by introducing a new metric, **Rejection Rate (RR)**.

These trade-offs are frequently overlooked in discussions surrounding OOD techniques, as they tend not to cater to specific model requirements. By shedding light on these trade-offs, we provide valuable insights for practitioners and researchers, enabling them to make more informed decisions when selecting OOD detection techniques for their models. For example, models connected through a pipeline might not suffer from high rejection rates or data loss if they ensure that the remaining data is of high quality and reliability. This leads to a preference for performance and a high rejection rate of OOD detection techniques.

Lastly, we examine existing OOD techniques on various tasks and datasets, showing that the performance of "state-of-the-art" OOD detection methods is contingent on external factors such as acceptable rejection rates and model performance.

II. Related Work

OOD detection has been studied under many different names, such as zero-day detection, open-set recognition, and rejection option classifiers [4], [7], [8], [10], [11], [12], [13], [14]. Shared by all is the classification model's ability to reject an input if it detects that it is from a class outside its training set. Thulasidasan et al. [10] add a rejection class, K+1, trained on a mix of OOD samples. This method is limited in practice because it demands extra OOD data—either a Generative Adversarial Network or by merging some classes—plus an architectural change and full retraining, with no guarantee that real-world OOD data will match what was seen in training. Devries and Taylor [11] instead train the model to output calibrated confidence scores, using misclassified In-Distribution (ID) samples as stand-ins for OOD. These are built in parallel to the original classification model, which does not affect the model's accuracy. Nonetheless, this requires knowledge of the internal workings of the model and retraining.

Another approach uses Siamese Neural Networks (SNN) as an OOD detection method for network traffic [12]. The authors use the ability of SNNs to find similarities in input to detect OOD, but these require clustering, data balancing, and training. He et al. [13] construct skew data and then train multiple one-class Random Forest classifiers based on the skew data. This is an expensive operation and gets increasingly more complex as the number of classes increases.

Out-of-distribution detector for neural networks (ODIN) [15] augments any pre-trained model without architectural changes by pairing temperature scaling with slight input perturbations. It computes a classifier's confidence, perturbs the input proportionally, and recomputes confidence. Because ID samples are more affected, they generally yield higher post-perturbation confidences than OOD samples, enabling threshold-based rejection. Our work is the first to test ODIN on encrypted network traffic for OOD detection.

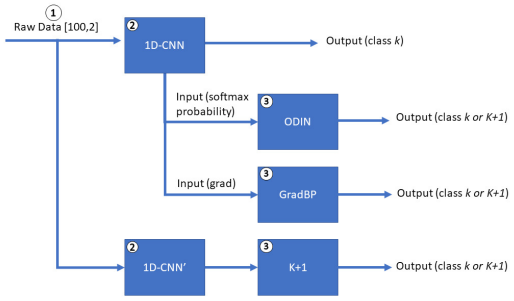


Fig. 1. Our Evaluation System

GradBP [4] treats the magnitude of the first back-propagation gradient, calculated as if training continued, as an OOD proxy. Novel inputs typically induce larger gradient magnitudes, allowing a single threshold to suffice for OOD detection. GradBP is space and time-efficient, works on pre-trained models, and has already been validated on encrypted traffic.

To conclude, several solutions to traffic classification exist, each characterized by its degrees of success and practicality. However, none of these works explores the influence of thresholds and their effect on OOD classification. As we will demonstrate, the influence of the threshold is critical for adequately evaluating OOD techniques.

Our main goal is to evaluate OOD techniques and compare their results for traffic classification, then experiment with the effects of the threshold. We chose a typical 1D-CNN as our base architecture, and three OOD techniques - ODIN [15], GradBP [4], and K+1 [16] as a baseline. ODIN and GradBP are model-agnostic and do not require model modification, retraining, or generating any additional data. Both ODIN and GradBP operate on the model output under the same space and time complexity class and require a threshold. All methods will be concatenated to our base architecture. Our evaluation system can be seen in Fig. 1.

III. System Architecture and Evaluation Framework

A. Basic System Components

Input. Each sample is a 100-step sequence of packet *direction* (0 = client $\rightarrow$ server, 1 = server $\rightarrow$ client) and min–max scaled *size* ($[0, 1]$), a standard low-cost traffic representation [4], [17] (Fig. 1,(1)).

Output. The classifier predicts a label $k \in [1, K]$; OOD detections are mapped to $k = K + 1$.

B. Out-of-Distribution Detectors

ODIN [15]. For a trained network with logits $f(x)$, we (i) rescale them by a temperature T, (ii) add a small perturbation $\epsilon \cdot \text{sign}\big(-\nabla_x \log S(x;T)\big)$, and (iii) reject any sample whose

TABLE I
TRAFFIC DATASETS USED IN THIS STUDY. B = BENIGN, M = MALWARE, TLS = DATASET IS PREDOMINATELY TLS-ENCRYPTED.

Dataset	Role	Classes	Samples	TLS	Notes
MTAB [21]	ID	2 (B/M)	29,000	Mix	MTA + benign
TEMU [20]	ID	8 (B)	20,000	Yes	Browser / misc.
USTCB [21], [22]	OOD	5 (M)	58,000	Mix	Diverse malware
VNAT [23]	OOD	6 (B)	5,500	Mix	VPN successor

largest softmax probability $S(x;T)$ falls below a threshold δ. The triplet (T, ϵ, δ) is chosen on validation data; rejected inputs are mapped to label $K+1$.

GradBP [4]. Keeping weights frozen, we back-propagate each test point once and measure the ℓ_2-norm of the resulting gradient at the penultimate model layer, $\|\delta^{L-1}\|_2$. Points with $\|\delta^{L-1}\|_2 > \epsilon_{\text{GR}}$ are considered OOD and likewise assigned $K+1$.

K + 1 [16]. The classifier head is expanded to $K+1$ outputs and fine-tuned on a small, heterogeneous set of OOD traces labelled $K+1$. At inference, any prediction of the extra node signals an outlier.

All OOD models are under label (3) in Fig. 1.

C. *Datasets*

To obtain a balanced and realistic evaluation spectrum, we use *MTAB* and *TEMU* as our two ID references and *USTCB* and *VNAT* as the OOD challengers. Together, these four datasets cover benign and malicious traffic, multiple application protocols, and both encrypted and plain-text flows, providing a rigorous test bed for evaluating ID performance and OOD generalisation. Specifically, **MTAB** (29 k flows, 2 classes) merges malicious mail-transfer (MTA) traffic with benign sessions from ISCX2016 [18], StratosphereIPS [19], and TEMU [20], yielding a mixed TLS/plain corpus labelled *benign* or *malware*. **TEMU** (20 k flows, 8 benign classes) contains browser and miscellaneous application traffic, most of it TLS-encrypted. **USTCB** (58 k flows, 5 malware classes) augments the USTC malware set—covering the *Cridex*, *Neris*, *Mi-uref*, and *Htbot* families—with the same benign sources as MTAB, giving a malware-heavy OOD benchmark. Finally, **VNAT** (5.5 k flows, 6 benign classes) offers an updated VPN/no-VPN corpus spanning streaming, chat, file-transfer, and other activities with a mix of encrypted and clear-text flows. Table I summarises the core statistics of each dataset, while the exact pairing of datasets in each experiment is shown later in Table II.

D. *Evaluation Metrics*

We report common performance indicators used for OOD evaluation [12], [13]:

Accuracy – proportion of correct decisions among the accepted samples.

True Detection Rate (TDR) – sensitivity to OOD inputs (OOD true positives divided by all OOD cases).

False Detection Rate (FDR) – type-I error on ID data (ID false positives divided by all ID cases).

Rejection Rate (RR) – The percent of samples rejected.

Recall and **Precision** – class-label sensitivity and positive predictive value, both computed on the accepted subset.

IV. EXPERIMENTAL DESIGN

We structured our evaluation around two distinct classification tasks: **Binary Classification** (i.e., distinguishing between benign and malicious traffic) and **Multiclass Classification** (i.e., identifying various well-known applications). Each classification task was evaluated using three separate experiments:

Experiment 1: This experiment assessed the base model's performance without OOD detection and unknown classes. It simulates a standard scenario where a model is developed and tested against a predefined dataset of known classes.

Experiment 2: In this experiment, we evaluated the base model's performance without OOD detection but with the introduction of unknown classes. This mimics deploying the Experiment 1 model in a real-world setting, where data may include previously unseen classes.

Experiment 3: The final experiment aimed to address the challenge of unknown classes by evaluating the base model with OOD detection capabilities.

The structure of our experiments can be seen in Table II. All OOD detection methods were tested on the same test set for robust evaluation. For comprehensive insights, including confusion matrices, tables, and threshold graphs, we have provided detailed results in our GitHub repository [24].

Lastly, when considering the preferability of one method or another, we limited the rejection rate to 50%. The limitation has two reasons: First, we could naively reject nearly all inputs without limit until few ID samples remain and achieve perfect accuracy. We do not consider such a result helpful in evaluating OOD detection performance. Second, there is no standard rejection rate, and the authors of ODIN and GradBP do not explain the effects of different rejection rates. It is standard practice to use FDR=95%TDR, such as in [13], [12], [4], [25], [26], [15], but this result is not useful for all use cases. ODIN [15] suggests hyperparameter tuning that will result in the correct identification of 95% ID samples, but does not explain the effect on rejection rates. GradBP [4] does not give any guidelines at all. As we will demonstrate, the optimal strategy is task-dependent and model-dependent.

V. EXPERIMENTAL RESULTS

A. ***Experiment 1*** *- without unknown classes, without OOD detection*

Our first experiment assesses our models' performance in both binary and multiclass classification tasks, focusing on scenarios where unknown classes and OOD data are not considered. The base models demonstrate remarkable performance despite utilizing only packet size and direction as time-series features. As shown in the Base column of Table III, the accuracy for both Binary and Multiclass tasks reaches impressive levels of 98% and 94%, respectively. This underscores the effectiveness of our models in classifying traffic even without the need for OOD detection mechanisms.

TABLE II
COMPARISON OF ARCHITECTURES AND EVALUATION METRICS ACROSS BINARY AND MULTICLASS CLASSIFICATION TASKS. BASE IS WITHOUT INCLUDING UNKNOWN CLASSES, AND BASE* IS WITH INCLUSION OF UNKNOWN CLASSES.

Exp.	Architecture	Task	Train/Test	Unknown	Evaluation Metrics
1	**Base**	Binary Multiclass	MTAB TEMU	– –	Accuracy, Recall, Precision –
2	**Base***	Binary Multiclass	MTAB TEMU	USTCB VNAT	Accuracy, Recall, Precision –
3	**Base + ODIN**	Binary Multiclass	MTAB TEMU	USTCB VNAT	Accuracy, Recall, Precision, True/False Detection Rate, Rejection Rate –
3	**Base + GradBP**	Binary Multiclass	MTAB TEMU	USTCB VNAT	Accuracy, Recall, Precision, True/False Detection Rate, Rejection Rate –
3	**Base' + (K+1)**	Binary Multiclass	MTAB+USTCB TEMU+VNAT	USTCB VNAT	Accuracy, Recall, Precision, True/False Detection Rate, Rejection Rate –

TABLE III
PERFORMANCE COMPARISON OF MODELS WITH (BASE*) AND WITHOUT (BASE) THE INCLUSION OF UNKNOWN CLASSES IN BINARY AND MULTICLASS CLASSIFICATION TASKS.

Metric	Binary Classification		Multiclass Classification	
	Base	Base*	Base	Base*
Accuracy	0.98	0.73	0.94	0.74
Recall	0.98	0.65	0.88	0.77
Precision	0.98	0.49	0.93	0.64

*B. **Experiment 2** - with unknown classes, without OOD Detection*

Next, we introduce previously unseen classes into the classification tasks, simulating real world deployment in which the model encounters new classes and no OOD detection is applied. This setting yields a pronounced drop relative to Experiment 1 across all metrics, as reported in the $Base^*$ columns of Table III. The decline exposes the model's vulnerability to novel inputs and underscores the need for robust OOD detection to mitigate this risk.

*C. **Experiment 3** - with unknown classes, with OOD detection*

Our third experiment evaluates OOD detection performance, especially the trade-offs of using a threshold-based technique on the accuracy and overall sample rejection rates. For this task, we implemented three detection techniques that will be evaluated: ODIN[15], GradBP[4], K+1[16].

In the following subsections, we demonstrate the model's performance across several rejection rates and justify choosing one threshold over another.

*1) **Classification Task - Binary, OOD - ODIN**:* Tables IV and VI show that as the rejection rate initially increases, nearly all rejected samples are OOD, as both accuracy and TDR rise. As the threshold tightens, we observe an exponential increase in FDR with minimal accuracy gains. This is essential because there might be a significant increase in the model's accuracy by allowing the model to reject slightly more ID samples. Increasing the rejection rate from 10% to 30% produced several effects: the TDR increased from 31% to 66%, while the FDR only increased from 3% to 18%, and accuracy increased from 79% to 87%. In contrast, going from a rejection rate of 30% to 50% had exponentially diminishing returns. The TDR increased from 66% to 75%, while FDR increased from about 18% to 41%, leaving the accuracy unchanged at 87%. At that point, accuracy rates began decreasing as the threshold tightened.

Despite allowing a rejection rate of up to 50%, the best accuracy figures came from rejecting only 31% of the overall samples. The accuracy score increased from 72% to 87%, with an FDR of 18% and a TDR of 67%, as seen in Table VI. These results demonstrate that despite ODIN being originally developed for images, it is generic enough to work on network traffic. However, there are apparent weaknesses as well. ODIN misses examples that are likely to be misclassified, and nearly all OOD samples missed were classified as benign. This weakness is particularly problematic because the purpose of the model is to detect malicious traffic.

Note that the rejection rate does not influence the K+1 technique; thus, it was not included in this analysis and does not appear in Table IV.

*2) **Classification Task - Binary, OOD - GradBP**:* Similarly to ODIN, as shown in Tables IV and VI, the FDRs are low as the threshold decreases, proving that nearly all samples rejected are OOD. Unlike ODIN, the GradBP results are much better at the maximum threshold rate of 50% (98% TDR). Nearly all OOD samples have been rejected, increasing the model's accuracy to 99% - higher than the base model without OOD data. The FDR is 32%, which means most ID samples are not rejected. The higher accuracy signals, again, that this method also rejects nearly all misclassified labels. Out of the missed OOD samples, only 29/31 were classified as benign by the model, compared to 659/661 in ODIN. We also compared these values at a 30% rejection rate; 193/362 OOD samples were detected as benign, a significantly better result than ODIN. As the threshold tightens, GradBP shows an exponential increase in practical terms. This means we can significantly increase the likelihood of all passed samples being ID if data loss is not a significant concern.

TABLE IV
EXPERIMENT 3: THE INFLUENCE OF REJECTION RATE ON THE PERFORMANCE OF ODIN AND GRADBP METHODS FOR THE BINARY CLASSIFICATION TASK.

Metric	10%		30%		50%		70%		90%	
	ODIN	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**
Accuracy	0.79	0.79	0.87	**0.93**	0.87	**0.99**	0.83	**1.00**	–	**1.00**
Recall	0.66	0.66	0.66	0.66	0.67	0.67	0.67	0.67	–	**1.00**
Precision	**0.54**	0.53	0.59	**0.62**	0.60	**0.66**	0.59	**0.67**	–	**1.00**
FDR	0.03	0.03	0.18	**0.12**	0.41	**0.33**	0.65	**0.60**	–	**0.87**
TDR	**0.31**	0.30	0.66	**0.82**	0.75	**0.98**	0.80	**1.00**	–	**1.00**

TABLE V
EXPERIMENT 3: INFLUENCE OF REJECTION RATE ON MODEL PERFORMANCE FOR MULTICLASS CLASSIFICATION USING ODIN AND GRADBP.

Metric	10%		30%		50%		70%		90%	
	ODIN	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**	**ODIN**	**GradBP**
Accuracy	**0.79**	0.78	**0.90**	0.87	**0.96**	0.93	**0.98**	0.95	–	**0.98**
Recall	**0.82**	0.77	**0.85**	0.76	**0.86**	0.85	**0.83**	0.82	–	**0.80**
Precision	**0.70**	0.66	**0.75**	0.64	**0.79**	0.74	**0.75**	0.73	–	0.69
FDR	0.07	0.07	**0.20**	0.22	**0.39**	0.41	**0.63**	0.64	–	0.88
TDR	**0.21**	0.20	**0.70**	0.59	**0.92**	0.84	**0.97**	0.93	–	0.99

TABLE VI
EXPERIMENT 3: PERFORMANCE SUMMARY FOR BINARY CLASSIFICATION USING ODIN, GRADBP, AND K+1 METHODS.

Metric	ODIN	GradBP	K+1
Accuracy	0.88	**0.99**	0.94
Recall	**0.67**	**0.67**	0.66
Precision	0.59	**0.66**	0.63
FDR	0.19	0.32	**0.02**
TDR	0.68	**0.98**	0.86
Rejection Rate	0.31	0.49	**0.23**

*3) **Classification Task - Binary, OOD - K+1**:* As seen in Table VI, the results from adding a rejection class were inconclusive. The accuracy achieved was 93.8%, which is lower than GradBP and ODIN. In the K+1 technique, there is no threshold to work with when adding a rejection class, so the rejection rate is fixed at 23%.

*4) **Classification Task - Multiclass, OOD - ODIN**:* As can be seen from Tables V and VII, there is full utilization of the rejection rate at about 50%, achieving an accuracy, TDR, and FDR of 96%, 92%, and 39%, higher than the base model when tested without OOD data. This ability means that ODIN rejects OOD samples and misclassifies known applications. For comparison, at a 30% rejection rate, the model had good accuracy, TDR, and FDR of 90%, 70%, and 20%. Table V (30% rejection rate, ODIN). Using the threshold to increase its effectiveness allows for a greater range of risk management; if losing most data is not a significant concern, we can increase the confidence that the samples that do pass are both ID and correctly classified.

*5) **Classification Task - Multiclass, OOD - GradBP**:* The results seen in Table VII for GradBP were interesting; at a 50% rejection rate, we observed an accuracy rate of 93%, lower than ODIN. TDR was 84%, also lower than ODIN, whereas the FDR was 41% compared to 39% in ODIN. The closest result was in recall, which has 85% compared to ODIN with 86%. The increase in rejection rate did not contribute much beyond 30%.

*6) **Classification Task - Multiclass, OOD - K+1**:* As shown in Table VII, the accuracy rate was 84%, with a recall of 78% and a precision of 66%. The rejection rate accounted for only 12% of the total samples, and the TDR was just 54%. However, the FDR remained remarkably low at only 1.5%, similar to the binary model.

TABLE VII
EXPERIMENT 3: PERFORMANCE SUMMARY FOR MULTICLASS CLASSIFICATION USING ODIN, GRADBP, AND K+1 METHODS.

Metric	ODIN	GradBP	K+1
Accuracy	**0.96**	0.93	0.84
Recall	**0.86**	0.85	0.78
Precision	**0.79**	0.75	0.66
FDR	0.39	0.40	**0.02**
TDR	**0.92**	0.84	0.54
Rejection Rate	0.50	0.49	**0.12**

VI. DISCUSSION AND CONCLUSIONS

Our empirical study of ODIN, GradBP, and K+1 across binary and multiclass traffic classification tasks paints a nuanced picture; there is no universally superior OOD method. Each technique excels only under specific operational conditions, a finding with immediate consequences for how OOD detection is benchmarked and deployed. In the multiclass setting, ODIN achieved the highest overall accuracy, reaching 96% in our experiments and preserving a favorable balance between true and false detection rates once the rejection rate exceeded roughly 30%. ODIN's temperature-scaled perturbation interacts advantageously with the richer softmax that accompanies

multiclass models, yielding robustness as the class cardinality increases. By contrast, in the binary task, GradBP emerged as the preferred method whenever practitioners could tolerate moderate to high rejection rates. Above the 30% rejection threshold, GradBP drove accuracy toward 99% and pushed the TDR past 98%, maintaining lower FDR rates than ODIN.

Although often dismissed as a baseline, K+1 demonstrated excellent precision at very low FDR. However, its practical value is undercut by the need for explicit OOD samples, a fixed rejection threshold, the computation costs of model retraining, and the possible effects on model performance. Rejection rates surfaced as the governing hyperparameter in the third experiment. With RR capped at 10%, ODIN and GradBP were statistically indistinguishable in the binary task, offering practitioners no basis for preferring one over the other. Once rejection rates were relaxed into the 30%-50% regime, the behavior of both methods diverged sharply: GradBP continued to accrue accuracy and detection gains, whereas ODIN effectively plateaued. In the multiclass task, the pattern reversed - a relaxed rejection rate benefited ODIN and put it significantly ahead of GradBP across all metrics, while GradBP's improvements saturated. When rejection rates were not relaxed, the performances of GradBP and ODIN were virtually identical. The bidirectional sensitivities confirm that rejection rates reshape each detector's risk-utility frontier and must be considered when benchmarking OOD detection methods. The operational consequences could be applied immediately; in domains where practitioners could tolerate elevated rejection rates at the price of suppressing false acceptances, candidate detection methods should be tested under different rejection rate circumstances to determine the ideal fit. Our findings show that OOD detection is far from a "plug-and-play" commodity. Its efficacy is contingent upon task granularity, traffic type, and tolerance for rejected samples.

Future research endeavors should explore the impact of rejection rates across a broader array of OOD detection methods and classification domains. Specifically, how strongly rejection rates improve OOD detection and performance across OOD detection performance and model accuracy. Future investigations should explore the connection between different models, domains, and rejection rates, potentially uncovering novel approaches for improved performance.

ACKNOWLEDGMENT

This work was supported by the Ariel Cyber Innovation Center, the Israel Innovation Authority, and the Trust.AI consortium.

REFERENCES

[1] D. Moore, K. Keys, R. Koga, E. Lagache, and K. C. Claffy, "The {CoralReef} software suite as a tool for system and network administrators," in *LISA*, 2001.

[2] S. Xu, J. Han, Y. Liu, H. Liu, and Y. Bai, "Few-shot traffic classification based on autoencoder and deep graph convolutional networks," *Scientific Reports*, vol. 15, no. 1, p. 8995, 2025.

[3] P. Tang, Y. Dong, S. Mao, H.-L. Wei, and J. Jin, "Online classification of network traffic based on granular computing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 8, pp. 5199–5211, 2023.

[4] L. Yang, A. Finamore, F. Jun, and D. Rossi, "Deep Learning and Traffic Classification: Lessons learned from a commercial-grade dataset with hundreds of encrypted and zero-day applications," *arXiv preprint arXiv:2104.03182*, 2021.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[6] X. Wang, Y. Wang, Y. Lai, Z. Hao, and A. X. Liu, "Reliable open-set network traffic classification," *IEEE Transactions on Information Forensics and Security*, 2025.

[7] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-Set Recognition: A good closed-set classifier is all you need," *CoRR*, vol. abs/2110.06207, 2021. [Online]. Available: https://arxiv.org/abs/2110.06207

[8] V. Franc, D. Průša, and V. Voracek, "Optimal strategies for reject option classifiers," *CoRR*, vol. abs/2101.12523, 2021. [Online]. Available: https://arxiv.org/abs/2101.12523

[9] Z. Fang, J. Lu, and G. Zhang, "Out-of-distribution detection with non-semantic exploration," *Information Sciences*, vol. 705, p. 121989, 2025.

[10] S. Thulasidasan, S. Thapa, S. Dhaubhadel, G. Chennupati, T. Bhattacharya, and J. Bilmes, "An effective baseline for robustness to distributional shift," in *ICMLA*. IEEE, 2021, pp. 278–285.

[11] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[12] Y. Chen, Z. Li, J. Shi, G. Gou, C. Liu, and G. Xiong, "Not Afraid of the Unseen: a Siamese Network based Scheme for Unknown Traffic Discovery," in *ISCC*, 2020, pp. 1–7.

[13] H. He, Y. Lai, Y. Wang, S. Le, and Z. Zhao, "A data skew-based unknown traffic classification approach for TLS applications," *Future Generation Computer Systems*, 2022.

[14] Z. Yang and W. Lin, "Unknown traffic identification based on deep adaptation networks," in *2020 IEEE 45th LCN Symposium on Emerging Topics in Networking (LCN Symposium)*. IEEE, 2020, pp. 10–18.

[15] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[16] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open Set Learning with Counterfactual Images," in *European Conference, Munich, Germany, September 8–14, 2018*, 2018, pp. 620–635.

[17] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 76–81, may 2019.

[18] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of Encrypted and VPN traffic using time-related features," in *ICISSP, Rome, Italy, February 19-21*, 2016, pp. 407–414.

[19] Stratosphere, "Stratosphere Laboratory Datasets," 2015, retrieved March 13, 2020, from https://www.stratosphereips.org/datasets-overview.

[20] R. Dubin, A. Dvir, O. Pele, J. Muehlstein, Y. Zion, M. Bahumi, and I. Kirshenboim, "Analyzing HTTPS Encrypted Traffic to Identify User's Operating System, Browser and Application," in *IEEE CCNC*, June 2017.

[21] A. Lichy, O. Bader, R. Dubin, A. Dvir, and C. Hajaj, "When a rf beats a cnn and gru, together—a comparison of deep learning and classical machine learning approaches for encrypted malware traffic classification," *Computers & Security*, vol. 124, p. 103000, 2023.

[22] W. Wang and D. Lu, "USTC-TFC2016," 2016. [Online]. Available: https://github.com/yungshenglu/USTC-TFC2016

[23] S. Jorgensen, J. Holodnak, J. Dempsey, K. de Souza, A. Raghunath, V. Rivet, N. DeMoes, A. Alejos, and A. Wollaber, "Extensible Machine Learning for Encrypted Network Traffic Application Labeling via Uncertainty Quantification," *arXiv preprint arXiv:2205.05628*, 2022.

[24] I. Meiri, "Ood not magic - github repository," 2025, available at: https://github.com/ArielCyber/OOD_Is_Not_Magic.

[25] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[26] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 351–356
DOI: 10.15439/2025F8576
ISSN 2300-5963 ACSIS, Vol. 43

Digital Twin Micro-Services Architecture to Support a Scenario-based Integration of Smart Mobility Services and Simulations

Damien Nicolas, Ruben J. Marques, Christophe Feltus
0000-0002-6388-032X, 0009-0004-5842-9544, 0000-0002-7182-8185
Luxembourg Institute of Science and Technology (LIST)
Avenue des Hauts–Fourneaux, 5, L–4362 Esch–sur–Alzette, Luxembourg
Email: *firstname.lastname*@list.lu

***Abstract*—This paper proposes a microservices-based digital twin architecture for flexible, scalable, and real-time integration of smart mobility services in urban centers. By decoupling functions into modular services embedded in scenario-driven simulations, the digital twin overcomes limitations of monolithic systems, enhancing adaptability and interoperability. Preliminary results demonstrate its effectiveness for dynamic urban mobility management.**

***Index Terms*—Digital Twin, Mobility, Reference architecture, Micro-services approach, Scenario-driven approach, Mobility Simulation.**

I. Introduction

AS CITIES adopt smart mobility solutions to tackle congestion, reduce emissions, and improve transport efficiency, the ecosystem of services—ride-sharing, public transport tracking, electric vehicle infrastructure, and autonomous systems—is rapidly expanding[1]. These services require real-time interaction, but most current integration strategies are rigid, tightly coupled, and difficult to scale or update. This limits cities' ability to adapt their mobility infrastructure as technologies and needs evolve[2]. Digital twins—virtual representations of physical systems—offer a promising approach to dynamically model, simulate, and manage urban mobility. Paired with microservices architecture, digital twins (DT) can evolve from static models into flexible, modular platforms supporting continuous integration of new services[3]. However, many existing DT implementations for mobility rely on monolithic or semi-monolithic systems lacking modularity and real-time interoperability. Adding new services often requires major redevelopment, causing bottlenecks and hindering innovation. There is a need for an architecture that enables seamless integration, fast iteration, and scalability without compromising performance or reliability[4]. This paper presents a microservices-based DT architecture designed for flexible, scalable integration of smart mobility services. Each service can operate independently while contributing to a unified city-wide model, enabling real-time coordination and dynamic updates with minimal friction. A key feature of the architecture is embedding simulation within workflow-driven scenarios and sub-scenarios, rather than as a separate module. This modular design enables configurable, extensible simulation pipelines that capture urban mobility's complexity and support context-specific experimentation and decision-making. Section III defines the concept of Simulation and Scenario. Section IV proposes the design principles for the architecture. Section V details the proposed architecture and its key components. Section VI describes the implementation of the case study to validate the approach. Section VII discusses the results, challenges, and insights gained and concludes with key takeaways and directions for future work.

II. Related work

The *Digital Twin* concept was first introduced in a NASA roadmap[1] [5] as a probabilistic, multi-physics simulation of a space system, combining physical models, sensor data, and operational history. Grieves [6] later formalised it through three core elements: *a physical product, a virtual counterpart, and the data links between them*. As summarised by [7], a DT is a dynamic virtual representation of a physical system, connected throughout its lifecycle for continuous data exchange. Without this temporal link, the model remains a static simulation rather than a true DT. In *Ceci n'est pas un jumeau numérique*, [8] describes the progression from design models to data-enriched *shadows*, with unidirectional data flow. Twin status requires feedback enabling the virtual to affect the physical. While bidirectional exchange is essential, [9] argues that full understanding requires five dimensions: physical entities, virtual models, data, services, and interconnections. Beyond representation, DTs support decision-making—through monitoring (e.g., vehicle status), simulation (e.g., traffic impacts), and prediction (e.g., mobility demand).

Micro-Services Architecture in Smart Systems enables scalability and maintainability by dividing applications into loosely coupled components that communicate via lightweight APIs. This modularity helps manage complexity in domains like cities, energy grids, or mobility, where subsystems—such as traffic control or route planning—can evolve independently.

[1] $https:\backslash\backslash www.nasa.gov\backslash sites\backslash default\backslash files\backslash 501321main_TA11-ID_rev4_NRC-wTASR.pdf$

TABLE I
TECHNICAL LIMITATIONS OF CURRENT APPROACHES

Challenge	Description
Legacy system integration	Difficulty integrating DTs with older infrastructure lacking IoT/sensor support
Scalability	Maintaining accuracy across expanding systems with diverse data volumes
Real-time synchronisation	Latency issues when combining Edge (fast) and Cloud (slow) DTs
Data quality and Governance	Errors in data pipelines lead to inaccurate DT representations

Recent use cases show improved iteration and fault isolation, such as in smart grids for energy forecasting and pricing [10], or in IoT cities for managing sensor and edge data streams [11]. Still, challenges like orchestration and service discovery persist but are mitigated with tools like Kubernetes and service meshes. For mobility-focused DTs, microservices enable real-time simulation, modular updates, and integration of varied services [12].

Integration Frameworks for Mobility Services face challenges due to data mismatches, conflicting services, and diverse tech stacks, limiting real-time performance. Automated methods offer potential but struggle with incomplete data, and domain-specific issues persist. In healthcare, DTs encounter uncertain model parameters and data ownership concerns, needing to balance AI insights with regulations like HIPAA [13].Urban DTs face difficulty simulating emergent behaviours across systems and resolving stakeholder conflicts [14]. Table I summarises key technical challenges in integrating SoS and DTs.

III. SIMULATION AND SCENARIO CONCEPT

This section defines simulation and scenario concepts, key to the proposed microservice architecture, enabling modular and scalable system modelling.

A. Simulation

The simulation is a central element of the digital twin concept. It enables the virtual reproduction of the real system's behaviour using various specialised simulation tools. In this work, four simulators have been integrated: SUMO[2], AIMSUN[3], MnMS[4], and MATSim[5]. These tools cover a wide range of scenarios and analyses in the field of mobility.

The simulation component was modelled using UML and consists of several inputs: parameters, context, initial conditions, constraints, and metrics for performance evaluation. The simulation produces outputs such as key performance indicators (KPIs) or visual results. These outputs can be displayed by the digital twin and support decision-making processes.

Table II provides a detailed description of the different concepts that make up a scenario.

[2]https://www.eclipse.dev/sumo
[3]https://www.aimsun.com
[4]https://github.com/licit-lab/MnMS
[5]https://www.matsim.org

TABLE II
SIMULATION CONCEPT DEFINITIONS

Name of the Concept	Description
Parameter	This class specifies parameters or settings that control the behaviour of the simulation. These parameters may include simulation duration, time step size, numerical methods, model parameters, and other configuration options.
Outcome	This class specifies the anticipated results, findings or insights that are expected to emerge from the simulation. This provides a basis for comparing the actual outcomes of the simulation with the desired or expected outcomes.
Metrics	This class outlines the criteria or measures used to evaluate the simulation (performance, effectiveness, etc.). It defines the key indicators, metrics, or outcomes that will be analysed to assess the success of the simulation in achieving its objectives.
Constraint	This class specifies the constraints, limitations or assumptions that apply to the simulation.
Initial Condition	This class defines the initial states or conditions at the beginning of the simulation. It includes parameters such as initial values, states, positions, velocities, and other relevant variables that characterise the system.
Context	This class contains the information related to the context of the simulation. This includes details about the environment, stakeholders, resources, constraints, and any other factors that may influence the simulation.

B. Scenario

The concept of a scenario is a core part of the proposed microservice architecture. A scenario can include multiple sub-scenarios to support complex simulation workflows. Each scenario consists of sub-scenarios, which are composed of individual steps called ScenarioStep. Each step represents a specific mobility-related service, such as calibration, stress test generation, traffic management, or simulation. This modular structure allows fine control over scenario behavior and execution. Additionally, each sub-scenario is defined by a workflow that manages the sequence and dependencies of its steps.

IV. DESIGN PRINCIPLES

A. Functional and non-functional Requirements

The functional and non-functional requirements stem from four pilot partners: the City of Amsterdam[6], Athens (via DAEM[7]), Helsinki (via Forum Virium Helsinki), and the Luxembourgish bus operator Sales-Lentz[8]. Each contributes with their domain expertise—from smart city development to digital infrastructure and autonomous transport. User-driven features are grouped by topic and labeled as essential or optional. Key functional needs include data collection from diverse sources, real-time dynamic simulations, scenario execution (e.g., traffic events), and 2D/3D visualization. Non-functional needs emphasize data fusion for enhanced insights and strong data privacy safeguards to prevent unauthorized access to personal

[6]https://www.amsterdam.nl/en/
[7]https://www.daem.gr/
[8]https://sales-lentz.lu/en/

TABLE III
FUNCTIONAL NEEDS BY THEME, WITH KEY (K) OR OPTIONAL (O) PRIORITY

Subj.	Features	K/O
Data	-As a data scientist (DS), I want to browse, upload and download data in different formats (JSON, CSV, XML) in the DT.	K
	-As a DS, I want data to be available according to well-defined access rules	O
	-As a DS, I want to get relevant infrastructure data (occupancy of parking lots, pricing, requests for parking places, etc) using open data	K
	-As a DS from Helsinki, I want the Helsinki DT data catalogue to be connected to the DT data management	K
Scenario	-As a civil servant (CS) or as a DS, I want to access a list of scenarios	K
	-As a DS, I want the scenarios to be labelled, timestamped, related to input, and generated by a simulation tool	K
	-As a DS, I want the scenarios to be stored in a database managed by the DT.	K
	-As a CS from Helsinki or as a DS, I want the simulation to be calibrated to better represent the traffic management strategies for Helsinki's harbour.	K
Simulation	-As a CS or as a DS, I want to realise various traffic simulations	K
	-As an employee from Luxembourg or as a DS, I want to simulate autonomous shuttle operations in the Luxembourg area. Therefore, I want: (1) to calibrate the model with data, (2) to collect and to store data from the national open data portal, (3) to simulate the shuttle mobility and the position of the stop stations	K
	-As a CS from Amsterdam or as a DS, I want to simulate the scenarios when the IJ-tunnel is open/closed.	K
	-As a CS from Helsinki or as a DS, I want to simulate different traffic management strategies and to simulate mobility environments when ferries arrive from Tallinn to Helsinki's harbour.	K
Visualization	-As a CS or as a DS, I want to visualise the different traffic management strategies on a 2D map	K
	-As a CS or as a DS, I want to visualise the different traffic scenarios on a 2D map	K
	-As a CS or as a DS, I want to visualise the KPIs associated with the scenario in a dashboard	K
	-As a CS or as a DS, I would like to make simple queries, such as to highlight the evolution of different KPIs	O
	-As a CS from Athens or as a DS, I want to view the different traffic management strategies in 3D	O
	-As a CS from Helsinki, I want to visualise 1 KPI that expresses the number of recommendations followed by the users (50 users for the pilot)	K
Feedback loop	-As an employee from Luxembourg, I would like the shuttle's position to be sent to the passengers' mobile application	K
	-As a CS from Helsinki, I want the trajectory recommendation to be sent to citizens arriving and living in the Helsinki harbour through a dedicated mobile application	K
Real time	-As a CS from Athens or as a DS, I want to use real-time data	O
	-As a CS from Helsinki, I want to visualise traffic management strategies, traffic flow in real-time, and a 3D viewer	O

data. These requirements shape the architecture supporting the Digital Twin (DT) systems used in the pilot projects.

B. Design of the digital twin

The design of a DT for mobility management requires a robust, scalable, and flexible architecture.

1) Microservices as an Architectural Foundation: Microservices offer flexibility, scalability, and resilience for data-intensive mobility systems ([15]). They support independent deployment and integration. Key design principles include **API-first** development, **domain-driven design** for mobility alignment, **containerization** for isolation, and **service decoupling** to reduce dependencies and ease maintenance. These ensure efficient, cloud-native, and modular architecture.

2) Layered Logical Architecture of the DT: To ease the deployment of the DT, we have decided to divide the architecture into layers, each supporting a dedicated functional principle. The seven layers considered are: 1) **Data Ingestion:** Collects real-time data from IoT sensors, GPS, traffic cameras, etc. 2) **Data Processing and Storage:** Processes and stores data in scalable systems (e.g., Apache Kafka, Cassandra) with harmonised formats.3) **Simulation and Analytics:** Provides insights via analytics and machine learning for forecasting and planning.4) **Integration:** Exposes APIs for system interaction and real-time responses.5) **User Interface:** Offers dashboards for operators to explore data and support decisions.6) **Security and Compliance:** Ensures data privacy and integrity with access controls and encryption. 7) **Scalability and Resilience:** Adapts to growth and evolving needs without major disruptions.

V. PROPOSAL ARCHITECTURE

A. Overview of the Digital Twin Micro-Services Framework

The digital twin platform adopts a high-level microservices architecture to ensure modularity, scalability, and maintainability. Each microservice functions as an autonomous unit with a specific role and may include its own user interface, business logic, and database.

Service communication is centralised through an API Gateway. Instead of direct IP-based interaction, all requests pass through the gateway, enabling unified access control, routing, and monitoring—enhancing both security and operational clarity. The gateway also enforces access policies and simplifies service discovery.

Inter-service communication includes authentication tokens, allowing services to verify the legitimacy of incoming requests. This ensures only authenticated and authorised interactions occur within the platform.

All services expose RESTful APIs, enabling standard HTTP communication and supporting independent development, deployment, and scaling.

To ensure platform independence and ease of deployment, all services are containerised. Containers encapsulate each service and its dependencies, enabling consistent deployment across environments and the use of orchestration tools like Kubernetes for automated deployment, scaling, and management.

B. Micro-Service Components

- **Mobility Data Ingestion:** The *Data Manager* provides a central point for accessing and managing data. It includes a user interface, business logic, and storage system. Users can upload or download datasets, while services can query or consume data for processing or simulation tasks.

- **Simulation & Prediction Engine (external):** This component offers unified access to simulation tools such as SUMO, Aimsun Next, MATSim, and MnMS. It abstracts the specifics of each tool, allowing higher-level services to request simulations through a standard, tool-agnostic interface.
- **Service Registry & API Gateway:** This module centralises all communication within the platform and serves as the access point for external clients. Services register their endpoints with the gateway, which manages routing, exposure control, and access policies. Inter-service communication passes through the gateway and includes authentication tokens to verify identity and permissions.
- **Scenarios Engine:** The *Scenarios Manager* orchestrates simulation scenarios by retrieving datasets from the Data Manager and initiating tasks through the Simulation Engine. It includes user interface components for scenario configuration and results visualisation.
- **User Interface / Visualisation Module:** A central front-end portal integrates visual components from services like the Data Manager and Scenarios Manager via `iframes`. This enables services to independently develop their interfaces while ensuring a unified and seamless user experience.

C. Integration Mechanism

The microservices described earlier form the core of the platform, providing essential functions such as data ingestion, scenario orchestration, simulation execution, and communication routing. Together, they ensure interoperability across all modules. The platform is designed for extensibility: developers can introduce new services using their preferred technology stack, provided each service encapsulates its internal components—such as business logic and storage—and exposes a clear HTTP RESTful API for system-wide interaction. To integrate a new service, it must be registered with the API Gateway, which centralises access control, routing, and service discovery. Developers specify which endpoints are exposed to the platform, ensuring new functionalities remain decoupled, manageable, and aligned with established security and communication standards. This modular design allows the platform to evolve, supporting the integration of new tools, analytics, or decision-support modules as the Digital Twin's requirements grow.

VI. Case Study Implementation

A. Implementation Context and Technology Stack

The reference architecture described in this paper has been instantiated in the context of the ACUMEN use case, which aims to enable advanced mobility functionalities such as traffic monitoring, dynamic route optimisation, public transport coordination, and seamless mobility services. The platform's modular microservices approach, centralised access control, and robust data handling, provides a reliable foundation to meet these operational requirements.

The implementation leverages a carefully selected technology stack to support these capabilities, ensuring scalability, interoperability, and maintainability. Table IV summarises the key technologies and tools used across the platform's core components and supporting infrastructure.

TABLE IV
Technology stack used in the implementation

Module	Technology / Tool	Purpose
API Gateway	Java, Spring Boot, Spring Cloud Gateway, MongoDB	Centralizes communication, manages routing, service registry and discovery
Authentication Service	Keycloak, Java	Handles user authentication, access control, and token management (JWT)
Data Manager	Java, Spring Boot, PostgreSQL, MinIO	Provides a central hub for dataset storage, retrieval, and sharing across services
Scenarios Manager	Java, Spring Boot, ArangoDB	Orchestrates scenario execution, including data fetching and simulation delegation
Simulations Engine	Python, ClickHouse	Interfaces with external tools to manage and process simulation requests
Front-End Shell	Vue.js, Vuetify	Hosts the main portal and integrates service-specific UIs via iframe-based inclusion
Simulation Tools	SUMO, Aimsun Next, MATSim, MNMS	Simulate real-world mobility scenarios and provide output for analysis
Containerization	Docker	Packages services into containers to ensure portability and consistency
CI/CD	GitLab CI/CD, Docker Compose	Automates build, test, and deployment workflows

B. First results of SUMO Simulation Scenario: Greek National Road 91 Closure

This section presents the first results of the scenario execution for a simulation using SUMO. The objective of this scenario is to simulate traffic evolution when the Greek National Road 91 in Athens is closed to traffic, as illustrated in Fig. 2. The scenario was subdivided into several sub-scenarios to visualise the impact of different types of closures (e.g., one lane from north to south, one lane from south to north, or both lanes simultaneously) and to determine the optimal mitigation strategy.

To perform these sub-scenarios, the necessary parameters were introduced. Specifically, the simulation was configured with the following inputs: *rush hour* was enabled, *weather conditions* were set to *rainy*, and an *incident* was introduced in the form of a *vehicle breakdown* located at the *midpoint of the tunnel*. A *dynamic, individually aware rerouting strategy* was applied. *Pedestrian presence* was considered *low*, and there were *no disruptions to public transport*. The simulation was run for *120 minutes* with detailed output including *vehicle*

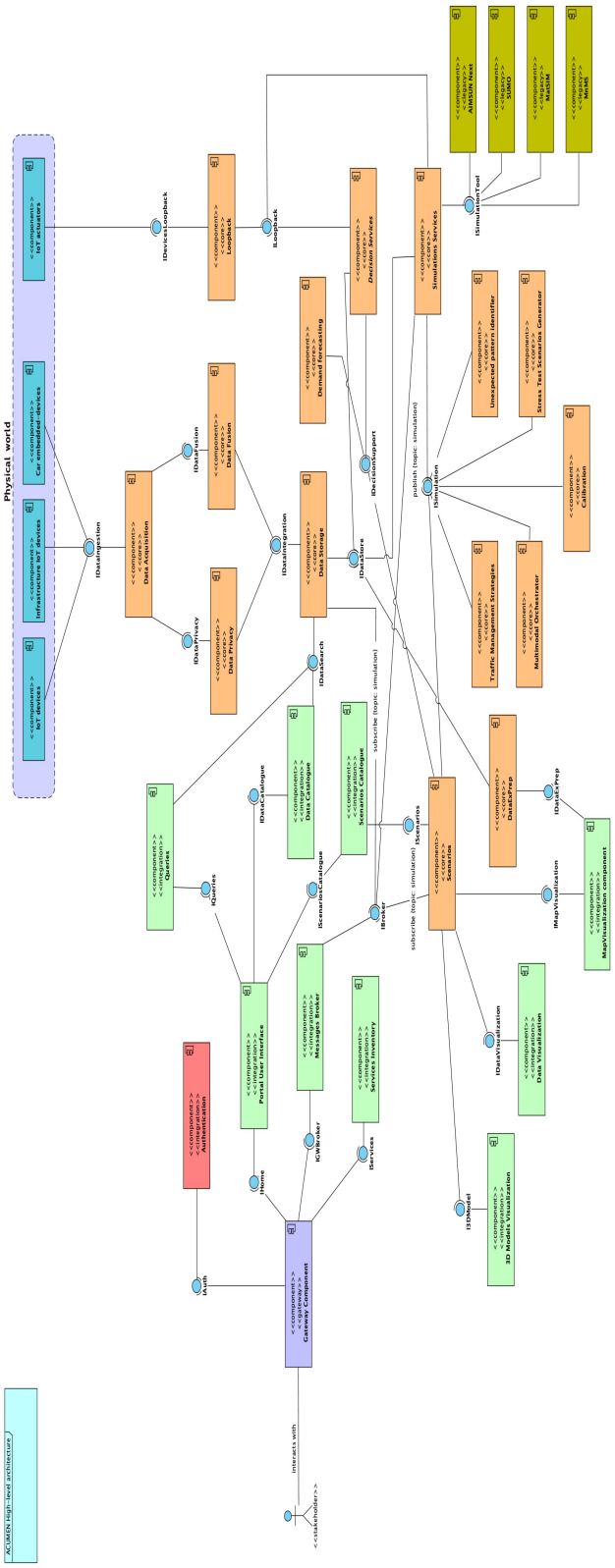


Fig. 1. ACUMEN microservices-based reference architecture

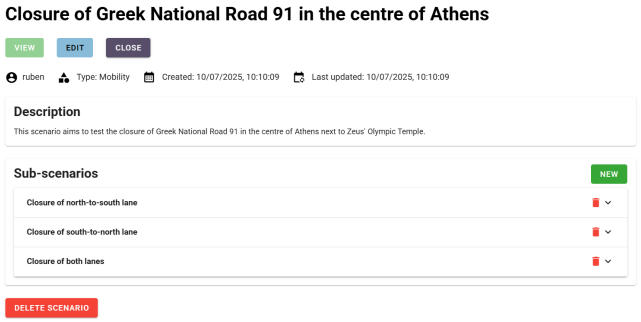


Fig. 2. Global view of the scenario

positions and teleports. The *traffic lights* followed an *adaptive signal control program*, as shown in Fig. 3.

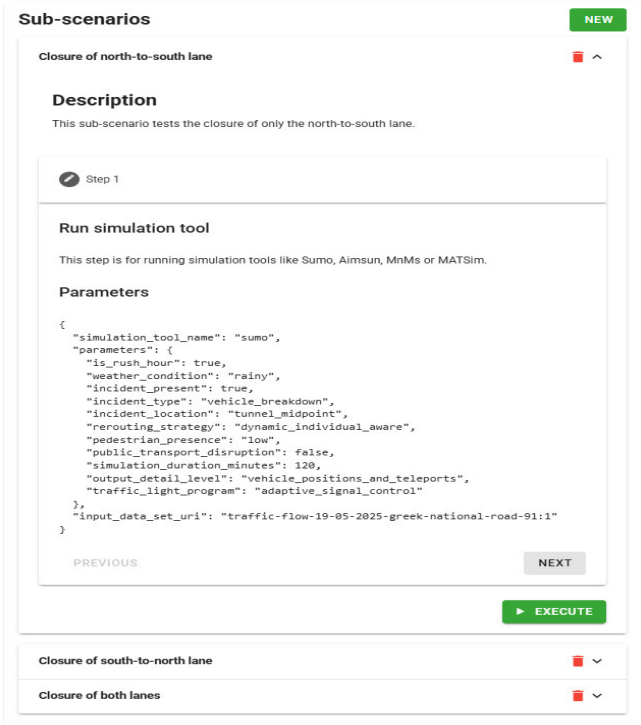


Fig. 3. Sub-scenario parameters

The scenario execution was initiated by clicking the *EXECUTE* green button shown in Fig. 3. After a few minutes of computation, simulation results were available for each sub-scenario, allowing a comparative visualisation of the different traffic configurations. As illustrated in Fig. 4, the results depict the movement of various types of vehicles: *cars in blue*, *buses in green*, and *trams in orange*. These visualisations help to understand how traffic flows redistribute across the city of Athens depending on the specific sub-scenario configuration.

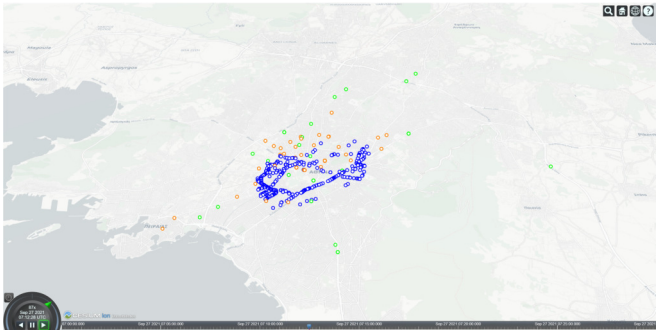


Fig. 4. Visual result of the simulation

VII. Conclusion and Future Works

A Digital Twin Microservices Architecture provides a scalable, modular framework ideal for smart mobility systems. It breaks down complex services—like traffic control or ride-sharing—into independent microservices, improving fault tolerance, development speed, and system updates. When combined with DT, real-world elements like vehicles and infrastructure are mirrored virtually, enabling real-time monitoring, predictive modeling, and system optimization in fast-changing urban environments (transport of dangerous good [16]).

The architecture promotes interoperability across diverse platforms used by different stakeholders, synchronizing data via standard APIs. A shared digital twin layer keeps services aligned with real-time conditions, supporting informed, coordinated decision-making. Its modularity also enables advanced analytics and machine learning, using rich data streams to forecast traffic, optimize routes, and assess infrastructure—without disrupting core operations.

However, several challenges remain. Integrating massive, heterogeneous data sources requires sophisticated pipelines, validation tools, and common standards to ensure accuracy. Orchestrating many microservices also demands complex coordination, secure communication, and real-time responsiveness. Additionally, the approach is resource-intensive, requiring significant investment in computing power, cloud infrastructure, and cybersecurity. As cities scale up smart mobility, managing cost, system complexity, and security becomes increasingly difficult. Overcoming these hurdles will require technical innovation, strong governance, and collaboration across sectors.

This paper proposes a microservices-based digital twin for smart mobility, featuring modular simulations for adaptability and interoperability. A case study on the Athens tunnel closures shows its effectiveness. Future work aims to target automation, self-adaptation, cloud-edge performance, data privacy, and AI explainability, replacing rigid systems with agile urban mobility solutions.

Acknowledgment

We gratefully acknowledge that project ACUMEN (project number 101103808), funded by the European Union as Research and Innovation Action through the Horizon Europe Cluster on Climate, Energy and Mobility, has been instrumental in shaping the ideas and insights presented in this work.

References

[1] C. Lai, F. Boi, A. Buschettu, and R. Caboni, "Microservices for multimobility in a smart city," *ICIW 2019: The Fourteenth International Conference on Internet and Web Applications and Services*, 2019.

[2] C. Goumopoulos, "Smart city middleware: A survey and a conceptual framework," *IEEE Access*, vol. 12, pp. 4015–4047, 2024. doi: 10.1109/ACCESS.2023.3349376

[3] E. Faliagka, E. Christopoulou, D. Ringas, T. Politi, N. Kostis, D. Leonardos, C. Tranoris, C. P. Antonopoulos, S. Denazis, and N. Voros, "Trends in digital twin framework architectures for smart cities: A case study in smart mobility," *Sensors*, vol. 24, no. 5, p. 1665, 2024. doi: 10.3390/s24051665

[4] A. Raghunandan, D. Kalasapura, and M. Caesar, "Digital twinning for microservice architectures," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023. doi: 10.1109/ICC45041.2023.10279802 pp. 3018–3023.

[5] M. Shafto, M. Conroy, R. Doyle, E. Glaessgen, C. Kemp, J. LeMoigne, and L. Wang, "Draft modeling, simulation, information technology & processing roadmap," *Technology area*, vol. 11, pp. 1–32, 2010.

[6] M. Grieves, "Digital twin: Manufacturing excellence through virtual factory replication. researchgate; unknown," 2015.

[7] J. Trauer, S. Schweigert-Recksiek, C. Engel, K. Spreitzer, and M. Zimmermann, "What is a digital twin?–definitions and insights from an industrial case study in technical product development," in *Proceedings of the design society: DESIGN conference*, vol. 1. Cambridge University Press, 2020, pp. 757–766.

[8] P. Bibow, M. Dalibor, C. Hopmann, B. Mainz, B. Rumpe, D. Schmalzing, M. Schmitz, and A. Wortmann, "Model-driven development of a digital twin for injection molding," in *International Conference on Advanced Information Systems Engineering*. Springer, 2020. doi: 10.1007/978-3-030-49435-3_6 pp. 85–100.

[9] Q. Qi, F. Tao, T. Hu, N. Anwer, A. Liu, Y. Wei, L. Wang, and A. Y. Nee, "Enabling technologies and tools for digital twin," *Journal of Manufacturing Systems*, vol. 58, pp. 3–21, 2021. doi: 10.1016/j.jmsy.2019.10.001

[10] A. F. Veloso, J. V. Júnior, M. M. d. N. Costa, R. A. Rabelo, and P. R. Pinheiro, "A comparative evaluation of monolithic and microservice architectures for load profiling services in smart grids," in *Handbook of Artificial Intelligence and Data Sciences for Routing Problems*. Springer, 2024, pp. 17–36.

[11] H. Siddiqui, F. Khendek, and M. Toeroe, "Microservices based architectures for iot systems-state-of-the-art review," *Internet of Things*, vol. 23, p. 100854, 2023. doi: 10.1016/j.iot.2023.100854

[12] M. Liu, S. Fang, H. Dong, and C. Xu, "Review of digital twin about concepts, technologies, and industrial applications," *Journal of manufacturing systems*, vol. 58, pp. 346–361, 2021. doi: j.jmsy.2020.06.017

[13] M. De Domenico, L. Allegri, G. Caldarelli, V. d'Andrea, B. Di Camillo, L. M. Rocha, J. Rozum, R. Sbarbati, and F. Zambelli, "Challenges and opportunities for digital twins in precision medicine from a complex systems perspective," *npj Digital Medicine*, vol. 8, no. 1, p. 37, 2025. doi: 10.1038/s41746-024-01402-3

[14] M. F. Jahangir, C. P. Leslie Schultz, and A. Kamari, "A review of drivers and barriers of digital twin adoption in building project development processes." *Journal of Information Technology in Construction*, vol. 29, 2024. doi: j.itcon.2024.008

[15] E. Ferko, A. Bucaioni, and M. Behnam, "Architecting digital twins," *IEEE Access*, vol. 10, pp. 50 335–50 350, 2022. doi: 10.1109/ACCESS.2022.3172964

[16] A. Imeri, C. Feltus, D. Khadraoui, N. Agoulmine, and D. Nicolas, "Solving the trust issues in the process of transportation of dangerous goods by using blockchain technology," in *Proceedings of the 11th International Conference on Security of Information and Networks*, 2018. doi: 10.1145/3264437.3264470 pp. 1–2.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 357–362
DOI: 10.15439/2025F8035
ISSN 2300-5963 ACSIS, Vol. 43

Enhancing API documentation by inter-endpoint dependency graphs

Panagiotis Papadeas
0009-0003-3012-0089
Software Engineering Lab, National Technical University of Athens. Zografos 15780, Greece
Email: papadeas@cs.ntua.gr

Dimitrios Gerokonstantis
0009-0004-0950-8414
Software Engineering Lab, National Technical University of Athens. Zografos 15780, Greece
Email: dgerokonstantis@cs.ntua.gr

Christos Hadjichristofi
0009-0000-7976-8749
Software Engineering Lab, National Technical University of Athens. Zografos 15780, Greece
Email: hadjichristofi_ch@mail.ntua.gr

Vassilios Vescoukis
0000-0002-5360-8349
Software Engineering Lab and Geoinformation Center, National Technical University of Athens. Zografos 15780, Greece
Email: v.vescoukis@cs.ntua.gr

***Abstract*—APIs are the main mechanism for integrating heterogeneous components in modern software systems. API documentation typically treats each endpoint as autonomous, leaving out information about other endpoints that produce or consume input or output data, which is useful in rapidly orchestrating valid call sequences. This makes the development, testing and maintenance of systems that use APIs to internally implement business logic, or integrate with third-party API-based services, more complex. This paper addresses this gap by introducing static and dynamic analysis methods to automatically discover and document inter-endpoint dependencies; static analysis uses existing API documentation, while dynamic analysis is based on logging API usage. Both methods are integrated into RADAR, a tool for extracting and visualizing graphs, to illustrate such inter-endpoint dependencies through exchanged data. Case study on the PayPal API demonstrates the value of adding dependency graphs to API documentation to support rapid development and documentation of software systems.**

***Index Terms*—APIs, inter-endpoint dependencies, API documentation.**

I. Introduction

IN MODERN software architectures, Application Programming Interfaces (APIs) serve as the primary mechanism for exposing the functionalities offered by entire software applications or even components, to other applications, components, or services of any kind. An API exposes a set of endpoints that, when invoked by a consumer, execute specific server-side functionality, which results in a response sent back to the client, regardless of the technologies used on either side.

APIs are developed both for integrating components within applications and as standalone software products. In either case, their complexity necessitates comprehensive documentation to support their development, maintenance, and integration. Typical API documentation usually follows standards such as OpenAPI[1], or comes in the form of de-facto formats introduced by tools, such as Postman[2] collections. Any such kind of documentation provides the schema of the request and response of every endpoint, data types and descriptions, and possibly full examples of API calls with data.

However, it usually treats each endpoint as an independent service, whereas an API is a collection of interacting endpoints that collectively can be used to implement pieces of business logic. To do this effectively, it is essential that developers have information not only about how to call each endpoint, but also about all its possible interactions with other endpoints that either consume output or provide input data. This work focuses on the discovery, modeling and representation of such interactions in a way that provides additional practically useful documentation for the integration of APIs into software components of any kind.

II. Related Work

Prior research has explored ways to facilitate the proper utilization of APIs by identifying the constraints that must be met when interacting with them. A wide area of research investigates techniques for identifying interdependencies among input parameters within (single) API calls, with the goal of facilitating the construction of syntactically and semantically valid API requests.

Oostvogels et al. [1] attempted to analyze and incorporate inter-parameter constraints within API documentation. These constraints define rules whereby the existence or valid values of certain input parameters in an API call depend on the presence or values of other parameters. To address the absence of structured representation for such constraints in existing documentation, a new specification language was created as an extension to the OpenAPI specification to incorporate this information. Martin-Lopez et al. ([2], [3], [4]) also focused on inter-parameter dependencies, proposing a Domain Specific

[1] OpenAPI: https://www.openapis.org/

[2] Postman: https://www.postman.com/

Language known as IDL (Inter-parameter Dependency Language) to describe these dependencies. Both studies focus on interdependencies of input parameters for a specific call rather than on inter-endpoint dependencies which is our area of interest.

Wu et al. [5] proposed an additional method that detects "dependency constraints on parameters", leading to the development of the INDICATOR system, which analyzes documentation and SDKs to identify constraint candidates and then validates them through testing. The INDICATOR system investigates these producer-consumer relationships based on the names and types of parameters. However, it does not focus on systematically extracting or documenting these relationships.

Another related area of research concentrates on identifying interdependencies across different operations of Web Services and APIs for testing. The work of Bai et al. [6] aims to automatically generate test cases from WSDL specifications with respect to the dependencies among various operations. One of the dependency types examined is input/output dependencies (IOD), which represent inter-endpoint dependencies. However, this approach limits the examination of inter-endpoint dependencies to aspects directly related to test case generation, which aligns with the focus of their study.

The study of Atlidakis et al. [7] introduces REST-ler, a tool that analyzes Swagger specifications for RESTful APIs, automatically identifies object types required by a request that have been produced by previous ones, generates tests with respect to these dependencies and dynamically infers which request sequences are valid. However, this inter-endpoint analysis aims mainly to automatically generate and execute meaningful and correct test cases, while adhering to these dependencies. Consequently, no documentation regarding the inter-endpoint dependencies is generated.

The study conducted by Bertolino et al. [8] aims to extract behavioral information regarding how users should invoke a Web Service. This method, called StrawBerry, examines the WSDL document and identifies dependencies between the operations of the Web Service based solely on types of input and output parameters. A behavior protocol automaton is generated, indicating the dependencies and data transfers among various operations of the Web Service. However, this analysis relies on the presence of a WSDL specification.

III. Inter-Endpoint Dependencies

An "inter-endpoint dependency" is an interaction between API endpoints. If invoking endpoint Y requires input data received from a call to endpoint X, then the required call order is (X, Y) and thus we say that Y "depends" on X. As an example, consider the hypothetical Flight Booking API illustrated in Fig 1, along with a sample invocation of its endpoints. The top-level nodes represent API endpoints, while the corresponding JSON response schemas and the example responses appear beneath each endpoint. Directed edges indicate the data flow between endpoints.

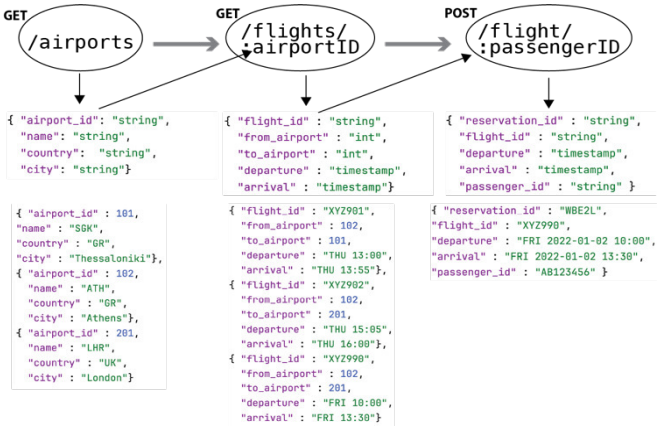


Fig 1. Production and consumption of data in API call sequences

The call to the /airports endpoint returns a list with all available airports, including their identifiers (airport_id). The following call to /flights takes as input, as a path parameter, the id of one of these airports (102, as shown in the example) and returns information about available flights departing from that airport. This is a body-path dependency, as a data element from the response body of the first call is used as a path parameter in the second call.

Subsequently, the /flight endpoint call takes as input, as the request body this time, all the details of one of the flights returned by the second call (the one with id=XYZ990, in the example) along with a path parameter that points to a passenger and performs the booking for that flight and passenger. This represents a body-body dependency, as the request body of the third call requires a data object included in the response body of the second call. Similarly, in body-query dependencies the value of a query parameter comes from a value contained in some attribute of a response body from a previous call. Notably, inter-endpoint dependencies can be identified by more than one attributes.

Examining inter-endpoint dependencies goes beyond typical API documentation by revealing the correct sequences and prerequisites of API calls. This knowledge eases the implementation of requests with valid inputs and helps discover entire paths that indicate the semantically correct order of endpoint invocation. As such, it can accelerate integration with third-party APIs by identifying valid call sequences. In microservice architectures, it also supports—and can even partially automate—the orchestration control flow. Finally, for API developers, it serves as a valuable reference for understanding data flow, maintaining consistency, and onboarding new team members.

This work focuses on the discovery and documentation of API inter-endpoint dependencies by introducing algorithms as well as a tool to discover, analyze, and visualize them using a directed graph, whose nodes represent the API endpoints and each directed edge represents the complete or partial consumption of data returned by the origin, by the destination endpoint.

IV. Discovery of Dependencies

Two approaches are demonstrated for discovering and analyzing inter-endpoint dependencies: static and dynamic, introduced in the diploma theses of Papadeas [9] and Gerokonstantis [10], respectively. The static approach is based on existing API documentation, whereas the dynamic captures and analyzes API calls made during the real-world utilization of the API.

4.1 Static analysis

Static analysis is entirely based on the information included in the API documentation. It aims to identify inter-endpoint dependencies using this documentation (structure, types), which typically contains details about the input and output parameters of each endpoint. The goal is to identify dependencies and generate a directed graph that shows all possible endpoint call sequences.

Static analysis takes its input from any standard API documentation, such as OpenAPI, or even from ad-hoc documentation such as Postman, Apidog[3], Firecamp[4] or other similar tools. It infers the proper sequence of API calls by analyzing the results of pre-executed calls included in the documentation examples. Lack of pre-executed calls and insufficient examples reduces the capability of this analysis to discover dependencies.

In our implementation, static analysis takes its input from a Postman collection to create a representation of the directed dependency graph. The Postman collection comes in as a JSON file that contains information about responses, request body, path and query parameters of each endpoint. The use of Postman collections has been chosen for illustration purposes only: any structured parsable object containing the same data can be used instead, and the results will still rely on the completeness of metadata included in the documentation.

A typical context where static analysis can be performed, is when we have some kind of documentation about an API, and we don't know anything about the proper sequence of endpoint calls or possible use cases of the API. The graph produced by the static analysis helps us get a first view of the API usage flow, as it shows how endpoints relate to each other.

The algorithm of the static analysis takes as input a postman collection file (Algorithm 1). For every endpoint of the API it stores information about its parameters including name, value and type of each parameter in a set named ParametersSet (lines 1-5 Algorithm 1). It also stores the response of each endpoint in a separate set named AttributesSet including the name, value and type of each attribute in the response (lines 6-10 Algorithm 1). Then it proceeds to compare the two sets to discover dependencies (lines 11-12 Algorithm 1).

Each dependency can be identified either by matching the values of parameters with the values of responses exchanged between endpoints in call examples (Algorithm 2), or by matching keys instead (Algorithm 3).

The algorithm compares the two sets to find every pair of attributes and parameters that have the same value (or name, respectively), type and are not part of the same endpoint (lines 1-4 Algorithm 2 and Algorithm 3 respectively). For each case where a match has been identified, a new edge is added to the graph, along with the relevant information regarding the dependent endpoints (lines 5-8 Algorithm 2, 3).

Algorithm 1. Static inter-endpoint dependency analysis

```
 1. Read postman_collection.json as Input
 2.
 3. foreach endpoint in Input do
 4.     foreach parameter in endpoint do
 5.         ADD(parameter(name, value, type)) TO SET
            named ParametersSet
 8.
 9. foreach endpoint in Input do
10.     foreach response in endpoint do
11.         ADD(attribute(name, value, type)) TO SET
            named AttributesSet
14.
15. DependencyGraphByValues(ParametersSet, AttributesSet)
16. DependencyGraphByNames(ParametersSet, AttributesSet)
```

Algorithm 2. DependencyGraphByValues

```
1. Procedure DependencyGraphByValues
2.
3. foreach parameter in ParametersSet do
4.     foreach attribute in AttributesSet do
5.         if parameter.value == attribute.value AND
              parameter.type == attribute.type AND
              endpoint(parameter) != endpoint(attribute)
           then
6.            ADD(dependency(endpoint(parameter),
                      endpoint(attribute)))
              TO LIST named Dependencies
7.
8. return Dependencies
```

Algorithm 3. DependencyGraphByNames

```
 1. Procedure DependencyGraphByNames
 2.
 3. foreach parameter in ParametersSet do
 4.     foreach attribute in AttributesSet do
 5.         if parameter.name == attribute.name AND
 6.            parameter.type == attribute.type AND
 7.            endpoint(parameter) != endpoint(attribute)
           then
 8.                ADD(dependency(endpoint(parameter),
                           endpoint(attribute)))
                   TO LIST named Dependencies
12.
13. return Dependencies
```

The second way of dependency identification, by comparing only the keys of parameters, has been proven to be useful in cases where the collection does not contain many examples of pre-executed calls and there isn't enough information regarding each endpoint. In that case, the only way a dependency could be inferred is by comparing keys between parameters and responses assuming that common keys could also imply dependent values.

Static analysis will usually discover a greater-than-actual number of inter-endpoint dependencies due to common key-value patterns across endpoints. For example, the frequent use of generic keys like "ID" can falsely suggest dependencies between unrelated endpoints. These findings are often hard to evaluate without access to actual API calls or a good understanding of the API's semantics. The goal of this analysis is to

[3] Apidog: API Development Tool - https://apidog.com

[4] Firecamp: Open Source Postman Alternative - https://firecamp.io

identify all possible dependencies, including dependencies which are considered “noise”.

Dynamic analysis

The dynamic analysis takes a different approach, aiming to discover inter-endpoint dependencies without using any kind of documentation. It collects information from HTTP requests and responses, by monitoring actual calls using a "man-in-the-middle" agent that captures HTTP traffic. The API traffic is logged and then the extraction of inter-endpoint dependencies is possible. Notably, this method follows a reverse engineering approach, analyzing already implemented but still undocumented APIs, to infer inter-endpoint dependencies.

Dynamic analysis also identifies all types of dependencies, just as static analysis does. However, in the case of body-path dependencies, since it is unclear which segments of a URL correspond to path parameters, the algorithm treats all URL segments as such. This allows it to identify the corresponding dependencies, introducing minimal noise.

In cases where the API is consumed through a web frontend, API calls can be captured by executing use cases while interacting with the frontend in a browser. In the context of an API ecosystem implementing orchestration, as in a microservices architecture, it is possible to capture API calls as part of the interaction between the orchestrator component and the internal components of the architecture.

In any of the approaches mentioned above, to capture the HTTP traffic resulting from real-world interactions with the API, a "mediator" is required, that can be implemented in several ways, such as the following:

1. a standalone mediator component to which API calls would be directed, responsible for forwarding the requests to the API server, returning the responses to the client, and recording information about the API calls; such a system was implemented in the context of this paper,

2. a browser extension that captures the incoming and outgoing traffic between the browser and the API while the user navigates through the User Interface interacting with the API,

3. an internal component within API Gateways, enabling the collection of multi-client data derived from the interactions of multiple external entities with the API.

Similar to the static method, the outcome of the dynamic analysis is a directed dependency graph of the same type. The input is a log file generated by the mediator, which contains information about the executed API calls.

The dynamic analysis algorithm comprises two main components: the parsing of the log file and extraction of value-specific information (Algorithm 4), and the identification of matches among the extracted values (Algorithm 5). Specifically, Algorithm 4 processes the input log file, assigns a sequence number to each API Call reflecting its execution order within the use case it belongs to, and constructs two dictionaries, REQUEST_VALUES and RESPONSE_VALUES, which store information about every value encountered in API request and response messages respectively. For each identified value, the dictionaries record the associated API call’s URL and sequence number, the identifier ("tag") of the corresponding use case, the value’s name, and its inferred data type.

The dictionaries constructed in Algorithm 4 serve as inputs to Algorithm 5, which is responsible for identifying matches between values appearing in request and response messages. Specifically, for each value V found in the response body of an API call RES, the algorithm examines whether V also appears as a request parameter (of any type) in any subsequent API calls (REQ_1, REQ_2, ..., REQ_n) within the same use case.

If such matches are found, the algorithm adds a directed dependency edge from RES to each matching REQ_i, provided the following conditions are met:

- the data type of V matches in both RES and REQ_i, as determined by type inference,
- RES and REQ_i correspond to distinct API endpoints, and
- REQ_i occurs after RES in the execution order of the use case.

Algorithm 4. Dynamic inter-endpoint dependency analysis

```
1. Initialization: Assign sequence numbers to the
   calls of every use case
2.
3. foreach call in Input do
4.     if call.body then
5.         foreach parameter in call.body do
6.             REQ_VALUES[parameter.value].APPEND({
7.               url: call.url, seq: call.sequence_number,
8.               tag: call.tag, paramName: parameter.name,
9.               paramType: parameter.type})
10.
11.    if call.response then
12.        foreach attribute in call.response do
13.            RES_VALUES[attribute.value].APPEND({
14.              url: call.url, seq: call.sequence_number,
15.              tag: call.tag, attrName: attribute.name,
16.              attrType: attribute.type})
17.
18. ComputeDependencyGraph(REQ_VALUES, RES_VALUES)
```

Algorithm 5. ComputeDependencyGraph

```
1.   Procedure ComputeDependencyGraph
2.     foreach value in RES_VALUES do
3.       if value in REQ_VALUES then
4.         foreach appearRes in RES_VALUES[value] do
5.           foreach appearReq in REQ_VALUES[value] do
6.             if appearRes.paramType ==
                  appearReq.paramType and
7.                appearRes.url ≠ appearReq.url and
8.                appearRes.seq < appearReq.seq and
9.                appearRes.tag == appearReq.tag
10.            then
11.            DEPENDENCY_GRAPH[
                      (appearRes.url, appearReq.url)
               ].APPEND({
12.              fromAttribute: appearRes.paramName,
13.              toAttribute: appearReq.paramName
14.            })
```

A key difference from the static method is that dynamic analysis restricts comparisons to subsequent API calls of the same use case, focusing on the dependencies most likely to be semantically correct within the context of the specific API.

While dynamic analysis can generate dependency documentation even for completely undocumented APIs, it also has several challenges, with the primary one being that the discovery of all dependencies requires the execution of calls between all dependent endpoints, which may not occur during the logging.

For visualizing the resulting dependency graph, a tool named RADAR[5] (REST API Dependencies and Analysis of Relationships) was implemented [11] and is available at radar.softlab.ntua.gr. RADAR takes as input either the Postman Collection of the API or a log file containing records of actual API calls, for static and dynamic analysis, respectively. After the user configures certain analysis parameters, RADAR performs either static or dynamic analysis to detect inter-endpoint dependencies, and generates a directed dependency graph, as discussed above.

V. Case Study

The two dependency analysis methods were applied, among others, to the PayPal API sandbox environment[6].

The PayPal API provides a wide range of functionalities, including the creation and management of products, orders, invoices, subscriptions, payments, and disputes. The complexity and size of such an API necessitate conducting dependency analysis to facilitate and guide its utilization. Below, we examine how static and dynamic inter-endpoint dependency analysis can assist a user of this API in creating and tracking a product order.

Static

For static analysis, we supply RADAR with the Postman Collection of the PayPal API[7], as downloaded from the Postman Public API Network, without any modifications. The section of the resulting dependency graph including endpoints related to product orders is shown in Fig 2 below. While the identified dependency path facilitates the process of creating an order and handling the payment, it lacks details regarding order tracking. Static analysis was unable to identify dependencies between the tracking endpoints and the others in the graph, as the examples in the Postman Collection were not adequate. Additionally, the static analysis graph contains some edges that are either redundant for executing the specific use case or even semantically incorrect.

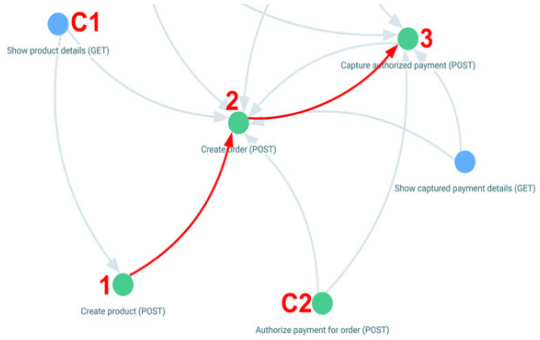


Fig 2. Order creation and tracking use case - Static dependency graph

For instance, displaying product details before creating a new product (Fig 2, edge C1→1) offers no practical value in the context of the "order management" use case, while authorizing a payment for an order before the order itself has been created, as the graph suggests (Fig 2, edge C2→2), seems not correct and should rather be classified as noise. As demonstrated below, dynamic analysis mitigates these issues.

Dynamic

Dynamic analysis involves logging interactions with the API relevant to product orders. To do this, we set up an example use case, in which we begin by creating/updating a product and generating an order for this. After that, we initiate the payment process by authorizing and capturing the payment, and conclude by adding tracking information for the order. This sequence is illustrated in Fig 3.

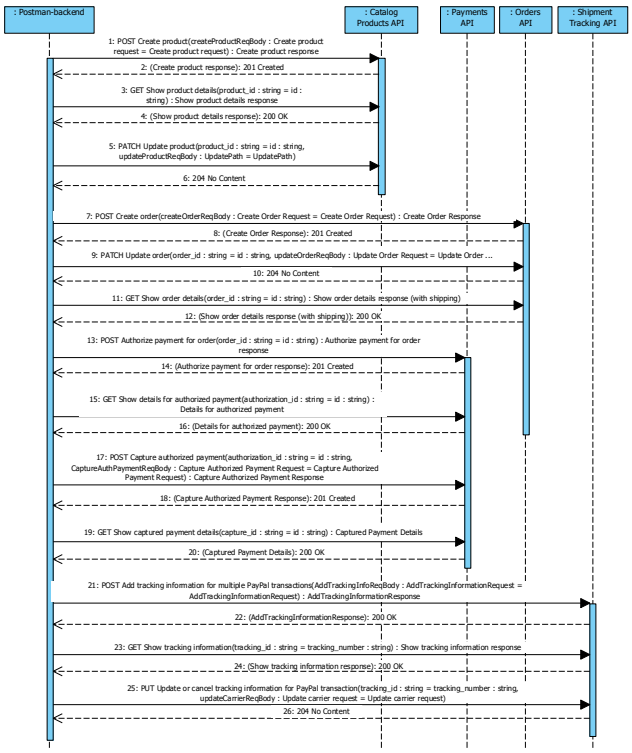


Fig 3. Order creation and tracking use case - sequence diagram

The dependency graph resulting from the dynamic analysis of the log file for this use case is shown in Fig 4, where API calls are enumerated based on their numbering in the sequence diagram above, and a key dependency path is highlighted.

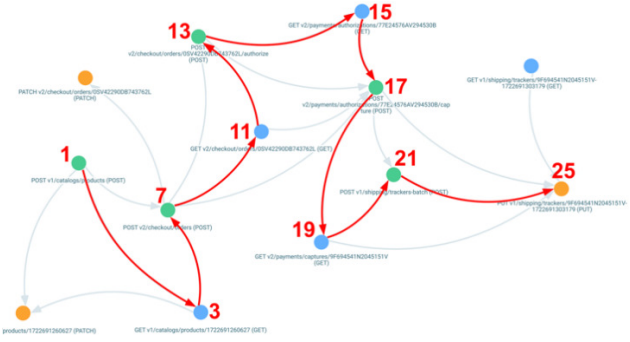


Fig 4. Order creation and tracking use case - dynamic dependency graph

[5] R.A.D.A.R by NTUA Softlab: https://radar.softlab.ntua.gr
[6] PayPal REST APIs: https://developer.paypal.com/api/rest (date accessed: 05/2024)
[7] PayPal API Postman Collection (date accessed: 05/2024): https://www.postman.com/paypal/paypal-public-api-workspace/overview

The highlighted dependency path can be traced back to the reference use case sequence diagram, providing the API user with guidance on how to use the API to implement the logic of the "order management" use case.

Specifically, the generated graph guides the user to first create a product and optionally retrieve information about it (calls 1, 3), then create an order for this product and possibly retrieve relative details (calls 7, 11), authorize and capture the payment (calls 13–19), and finally add and potentially update tracking information (calls 21, 23). In contrast to static analysis, the dynamic dependency graph proved to be more comprehensive while avoiding misleading edges.

VI. Discussion

Like all software documentation, API documentation must be complete, detailed, and kept up to date to support communication among technical staff and stakeholders. As APIs play a central role in software integration, better documentation leads to more efficient development and fewer errors and backtracks.

This work aims to enhance API documentation by adding information on data producers and consumers among endpoints, representing their dependencies. Such an "endpoint dependency graph" helps developers better understand and integrate complex third-party or in-house APIs. We propose two methods to identify these dependencies based on available documentation and runtime access, providing a roadmap to accelerate API integration.

The effectiveness of static analysis in identifying dependencies relies on the information in the API's static documentation, particularly the examples included. On the other hand, the effectiveness of dynamic analysis in the same task depends on the extent to which the executed use cases cover a broad range of the API's functionalities, theoretically all of them.

Generally, the dynamic approach produces less, yet not zero, noise, particularly in body-path dependencies derived from URL segments that are not actually path parameters. In static analysis, which runs exhaustive comparisons between every pair of endpoints, noise is created by ambiguities that do not allow to filter out irrelevant dependencies.

Overall, considering the statistics derived from the results of the two analyses across several APIs—including the OpenAI API and the previously examined PayPal API—we observe that dynamic analysis detects fewer edges in the generated graph: 52 vs. 92 for OpenAI and 209 vs. 724 for PayPal. Furthermore, dynamic analysis identifies more dependent attributes for each edge, offering a deeper understanding of how two endpoints are interdependent. For example, for the OpenAI API, an average of 1.22 dependent attributes per edge was found in the static analysis and 2.53 in the dynamic analysis, while for the PayPal API, 2.13 dependent attributes per edge were found in the static analysis and 14.19 in the dynamic analysis.

Considering the limitations of each method, as well as the conclusions drawn from their comparative study and application to the API examined, it becomes clear when each method is most suitable: when any static documentation such as a Postman Collection with enough examples is available, dependency documentation for the API can be generated using static analysis. When logging is possible, dynamic analysis can offer higher accuracy and completeness of the results in the context of the observed use cases, but at the cost of greater effort that comes with logging. RADAR supports both methods and although it has space for improvement, it produces useful data and visualizations of inter-endpoint API dependencies.

References

[1] N. Oostvogels, J. De Koster, and W. De Meuter, "Inter-parameter Constraints in Contemporary Web APIs," 2017, pp. 323–335. doi: 10.1007/978-3-319-60131-1_18.

[2] A. Martin-Lopez, "Automated analysis of inter-parameter dependencies in web APIs," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings*, New York, NY, USA: ACM, Jun. 2020, pp. 140–142. doi: 10.1145/3377812.3382173.

[3] A. Martin-Lopez, S. Segura, C. Muller, and A. Ruiz-Cortes, "Specification and Automated Analysis of Inter-Parameter Dependencies in Web APIs," *IEEE Trans Serv Comput*, vol. 15, no. 4, pp. 2342–2355, Jul. 2022, doi: 10.1109/TSC.2021.3050610.

[4] A. Martin-Lopez, S. Segura, and A. Ruiz-Cortés, "A Catalogue of Inter-parameter Dependencies in RESTful Web APIs," in *Service-Oriented Computing: 17th International Conference, ICSOC 2019, Toulouse, France, October 28–31, 2019, Proceedings*, Berlin, Heidelberg: Springer-Verlag, 2019, pp. 399–414. doi: 10.1007/978-3-030-33702-5_31.

[5] Q. Wu, L. Wu, G. Liang, Q. Wang, T. Xie, and H. Mei, "Inferring dependency constraints on parameters for web services," in *Proceedings of the 22nd international conference on World Wide Web*, in WWW '13. New York, NY, USA: ACM, May 2013, pp. 1421–1432. doi: 10.1145/2488388.2488512.

[6] Xiaoying Bai, Wenli Dong, Wei-Tek Tsai, and Yinong Chen, "WSDL-Based Automatic Test Case Generation for Web Services Testing," in *IEEE International Workshop on Service-Oriented System Engineering (SOSE'05)*, IEEE, 2005, pp. 215–220. doi: 10.1109/SOSE.2005.43.

[7] V. Atlidakis, P. Godefroid, and M. Polishchuk, "RESTler: Stateful REST API Fuzzing," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, IEEE, May 2019, pp. 748–758. doi: 10.1109/ICSE.2019.00083.

[8] A. Bertolino, P. Inverardi, P. Pelliccione, and M. Tivoli, "Automatic synthesis of behavior protocols for composable web-services," in *Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, in ESEC/FSE '09. New York, NY, USA: ACM, Aug. 2009, pp. 141–150. doi: 10.1145/1595696.1595719.

[9] P. Papadeas, "Automatic generation of dependency documentation between endpoint calls of a REST API," National Technical University of Athens, 2023. Accessed: Apr. 27, 2025. [Online]. Available: http://artemis.cslab.ece.ntua.gr:8080/jspui/handle/123456789/18728

[10] D.-D. Gerokonstantis, "Dynamic Analysis of Inter-endpoint Dependencies in RESTful APIs," National Technical University of Athens, 2024. Accessed: Apr. 27, 2025. [Online]. Available: http://artemis.cslab.ece.ntua.gr:8080/jspui/handle/123456789/19419

[11] D. Lalias, "Web application for visualizing dependencies between endpoints of a REST interface," National Technical University of Athens, 2024. Accessed: Apr. 27, 2025. [Online]. Available: http://artemis.cslab.ece.ntua.gr:8080/jspui/handle/123456789/19020

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 363–368
DOI: 10.15439/2025F5576
ISSN 2300-5963 ACSIS, Vol. 43

Hybrid U-Net segmentation of vessels in fundus eye images

Lesław M. Pawlaczyk
WSB Merito University
in Chorzow, Poland
ul. Sportowa 29, 41-506 Chorzów, Poland
Email: leslaw.pawlaczyk@chorzow.merito.pl

***Abstract*—We present a novel multi-stage method for colour image segmentation, with a primary focus on vessels segmentation in retinal fundus images, using a U-Net based architecture. Our approach tackles challenges posed by varying image resolutions through a coarse-to-fine segmentation pipeline. It begins with a rough segmentation at varying scales, guided by a traditional CNN and progressively refines results to find a target resolution. It culminates with detailed segmentation at a target scale with a smaller window sliding step, compared to previous stages. We train and validate our method using four publicly available datasets FIVES, DRHAGIS, HRF, and STARE - and demonstrate superior performance compared to traditional sliding window techniques. Notably, our model achieves high accuracy with relatively few training images. The entire framework is open-sourced and adaptable to a wide range of image segmentation tasks.**

***Index Terms*—U-Net, segmentation, deep learning, vessel segmentation, medical imaging, fundus images**

I. Introduction

THE process of image segmentation has been for many years a challenge, that was only partially solveable. This changed with an advent of deep learning and introduction of convoutional neural networks. In the year 2012 Krizhevsky et al [1] introduced ImageNet, which achieved a new level of image recognition accuracy. It won the ImageNet competition that year, and inspired a plethora of other researchers to improve upon it. This milestone in image classification allowed a development of new class of segmentation algorithms.

One of the challenges in research was to segment biomedical images. It was partially solved by Ronneberger et al. [2] by introduction of *U-Net* architecture. The *U-Net* is designed to work with very few training images and to yield more precise segmentations. It consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. However it comes with drawbacks, such as sensitivity to input image resolution and segmentation window position, and high demand for GPU memory which is limiting the segmentation window size (*WS*).

In this article, we address challenges in image segmentation by proposing a multi-step framework for training a *U-Net* adaptable to varying image resolutions. We evaluate segmentation results at multiple scales to select the most effective output and analyze how resolution affects performance when the model is applied at different scales. We also assess the annotation effort required to achieve satisfactory results. Experiments on four public vessel segmentation datasets enable us to test how well a *U-Net* trained on one dataset generalizes to others.

The article has six chapters. The first introduces the topic. The second reviews recent work on *U-Net* segmentation in medical imaging. The third explains our approach with illustrations. The fourth describes our experiments and results. The fifth highlights key findings. The final chapter states there are no competing interests.

The main contributions of this work are:

- A multi-scale rough segmentation step that processes images at multiple predefined scales using a sliding window approach.
- A *CNN*-based scoring mechanism that selects the optimal segmentation scale for each image based on predicted segmentation quality.
- Experimental validation of cross-dataset generalization, demonstrating that models trained on one dataset can perform well on others.
- An analysis of how the number of training images affects segmentation quality, providing insight into the annotation effort required for satisfactory performance.
- An open-source implementation of the method, available on GitHub [3].

II. Related works

Ren et al. [4] proposed an improved *U-Net*-based method for retinal vessel image segmentation, enhancing the original architecture to achieve greater accuracy and robustness. Their improvements include modifications to both the network structure and the training process. Liu et al. [5] introduced a three-path *U-Net* model for retina image segmentation, which leverages multiple pathways to capture features at different levels, resulting in improved segmentation performance. Das et al. [6] evaluated the performance of the standard *U-Net* for retinal blood vessel segmentation, confirming its effectiveness for this task. Yun et al. [7] proposed a Multi-Path Recurrent *U-Net* for segmenting retinal fundus images, incorporating both multiple paths and recurrent units to enhance feature extraction. Similarly, Huang et al. [8] developed an improved *U-Net* architecture based on residual modules, achieving increased

Topical area: Advanced Artificial Intelligence in Applications

robustness and better segmentation results in retinal vascular images.

Many existing studies aim to enhance the performance of *U-Net* across a range of applications. A common theme among these efforts is the focus on segmenting disconnected or localized objects [9], often overlooking use cases like vessel segmentation, where structures are continuous and span the entire image. Additionally, relatively few works address the challenge of handling images with varying resolutions - a critical factor in real-world medical imaging scenarios. Notably, many proposed improvements to *U-Net* result in only marginal performance gains, typically in the range of 1–2% absolute accuracy, as the original *U-Net* already performs strongly, often achieving over 95% accuracy under favorable conditions.

III. Methodology

In recent years, numerous high-quality datasets for fndus eye images (*FEI*) were published. We selected four publicly available datasets for our experiments: *FIVES* [10], consisting of 1,200 images with a resolution of 2048×2048 pixels; *DRHAGIS* [11], containing 80 images at 4752×3168 pixels; *HRF* [12], which includes 45 images at 3504×2336 pixels; and *STARE* [13], from which we used 20 images at 700×605 pixels resolution.

A. Training data preparation

In preparing data for *U-Net* training, it is essential to extract square windows (ideally with dimensions of 2^n) from the input images. These windows are then fed into the training pipeline to iteratively refine the model's performance. Although *U-Net* is known for its efficiency and accuracy under favorable conditions, a significant challenge lies in determining which parts of the image should be selected for training.To address this challenge, several key questions must be considered:

- Which locations should be chosen to ensure that windows cover both object and background regions effectively?
- How many windows should be sampled from each image?
- How many of these windows should predominantly contain the object of interest, and how many background?

Below, we present our approach to answering the forementioned questions by constructing a mechanism for automatic data generation. We use the *FIVES* dataset [10] as the primary input for training the *U-Net* model. The original dataset, consisting of 1,200 images, is divided into two subsets. The first subset, referred to as the *U-Net* Main Training Set (*UNETMTS*), includes 900 images and is used for model training. The second subset, called the *U-Net* Segmentation Validation Set (*UNETSVS*), comprises the remaining 300 images and is used to validate segmentation performance.

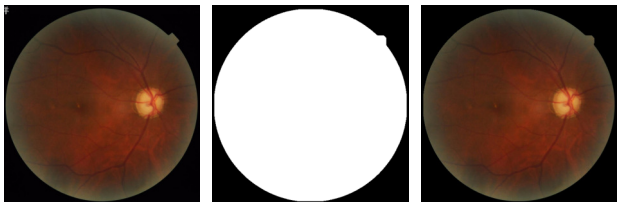
1) Mask finding: Each image is processed as follows:

1) Convert from RGB to grayscale.
2) Compute a threshold:

$$\text{threshold} = \frac{\text{mean}(\text{grayscale_image})}{3} - 5$$

and binarize to generate a mask.
3) Apply flood fill from all four corners to produce an additional mask.
4) Combine the two masks by pixel-wise summation.
5) Apply erosion and dilation with a 20×20 kernel.

Finally, the mask is applied via logical AND operation. Pixels within its outer layer are blacked out; others remain unchanged (see Fig. 1).



(a) Input image (b) Mask (c) Masked image

Fig. 1. Image Masking Results

After the mask is generated, the image is trimmed from the left, right, top, and bottom edges to ensure that no completely black vertical or horizontal lines remain.

2) Random windows: The initial set, *UNETMTS*, is reduced to a specified value referred to as *TRDTSZ* (training data size), which is initially set to 900. However, this value is varied and analyzed in later sections of this article (see Section IV) to investigate whether the full set of 900 images is necessary, or if smaller subsets can still yield high segmentation accuracy. The goal is to investigate how much training data the *U-Net* model requires to achieve optimal results. Once the value of *TRDTSZ* is selected, *UNETMTS* is randomly divided into two subsets: a training set (*DPSTR*) and a test set (*DPSTS*), using a 75% to 25% split.

From each image, a set of training windows is randomly sampled. Two types of windows are defined: object-type windows (*OBTW*) and background-type windows (*BKTW*). A window is classified as *OBTW* if the number of pixels corresponding to the ground truth label exceeds a minimum threshold (e.g., 15%). Conversely, a window is labeled as *BKTW* if the proportion of object pixels does not exceed a maximum threshold (e.g., 5%). We also define target ratio of object to background windows per image (e.g., 50%), to ensure a representative sampling that reflects typical pixel distribution between object and background regions in *FEI*.

The underlying motivation for this windowing strategy is to balance object and background representation during training. This balance is reinforced by employing a weighted binary cross-entropy loss function in the *U-Net* implementation, which helps to address class imbalance. The size of each sampled window (*WS*) is typically set to either 64×64 or 128×128 pixels. Input images are rescaled after mask generation to 512×512 pixels. Larger values of *WS* and image resolution significantly increase GPU memory consumption - an important consideration given that our hardware setup was a GPU with only 8GB of memory.

The percentage values used in this step are based on intuition and can be adjusted depending on the type of images, particularly when the ratio of object pixels to background pixels varies. For instance, these values would differ when segmenting structures such as the eye cup (*EC*) or optic disc (*ED*) in other *FEI* datasets. The goal of this strategy is to capture a wide range of spatial arrangements of the target object within randomly selected windows, thereby enriching the training dataset for the *U-Net* model used in later stages of the system.

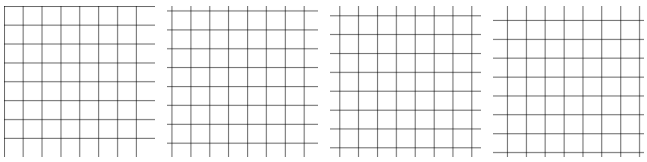
B. U-Net network training

The windows extracted in Section III-A are grouped into training batches, with a predefined number of epochs and a selected (*WS*) that matches the input dimensions of the *U-Net* model. The input windows, referred to as *X* windows, consist of RGB channels, with pixel values normalized to the range [0, 1]. The corresponding output windows *Y* windows, represent the segmentation masks, and are converted into binary image.

C. U-Net preliminary multi resolution segmentation

This step begins with a rough segmentation step (*RSS*). Input image is scaled to a resolution of 512×512, then segmented using a sliding window technique. Its primary goal is to produce preliminary segmentation results which are passed to the next stage Section III-D, where a dedicated *CNN* model is trained how to evaluate segmentation result.

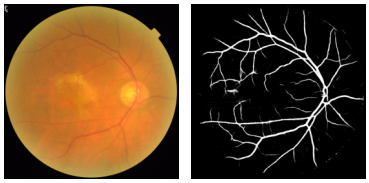
1) Windows mesh creation: Once the image has been prepared for segmentation - by generating a mask as described in Section III-A1 - it typically contains a black border surrounding the scene, with the retina centered. This preprocessed image is then analyzed across three color channels: red, green, and blue. Each color channel is divided into a collection of meshes (i.e., image patches) by sliding a window across the image. Mesh extraction begins with an offset of 0 and proceeds up to $WS - 1$, using a predefined sliding step denoted as *SS*. While it is not required for $WS \bmod SS = 0$, it is essential that $SS < WS$ to ensure overlapping patches and continuous coverage of the entire image. The idea of meshes is shown in Fig. 2



(a) Offset 0 (b) Offset 32 (c) Offset 64 (d) Offset 96

Fig. 2. Sample meshes generated for an image of size 1024×1024 pixels, *WS* of 128×128 pixels, and *WSS* set to 32

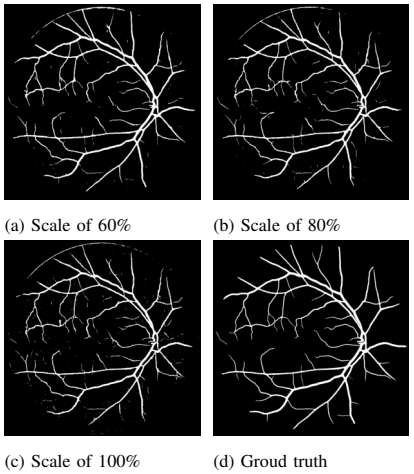
Each mesh allows the input image to be segmented using a different window configuration. This approach is crucial for capturing diverse spatial variations within the image, as each mesh is processed independently by the trained *U-Net* model. For each mesh, the *U-Net* generates a probability map, assigning to each pixel the likelihood of belonging to the object or background. Due to overlapping meshes, a single pixel may appear in multiple segmentations, resulting in multiple probability estimates. These values are averaged to generate a heatmap, which represents the overall likelihood of each pixel being part of the object. In this heatmap, brighter pixels correspond to higher confidence in object presence. Finally, the heatmap is binarized using a fixed threshold of 127, producing a black-and-white segmentation mask. An example of this process is shown in Fig. 2, where four meshes contribute to the heatmap visualized in Fig. 3. This method is referred to as the Sliding Window Algorithm (*SWA*).



(a) Input image (b) Pre binarization heatmap

Fig. 3. Input image from [10] along with generated heatmap

2) Data preparation for CNN training: As described in Section III-A, the main dataset [10] was partitioned to include a subset named *UNETSVS*, which is used for training the *CNN* in the subsequent evaluation stage. All images in this subset are processed through the (*RSS*). We perform segmentation at multiple scales, ranging from 20% to 100% of the original image size, in increments of 10%. This produces a series of black-and-white segmentation outputs for each image at varying resolutions. Figure 4 presents the segmentation results for three selected scales: 60%, 80%, and 100%. It is apparent from the visual comparison that the quality of the segmentations varies significantly across different scales.



(a) Scale of 60% (b) Scale of 80%

(c) Scale of 100% (d) Groud truth

Fig. 4. Different scales of input image for *CNN* segmentation assessment training

To objectively assess which of the segmentation results is most optimal, we design and train a simple *CNN*-based scoring network. This network consists of a single output neuron that produces a score between 0 and 1, where 0 indicates the

poorest segmentation quality and 1 indicates the best (see Fig. 7). The trained *CNN* is employed in the final segmentation stage, as described in Section III-E.

D. CNN training for segmentation results assessment

Using the *UNETSVS* set, we segment each image into nine black-and-white binarizations, producing a total of $300 \times (9 + 1) = 3000$ samples for training the *CNN*. The additional term in the formula accounts for the ground truth image (*GT*), which serves as the expected output from the segmentation algorithm. This ground truth image is also used to compute the Jaccard index [14], which measures the similarity between the segmented output and its corresponding *GT*. The index yields a value between 0 and 1, where 0 indicates a completely incorrect segmentation and 1 indicates a perfect match. The *CNN* architecture used for evaluation consists of a simple stack of convolutional layers, followed by batch normalization and pooling layers. Once the *CNN* scorer is trained, it can be used to evaluate and select the most accurate segmentation result. This final decision-making step is carried out as described in Section III-E.

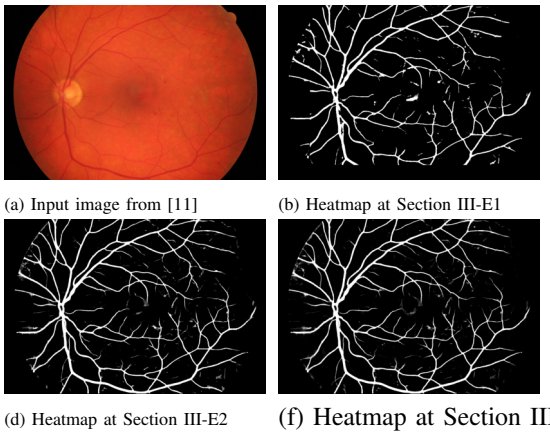
E. Optimal segmentation step

The final segmentation comprises of three stages.

1) Rough stage of final segmentation: The final segmentation algorithm builds upon the approach described in Section III-C. We continue to perform the Rough Segmentation Step (*RSS*), but in this phase, segmentation is carried out at scales ranging from 20% to 100%, with a step size of 20%. From the five resulting segmentation candidates, the one deemed most accurate by the trained *CNN* (introduced in Section III-D) is selected as the final output. The scale at which this optimal segmentation occurs is referred to as the *RSOR*. It is important to note that for smaller images (up to 800×800 pixels), we increase the maximum scale to 150%. This adjustment accounts for the fact that *U-Net* may underperform when operating on very low-resolution inputs.

2) Detailed stage of final segmentation: The *RSOR* ratio plays a crucial role, as it serves as the starting point for the subsequent stage, known as the Detailed Segmentation Step (*DSS*). This step further refines the search for the optimal segmentation scale. Specifically, segmentation is performed at scales ranging from $RSOR - 10\%$ to $RSOR + 10\%$, using a step size of 4%. As before, the trained *CNN* from Section III-D is used to evaluate the quality of each segmentation. The scale identified as producing the best result is referred to as the *DSOR*.

3) Most detailed stage of final segmentation: With the *DSOR* ratio now established, we proceed to the Most Detailed Segmentation Step (*MDSS*). In this final stage, the input image is rescaled according to the selected *DSOR* value. Unlike in Section III-C, where the window sliding step (*SS*) was set to 64, we now adopt a finer step size of 32 to enhance segmentation precision. The results of this final segmentation step, applied to a sample image from the *DRHAGIS* dataset [11], are shown in Figure 5. It is important to note that both the binarized segmentation outputs and the probability heatmaps, originally generated at the scaled resolution, are rescaled back to the original image dimensions for consistency and visualization.



(a) Input image from [11] (b) Heatmap at Section III-E1

(d) Heatmap at Section III-E2 (f) Heatmap at Section III-E3

Fig. 5. Input image from [11] along with generated heatmaps at 3 different stages with *SS* at 128, 64 and 32

All configuration parameters associated with window size (*WS*), sliding step (*SS*), and the scale percentages used for segmentation are fully adjustable by the user of the implementation provided in [3]. In Section IV, we present a series of experiments evaluating segmentation accuracy using the Jaccard Index across a dataset composed of three sources: *DRHAGIS* [11], *HRF* [12], and *STARE* [13]. This dataset is referred as the Combined Validation Dataset (*CVDS*).

IV. Experiments

In our experiments, we aimed to evaluate how well the proposed approach generalizes when a *U-Net* model trained on one dataset is applied to a completely different dataset.

We conducted a total of three experiments:

1) In the first experiment, both the *U-Net* and *CNN* models were trained on 100 randomly selected images from the *FIVES* dataset [10]. The trained models were then tested on the Combined Validation Dataset (*CVDS*), which consists of 150 images.
2) In the second experiment, 900 randomly selected images from the *FIVES* dataset were used to train both the *U-Net* and *CNN* models.
3) In the third experiment, we used the same *U-Net* model trained in the first experiment, but applied a naive Sliding Window Algorithm (*SWA*), as described in Section III-C.

The purpose of these experiments was to assess whether the full pipeline, outlined in Sections III-A through III-E, achieves superior segmentation performance compared to the traditional Sliding Window Algorithm (*SWA*). The central hypothesis is that identifying an optimal resolution, at which the *U-Net* model performs best for each individual image, leads to more accurate segmentation results then applying the model directly, at a fixed resolution in a brute-force manner. The *U-Net*

architecture used for segmentation is shown in Figure 6. As illustrated, the structure is simplified compared to the original version proposed in [2]. The accompanying *CNN* (Figure 7) was designed to be lightweight.

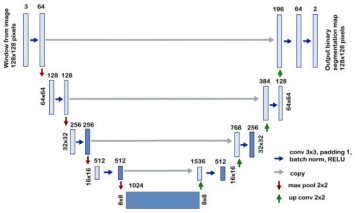


Fig. 6. *U-Net* used across all segmentation stages

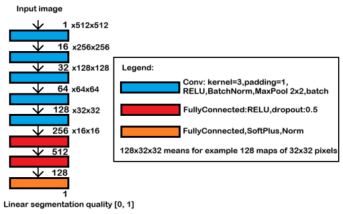


Fig. 7. *CNN* used for segmentation quality assessment

A. First experiment

In the first experiment we randomly selected 100 images from *FIVES* dataset. Results are shown in Table I.

TABLE I
EXPERIMENT 1 WITH JACCARD METRICS AND *TTE* FOR VARIOUS PARAMETER SETTINGS OF *WSS*.

RWSS	DWSS	OWSS	TTE	AJI	MNJI	MXJI
128	128	16	3316	0.6114	0.2722	0.7705
128	16	8	53107	0.6123	0.2823	0.7672
128	16	16	5641	0.6145	0.274	0.7650
128	32	8	10527	0.6148	0.2713	0.7672
128	64	8	3822	0.6147	0.2722	0.7705
128	**64**	**32**	**2704**	**0.6125**	**0.2714**	**0.7694**
128	**64**	**8**	**12877**	**0.6153**	**0.2713**	**0.7713**
16	16	8	35212	0.6078	0.2823	0.7570
16	32	8	18842	0.6089	0.2823	0.7570
16	64	8	7653	0.6097	0.2823	0.7516
32	32	8	14048	0.6082	0.2823	0.7436
32	64	8	11784	0.6099	0.2823	0.7516
64	16	8	16839	0.6137	0.2823	0.7649
64	16	16	5627	0.6147	0.2829	0.7634
64	32	8	9680	0.6151	0.2819	0.7665
64	32	16	4091	0.6153	0.2819	0.7668
64	64	8	7938	0.6143	0.2823	0.7601

The common parameters selected for the segmentation process are as follows:

- Minimum rough segmentation scale: 20%,
- Maximum rough segmentation scale: 100%,
- Rough segmentation scale step: 20%,
- Rough segmentation range for optimal resolution: $\pm$10%,
- Detailed segmentation range for optimal resolution: $\pm$4%.

The meanings of the column headings used in the evaluation tables are:

- *RWSS*: Rough Segmentation Window Step,
- *DWSS*: Detailed Segmentation Window Step,
- *FOWSS*: Final Optimal Window Step for Segmentation,
- *TTE*: Total Time Elapsed (in seconds),
- *AJI*: Average Jaccard Index,
- *MNJI*: Minimum Jaccard Index,
- *MXJI*: Maximum Jaccard Index.

The highest achieved *AJI* was 0.6153; however, this result came at the cost of a total time elapsed (*TTE*) of 12,877 seconds. In contrast, the fastest configuration, with a *TTE* of just 2,704 seconds, was obtained using the parameters *RWSS* = 128 pixels, *DWSS* = 64 pixels, and *FOWSS* = 32 pixels. This setup yielded an *AJI* of 0.6125, which is only marginally lower than the best result. As such, it represents a significantly more efficient trade-off between segmentation accuracy and computational performance.

B. Second experiment

In the second experiment we randomly selected 900 images from *FIVES* dataset. Exactly the same experiments in terms of values for **RWSS**, **DWSS** and **OWSS** were conducted and can be previewed in details in [3].

The best achieved *AJI* in this experiment was 0.5957, obtained with a total time elapsed (*TTE*) of 8,320 seconds. The fastest configuration, with a *TTE* of just 6,208 seconds, was achieved using the parameters *RWSS* = 128 pixels, *DWSS* = 64 pixels, and *FOWSS* = 32 pixels. However, this setup yielded an *AJI* of only 0.5813, which is noticeably lower than the optimal value. This indicates a trade-off between speed and accuracy, with the faster configuration sacrificing some segmentation performance.

C. Third experiment

In the third experiment we decided to try and apply only a part of the Section III-E3. We specified the *SS* to values of 8, 16, 32, 64 and 128. You can see the segmentation results in Table II. The results in this experiment are notably inferior to those presented in Section IV-A. The best outcome observed corresponds to *SS* of 8, resulting in *TTE* of 24,227 seconds and *AJI* of 0.5639. While this approach demonstrates high execution speed for *SS* values ranging between 32 and 128 pixels, the resulting *AJI* scores remain unsatisfactory. In contrast, the fastest configuration from Section IV-A achieved an *AJI* of 0.6125, clearly demonstrating the superiority of the hybrid approach described in this article over the traditional *SWA*-based method.

V. CONCULSIONS

The experiments conducted in this article reveal several notable observations regarding the proposed *HUS* algorithm:

- When using only 100 training images from the *FIVES* dataset [10], we achieved a higher accuracy of 0.6153 compared to 0.5957 obtained with 900 training images. This suggests that in Section IV-B, the *U-Net* and *CNN*

TABLE II
EXPERIMENT 3 WITH JACCARD METRICS AND *TTE* FOR VARIOUS PARAMETER SETTINGS OF *WSS*.

WSS	TTE	AJI	MNJI	MXJI
128	454	0.4878	0.1824	0.6490
64	591	0.5226	0.1920	0.6918
32	1354	0.5591	0.2582	0.7131
16	4504	0.5626	0.2626	0.7153
8	24727	0.5639	0.2642	0.71705

models may have become overly specialized to the training data, resulting in decreased generalization to the test set.

- In Experiments IV-A and IV-B, we demonstrated that the hybrid method outperforms the classic *SWA* approach by nearly 0.05 in *AJI*, while maintaining comparable execution speeds.
- Our experiments with *HUS* also indicate that high values of the sliding step (*SS*) are not necessarily critical for achieving optimal performance. The results suggest a possible correlation between hierarchical window sizes. Specifically, for a window size of 2^n, typically associated with the first convolutional layer of *U-Net*, the following configuration yields strong results: $RWSS = 2^n$, $DWSS = 2^{n-1}$, and $FOWSS = 2^{n-2}$.

The *HUS* algorithm [3] effectively addresses limitations associated with traditional *U-Net* segmentation:

- The requirement for training and test datasets to have similar resolutions.
- The substantial memory consumption of *U-Net* when processing high-resolution images.
- The sensitivity of classical *U-Net* to variations in the spatial location of the target object.

VI. DECLARATION OF NO COMPETING INTEREST

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E.,*ImageNet Classification with Deep Convolutional Neural Networks*,Advances in Neural Information Processing Systems (NIPS), volume 25, pages 1097–1105, 2012,doi:10.1145/3065386,https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[2] Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas,*U-Net: Convolutional Networks for Biomedical Image Segmentation*,Medical Image Computing and Computer-Assisted Intervention (MICCAI),pages 234–241, 2015, Springer,doi:10.1007/978-3-319-24574-4_28,https://arxiv.org/abs/1505.04597

[3] Lesław M. Pawlaczyk, *U-Net segmentation implementation*,2025, https://github.com/palles77/UnetSegmentation,Accessed: January 29, 2025

[4] Ren, Kan and Chang, Longdan and Wan, Minjie and Gu, Guohua and Chen, Qian,*An improved U-net based retinal vessel image segmentation method*,Heliyon, volume 8, number 10, pages e11187, 2022, Elsevier,doi:10.1016/j.heliyon.2022.e11187,https://www.sciencedirect.com/science/article/pii/S2405844022024756

[5] Liu, Rui and Pu, Wen and Nan, Hong and Zou, Yao,*Retina image segmentation using the three-path Unet model*,Scientific Reports, volume 13, number 1, pages 22579, 2023, Nature Publishing Group,doi:10.1038/s41598-023-50141-0,https://www.nature.com/articles/s41598-023-50141-0

[6] Das, Smita and Chakraborty, Suvadip and Mishra, Madhusudhan and Majumder, Swanirbhar,*Assessment of retinal blood vessel segmentation using U-Net model: A deep learning approach*,Franklin Open, volume 8, pages 100143, 2024, Elsevier,doi:10.1016/j.fraope.2024.100143,https://www.sciencedirect.com/science/article/pii/S2773186324000732

[7] Yun, Jiang and Wang, Falin and Gao, Jing and Cao, Simin,*Multi-Path Recurrent U-Net Segmentation of Retinal Fundus Image*,Applied Sciences, volume 10, number 11, pages 3777, 2020, MDPI,doi:10.3390/app10113777,https://www.researchgate.net/publication/341810610_Multi-Path_Recurrent_U-Net_Segmentation_of_Retinal_Fundus_Image

[8] Huang, Ko-Wei and Yang, Yao-Ren and Huang, Zih-Hao and Liu, Yi-Yang and Lee, Shih-Hsiung,*Retinal Vascular Image Segmentation Using Improved UNet Based on Residual Module*,Bioengineering, volume 10, number 6, pages 722, 2023, MDPI,doi:10.3390/bioengineering10060722,https://www.mdpi.com/2306-5354/10/6/722

[9] Westwańska, W. and Respondek, J., *Counting instances of objects in color images using U-Net network on example of honey bees*, in Proceedings of the 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), Annals of Computer Science and Information Systems, vol. 18, pp. 87–90, Leipzig, Germany, September 2019, doi:10.15439/2019F94, https://ieeexplore.ieee.org/document/8859742

[10] Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, Juan Ye, *FIVES: A Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation*, 2022 Data Descriptor in Scientific Data, volume 9, number 475, pages 475, 2022, Nature Portfolio,doi:10.1038/s41597-022-01590-2,https://www.nature.com/articles/s41597-022-01564-3,https://figshare.com/ndownloader/files/34969398

[11] Holm, Sven, Russell, Greg, Nourrit, Vincent, McLoughlin, Niall, *DR HAGIS - a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients*, 2017 *Journal of Medical Imaging (Bellingham)*, volume 4, number 1, pages 014503, 2017, SPIE,doi:10.1117/1.JMI.4.1.014503,https://pmc.ncbi.nlm.nih.gov/articles/PMC5299858/pdf/JMI-004-014503.pdf,http://personalpages.manchester.ac.uk/staff/niall.p.mcloughlin/DRHAGIS.zip

[12] Budai, Attila, Bock, Rüdiger, Maier, Andreas, Hornegger, Joachim, Michelson, Georg,*Robust Vessel Segmentation in Fundus Images*,2013 *International Journal of Biomedical Imaging*,volume 2013, article 154860, 2013, Hindawi,doi:10.1155/2013/154860,http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2013/Budai13-RVS.pdf,https://www5.cs.fau.de/fileadmin/research/datasets/fundus-images/all.zip

[13] Hoover, A. D., Kouznetsova, V., Goldbaum, M.,*Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response*,2000 *IEEE Transactions on Medical Imaging*,volume 19, number 3, pages 203–210, 2000, IEEE,doi:10.1109/42.845178,https://cecas.clemson.edu/~ahoover/stare/probing/index.html

[14] Jaccard, P., *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*, 1901, *Bulletin de la Société Vaudoise des Sciences Naturelles*, volume 37, pages 547–579, Société Vaudoise des Sciences Naturelles.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 369–374
DOI: 10.15439/2025F3919
ISSN 2300-5963 ACSIS, Vol. 43

Enhancing Low-Cost Air Quality Sensors with AI for Smart Green Routing

Laura Po, Federica Rollo, Martina Casari
0000-0002-3345-176X, 0000-0002-3834-3629, 0000-0003-0406-3036
"Enzo Ferrari" Engineering Department
University of Modena and Reggio Emilia
Modena, Italy

Matteo Angelinelli, Giorgio Pedrazzi, Roberta Turra
0000-0002-3382-1900, 0000-0001-8896-3044, 0000-0001-5973-3805
HPC Department, Cineca
Casalecchio di Reno, Bologna, Italy

Chiara De Pascali, Luca Nunzio Francioso
0000-0002-2673-4593, 0000-0002-4500-0421
National Research Council, CNR
Institute for Microelectronics and Microsystems
IMM Lecce, Italy

***Abstract*—Urban air pollution poses severe health risks, demanding accurate monitoring and exposure-reducing solutions. Low-cost air quality sensors (LCS) provide high spatial resolution but suffer from accuracy limitations that hinder their reliability. This paper presents the AIQS project (AI-enhanced air quality sensor for optimizing green routes), an ongoing initiative that combines artificial intelligence, sensor hardware optimization, and pedestrian routing innovation to address these challenges. AIQS applies machine learning techniques, including Multilayer Perceptrons and fuzzy logic, to correct sensor readings. In parallel, hardware-level optimizations, such as fluid dynamics simulations and pre-treatment modules, are explored to enhance sensor performance. The corrected AQ data is then incorporated into a configurable routing tool capable of estimating pollutant exposure and computing low-exposure pedestrian paths in urban environments. First evaluations, shows that our correction models achieve up to 0.92 R^2 against reference data across diverse urban environments. The corrected data drives a configurable routing tool that computes paths minimizing cumulative pollution exposure while balancing user preferences (e.g., proximity to green spaces). Preliminary validation in Modena, Italy demonstrates viable "green routes".**

***Index Terms*—Air quality monitoring, Low-cost sensors, Artificial Intelligence, Machine Learning, Data correction, Fuzzy logic, Green routing, Urban mobility, Pollution exposure.**

I. Introduction

URBAN air pollution remains one of the most pressing environmental health risks worldwide, contributing to respiratory illnesses, cardiovascular diseases, and premature deaths, particularly in densely populated areas [1]. Monitoring air quality (AQ) is therefore crucial for supporting mitigation strategies and informing public health interventions. In recent years, low-cost air quality sensors (LCS) have emerged as a promising tool for expanding AQ monitoring coverage, especially in urban environments where spatial resolution is key [2]. However, despite their affordability and ease of deployment, LCS suffer from significant limitations, including low measurement accuracy, susceptibility to environmental conditions such as humidity, and a lack of standardized calibration procedures [3].

To address these challenges, data correction algorithms have been introduced to improve the accuracy of sensor measurements, with growing interest in applying artificial intelligence (AI) and machine learning (ML) techniques to this task [4], [5]. Concurrently, urban navigation systems are increasingly being explored as tools for reducing human exposure to pollution by suggesting "green routes"—paths that prioritize air quality rather than shortest time or distance [6]. Yet, these systems are only as reliable as the data they consume; inaccurate AQ measurements from low-cost sensors risk undermining the effectiveness of health-oriented routing recommendations.

In this paper we present the AIQS project (AI-enhanced air quality sensor for optimizing green routes) that directly addresses this gap by combining hardware improvements, AI-driven data correction methods, and routing algorithm integration to enhance the trustworthiness and usability of AQ data in real-world applications. Building upon the open-source MitH framework for humidity correction [5], the project aims to raise the technology readiness level (TRL) of the solution to TRL 7, demonstrating its applicability across various urban environments through partnerships with environmental agencies and industrial stakeholders. In doing so, AIQS represents a comprehensive, interdisciplinary effort to bring together environmental sensing, machine learning, and human-centered navigation in support of healthier, more sustainable cities.

The paper is structured as follows. Section II reviews existing approaches to improve the accuracy of LCS through AI-driven data correction techniques and discusses urban navigation systems designed to reduce exposure to air pollution. Section III presents the integrated strategy of the AIQS project, that combines AI-based data correction, hardware optimization, and pollution-aware routing. Section IV reports preliminary results from data correction experiments highlighting improvements in sensor accuracy achieved through transformer-based models and validating the approach across multiple datasets from both Italian and international sources,

moreover the pollution-aware routing system and an evaluation in a real urban context is shown. These findings support the development of health-conscious routing systems within the broader vision of sustainable smart cities. Finally, Section V offers concluding remarks and suggests directions for future work.

II. Related Work

A. AI for Air Quality

Recent advances in ML for air quality monitoring have demonstrated promising approaches to enhance low-cost sensor accuracy. Hybrid models combining multiple algorithms consistently outperform single-technique approaches for AQI prediction [7], while neural networks have shown significant improvements in PM10 calibration [8]. Practical implementations of end-to-end IoT pipelines for sensor data acquisition and cleaning have also been demonstrated in real-world deployments [9].

For short-term forecasting, deep learning models like LSTM have proven to be effective when integrated with meteorological data. Arsov et al. [10] showed that LSTM networks outperform traditional ARIMA models for PM_{10} prediction (+3 hours) with MSE as low as 0.01. These approaches collectively highlight the importance of data preprocessing and multi-source integration to address LCS limitations.

B. Green Navigation Systems

The integration of air quality data into urban navigation systems is emerging as a key strategy for reducing pollution exposure. IoT technologies enable smart city solutions that combine heterogeneous data sources to support sustainable mobility [11]. Recent frameworks leverage traffic-aware predictions and routing algorithms like Dijkstra's to reduce pedestrian PM2.5 exposure by 11-15% [12], while prototypes in cities like Barcelona demonstrate the feasibility of real-time green routing [13]. Despite these advances, most navigation systems still lack robust environmental health considerations.

III. Integrated Approach

The AIQS project adopts an interdisciplinary approach that combines AI-based data correction, sensor hardware optimization, and the development of a pollution-aware routing tool. Its goal is to deliver a validated, open-source system that enhances LCS reliability and supports healthier, user-configurable pedestrian navigation in urban environments.

The main goals are:

- Improving sensor data accuracy: Develop and apply AI and ML techniques, to dynamically correct sensor readings and improve reliability in various environmental contexts.
- Optimize sensor hardware: Enhance the physical design and operation of low-cost AQ sensors through fluid dynamics simulations and the development of pre-treatment modules, such as air dehydration systems and anti-coalescence components.
- Deliver a validated demonstrator: Advance the MitH framework to TRL 7 by integrating AI algorithms and hardware improvements into a real-world demonstrator, tested in an urban pilot and released as an open-source tool for broad adoption.
- Develop a green routing tool: Design and implement a flexible routing tool that integrates air quality data to compute pollution-aware pedestrian routes. It will estimate individual exposure to pollutants based on real-time AQ levels along each road segment, and compute routes that minimize cumulative exposure. The tool will be adaptable to different urban contexts and configurable by users, who will be able to tune preferences such as avoiding polluted areas, favoring proximity to green spaces, or prioritizing walking routes over other options.

A. Data Collection & AI Correction

To enhance the accuracy of low-cost air quality sensors, we apply AI-based correction techniques using data collected from co-located low-cost sensors and regulatory reference stations (e.g., ARPA) across multiple Italian and international cities. These co-located datasets allow us to calibrate sensor readings by learning correction functions that map raw low-cost measurements to validated reference values under shared environmental conditions. Meteorological variables are also included to improve model robustness. Our approach combines dynamic correction functions to mitigate humidity effects, anomaly detection to filter spurious data, and machine learning algorithms, such as Multilayer Perceptrons and fuzzy logic systems, to refine sensor outputs. These components are being integrated into the open-source MitH framework [5], which we aim to advance to Technology Readiness Level (TRL) 7. Model training and evaluation are supported by a high-performance computing infrastructure with GPU capabilities.

B. Pollution-Aware Routing

LCS measurements provide local PM concentrations. Through spatial interpolation strategies [14], we generate city-wide air quality maps that enable health-conscious navigation by estimating pollution levels across urban areas.

These AQ datasets drive our configurable routing system, which computes pedestrian paths minimizing cumulative pollutant exposure using OpenStreetMap networks. The tool supports user-customizable preferences:

- Minimizing pollution exposure
- Prioritizing proximity to green spaces
- Selecting pedestrian-only segments

Currently being validated in Modena through stakeholder partnerships, the open-source implementation[1] represents an adaptable framework for human-centered, pollution-aware navigation in diverse urban contexts.

[1]https://github.com/federicarollo/GRAFMOVE

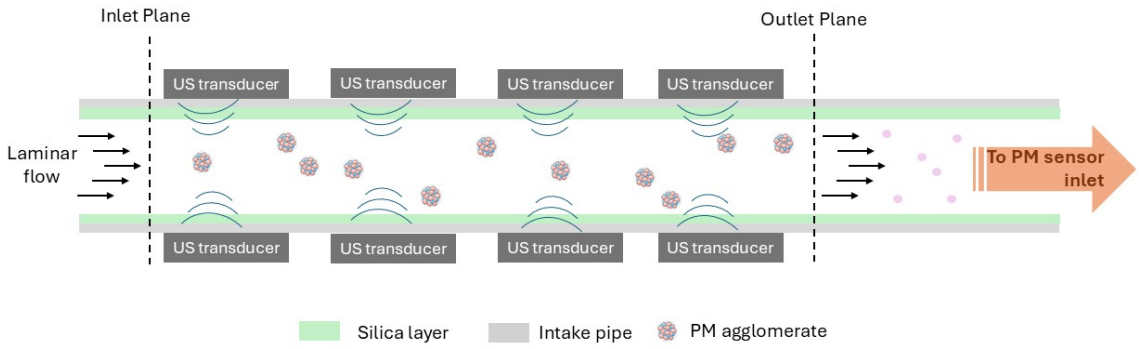


Fig. 1: Low-power air preconditioning module for particulate matter.

C. Hardware Optimization for Humidity Mitigation

A compact pretreatment module is being developed to minimize humidity-induced artifacts in LCS sensors. As shown in Figure1, ambient air enters through an intake pipe whose inner surface is coated with a thin silica aerogel layer acting as a passive desiccant. Along the pipe, ultrasonic (US) transducers are mounted transversely and driven in phase to generate localized oscillations and mild heating. This action promotes de-aggregation of PM agglomerates and accelerates moisture desorption from particles, while simultaneously regenerating the silica layer by enhancing vapor transport. The laminar flow then reaches the PM sensing chamber with reduced aggregation and humidity effects.

Finite-Element simulations guide the optimization of acoustic and thermal fields to achieve these effects with low power consumption, and a PWM-controlled heater can be added if ultrasonic energy alone is insufficient. Ongoing evaluation quantifies the reduction of humidity-related artifacts by comparing sensor responses with and without the module under controlled conditions.

IV. Results

A. Data Collection

Several key activities were carried out to enhance the accuracy and reliability of low-cost air quality sensors using artificial intelligence and high-performance computing resources.

We completed a comprehensive dataset collection [15] that includes:

- 4 Italian air quality datasets (e.g., ARPA regional agencies),
- 10 international datasets across various urban environments,
- Meteorological variables, including temperature, relative humidity, atmospheric pressure, and wind speed/direction, collected specifically for each city to support model robustness and correction accuracy.

All datasets include measurements from co-located low-cost sensors and regulatory reference stations, which is a critical requirement for developing accurate correction models. Co-location ensures that both LCS and RS measure the same environmental conditions at the same time and place, enabling a direct comparison between the two. This alignment is essential for calibrating LCS readings, as RS data provide a validated high-quality ground truth against which LCS measurements can be adjusted. The availability of synchronized data supports the development of reliable correction algorithms that significantly enhance the accuracy and trustworthiness of LCS output. Furthermore, all datasets feature at least an hourly measurement frequency, ensuring adequate temporal resolution for both real-time applications and time series analysis.

B. AI-Based Sensor Data Correction

In the preliminary phase of our study, we conducted evaluations of various anomaly detection and correction algorithms using some of the air quality datasets collected. The goal was to improve the quality of the sensor data and improve the performance of the model in subsequent prediction and correction tasks. These initial experiments provided insight into the effectiveness of both statistical and AI-based methods in identifying and correcting anomalies, as well as in accurately modeling time series data.

a) Anomaly Detection: To identify anomalous patterns in the AQ time series data, we applied a range of traditional outlier detection techniques. These included density-based spatial clustering of noise applications (DBSCAN), the inter-quartile range (IQR) method, and the three-sigma rule. Each method contributed to detecting noise and irregularities across different temporal contexts. For time series repair, we employed K-Nearest Neighbors (KNN) interpolation, which proved effective in reconstructing missing or corrupted values. Our preliminary results indicate that the inclusion of these anomaly detection and repair steps in the pre-processing pipeline leads to measurable improvements in model performance. In particular, we observed increases of up to 0.1 in the coefficient of determination (R^2) in some cases, highlighting the value of robust data preparation.

b) AI Algorithms for Correction: To correct and calibrate sensor readings, we tested a wide spectrum of traditional and advanced machine learning models. Among traditional

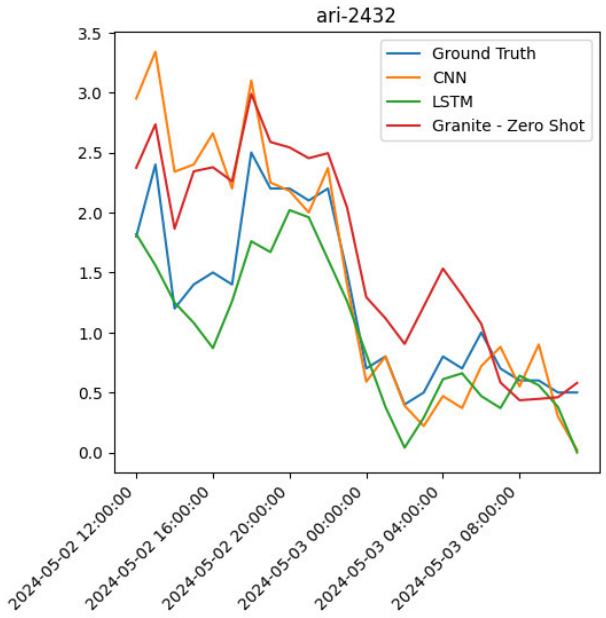


Fig. 2: A 24-hour forecast produced using different deep learning approaches on a randomly chosen AQ sensor.

regression techniques, we employed Linear Regression, Lasso, Ridge, and Elastic Net, all of which served as baseline models for evaluating correction accuracy.

We then explored supervised machine learning models, including Support Vector Regression (SVR), Decision Trees (DT), Random Forest (RF), XGBoost, and LightGBM. These models demonstrated varying degrees of effectiveness in correcting sensor data, with ensemble methods generally outperforming simpler approaches.

In addition to these methods, we investigated a fuzzy logic-based correction strategy using the Adaptive Neuro-Fuzzy Inference System (ANFIS) [16]. ANFIS achieved performance levels comparable to those of neural network models, obtaining a value R^2 of approximately 0.68 on the primary test dataset. Further experimentation with ANFIS on five additional datasets, focusing on model interpretability, yielded Pearson correlation coefficients exceeding 0.75 in all cases and reaching up to 0.97 in the best scenario.

Neural network models, particularly Multilayer Perceptrons (MLPs) [17], were also tested. These models achieved strong performance, with an R^2 value of 0.92 in one of the urban scenarios. Moreover, we implemented dynamic correction functions to account for the hygroscopic effect of humidity using both empirical and data-driven approaches. These corrections led to improvements of up to 0.3 in R^2 values during the autumn period, demonstrating the importance of incorporating environmental factors in correction models.

c) Deep Learning Architectures:: We also performed a series of preliminary tests using deep learning architectures, evaluating their potential in modeling AQ time series. These included Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), Temporal Convolutional Networks (TCNs), and Transformer-based models. These architectures were trained and tested on domain-specific AQ datasets, with particular attention to their ability to reproduce both general trends and specific point values.

Among these, the CNN and LSTM models showed the most promise in accurately modeling the observed behaviors. Our early results suggest that, while these models are capable of capturing overall trends, fine-grained prediction of individual values remains challenging in cases of high temporal variability. Further optimization, particularly in terms of network architecture and hyperparameter tuning, is necessary to enhance performance and potentially surpass the results obtained from simpler models.

In addition, we evaluated the performance of foundation models designed for time-series analysis. Two such models, TimesFM [18] and Granite [19], were tested. TimesFM was evaluated in zero-shot mode, while Granite was tested both in zero-shot and fine-tuned settings. Initial findings suggest that these foundation models can achieve performance comparable to traditional deep learning models without requiring extensive training. Notably, fine-tuning Granite on specific AQ datasets further improved its accuracy, indicating the potential for transfer learning in environmental sensing applications. However, further investigation is needed to draw definitive conclusions, as specific metrics are still under analysis.

Figure 2 illustrates an example of 24-hour time series from one AQ sensor. In these examples, both the trained architectures and the foundation model successfully capture the ground truth trend. While sharp variations (left panel) present challenges in reproducing point values, smoother patterns (right panel) are more accurately modeled. Figure 3 provides a comparative overview of the R^2 values obtained from all tested deep learning models. It is worth emphasizing that preliminary tests highlight the critical importance of defining appropriate training, validation, and test sets, as their configuration significantly impacts model performance. Ongoing experiments aim to systematically address this issue and validate the robustness of the observed results.

Among the traditional architectures, *LSTM*, *GRU*, and *CNN* demonstrate exceptionally high performance with R^2 values approaching 1.0, suggesting that these models are highly capable of capturing temporal dependencies and nonlinear patterns in the data. The *Transformer*, despite its theoretical strengths, shows comparatively lower performance ($R^2 \sim 0.87$), possibly due to the need for more extensive hyperparameter tuning or larger datasets to fully leverage its self-attention mechanisms.

The *TCN* model performs well with an R^2 of 0.95, slightly below the recurrent models but still in a strong range. This suggests TCNs can be a competitive alternative.

Regarding foundation models, *TimesFM - Zero Shot* achieves an R^2 of 0.93, which is notable given that it was used without any fine-tuning. Similarly, *Granite - Zero Shot* performs slightly better ($R^2 \sim 0.96$), indicating strong out-of-the-box generalization capabilities. However, the *Granite* -

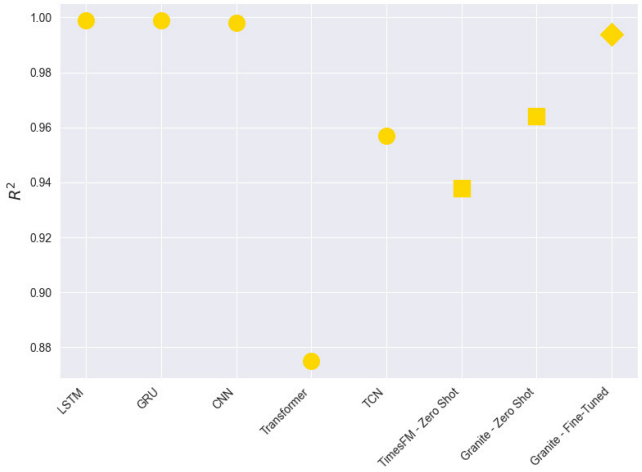


Fig. 3: Comparison of R^2 metrics computed over several deep learning architectures on a test set includes all the available AQ sensors.

Fine-Tuned model outperforms all other approaches with an R^2 of 0.99, highlighting the significant boost in performance that can be gained from domain-specific fine-tuning.

C. Pollution-Aware Routing System

We extracted and analyzed cycleway and footpath data from OpenStreetMap (OSM) and Open Data portals, focusing on the city of Modena. The data was modeled as a directed graph, where $Nodes$ represent OSM-defined junctions or road shape points, with properties such as the OSM identifier and the GPS coordinates, $Edges$ correspond to roads, with properties including: *geometry* representing the shape of the road, *distance*, i.e., the geometric length of the road, *green_area* indicating whether the segment traverses a green space (e.g., parks).

Only pedestrian-accessible roads (e.g., sidewalks, footpaths) and cycleways were included. The graph was implemented in Neo4j, using the spatial plugin.

To enable pollution-aware routing, we sourced open air quality data for Modena. Due to the absence of hyperlocal measurements (e.g., particulate matter sensors), we generated synthetic PM_{10} data by analyzing historical measurements from ARPAE's two regulatory stations in Modena, applying Inverse Distance Weighting (IDW) interpolation using the GDAL library[2] to estimate PM_{10} levels across the road network, assigning each edge an air_quality property, calculated as the average PM_{10} concentration within a configurable buffer around the road.

Using the enriched graph, we developed a routing algorithm to prioritize low-pollution pedestrian paths leveraging on the routing algorithms (Dijkstra and A*) of the APOC library[3].

Preliminary tests were performed using randomly selected locations within the urban area of Modena, computing optimal routes based on edge properties including *distance*, *green_area*, and *air_quality*. Figure 4 presents a comparison between two routes connecting the San Paolo kindergarten and the "Fontana dei due fiumi". The left route represents the distance-optimized solution (shortest path), while the right route maximizes green space exposure along the path.

[2]https://gdal.org/en/stable/programs/gdal_grid.html
[3]https://github.com/neo4j-contrib/neo4j-apoc-procedures

V. Conclusion

This paper presented the AIQS project, an integrated approach to enhance the reliability of low-cost air quality sensors (LCS) using AI-driven correction and to enable pollution-minimizing "green routing" for pedestrians.

The AI-based sensor correction ca be performed using Multilayer Perceptrons (MLPs), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), or foundation models (e.g., TimesFM, Granite), and achieves up to 0.92 R^2 accuracy against reference data across diverse urban environments. The configurable routing tool leveraging OpenStreetMap and Dijkstra/A* algorithms to compute paths minimizing cumulative pollution exposure while balancing user preferences has been validated in Modena and confirmed the opportunity to generate actionable green routes.

Upcoming efforts will focus on finalizing the integration of the AI models into the open source MitH framework. The framework will be validated on at least 6 real AQ datasets to ensure robustness across diverse conditions. On the hardware side, planned field trials will rigorously assess the effectiveness and durability of the proposed pre-treatment module in real environments. In parallel, the routing system will evolve into a deployable web application with multi-objective optimization, adaptable to different urban contexts. By addressing both data accuracy and usability, AIQS contributes to the development of healthier, cleaner, and more sustainable cities.

By addressing both data accuracy and application usability, AIQS aims to contribute to the development of healthier, cleaner, and more sustainable cities through AI-enhanced environmental monitoring and decision-making tools.

Acknowledgment

This work was carried out as part of the AIQS project (AI-enhanced Air Quality Sensor for Optimizing Green Routes), under the project code DIP_AIQS_PO_2025_PNRR_ECOS_SK4AF_E93C22001100001. AIQS was funded through a closed call within the initiative "Ecosystem for Sustainable Transition in Emilia-Romagna" (ECOSISTER), financed under the National Recovery and Resilience Plan (PNRR) – Mission 4 "Education and Research", Component 2 "From Research to Business", Investment 1.5 "Creation and strengthening of innovation ecosystems, building territorial R&D leaders" – funded by the European Union – *NextGenerationEU* (Grant Agreement No. 0001052, dated 23/06/2022 – Project ECS_00000033 – CUP E93C22001100001).

We gratefully acknowledge the Regional Agency for Environmental Protection of the Aosta Valley, Arpae Emilia-Romagna, and the Provincial Agency for Environmental Protection of Trento for providing data from their reference

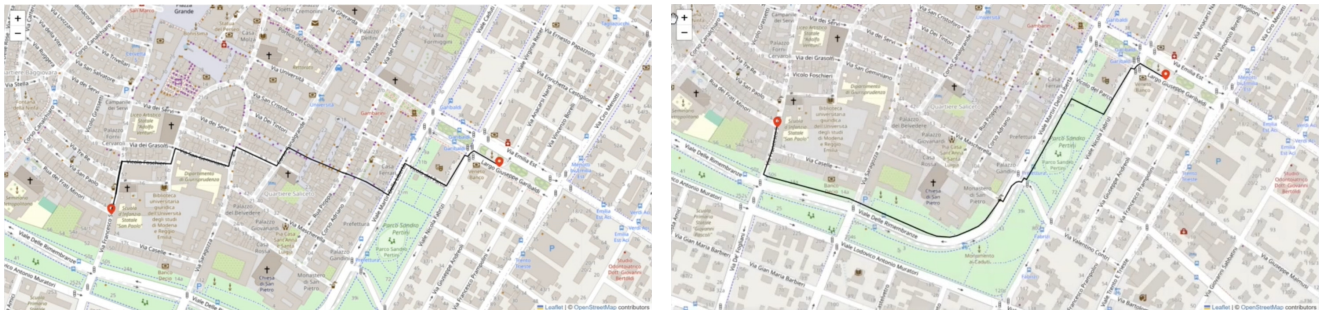


Fig. 4: Comparison of optimized routes based on shortest path (left) and greenest path (right) sharing the same origin and destination points.

monitoring stations. These stations were co-located with low-cost sensors donated by Wiseair Srl, whose contribution of sensors is also sincerely appreciated.

REFERENCES

[1] W. H. Organization, "Ambient air pollution: A global assessment of exposure and burden of disease," https://www.who.int/publications/i/item/9789241511353, 2018.

[2] B. Alfano *et al.*, "A review of low-cost particulate matter sensors from the developers' perspectives," *Sensors*, vol. 20, no. 23, p. 6819, 2020.

[3] A. Molnár *et al.*, "Aerosol hygroscopicity: Hygroscopic growth proxy based on visibility for low-cost pm monitoring," *Atmospheric Research*, vol. 236, 2020.

[4] M. Casari, L. Po, and L. Zini, "Airmlp: A multilayer perceptron neural network for temporal correction of pm2.5 values in turin," *Sensors*, vol. 23, no. 23, p. 9446, 2023.

[5] M. Casari and L. Po, "Mith: A framework for mitigating hygroscopicity in low-cost pm sensors," *Environmental Modelling & Software*, vol. 173, p. 105955, 2024.

[6] S. Mahajan *et al.*, "Car: The clean air routing algorithm for path navigation with minimal pm2.5 exposure on the move," *IEEE Access*, vol. 7, pp. 147 373–147 382, 2019.

[7] T. Desai *et al.*, "Comparative analysis of machine learning algorithms for air quality index prediction," in *Machine Learning for Computational Science and Engineering*, 2025.

[8] P. Zhivkov, "Optimization and evaluation of calibration for low-cost air quality sensors: Supervised and unsupervised machine learning models," *Annals of Computer Science and Information Systems*, vol. 25, pp. 255–258, 2021. [Online]. Available: https://doi.org/10.15439/2021F95

[9] F. Rollo, B. Sudharsan, L. Po, and J. G. Breslin, "Air quality sensor network data acquisition, cleaning, visualization, and analytics: A real-world iot use case," in *UbiComp/ISWC '21: 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and 2021 ACM International Symposium on Wearable Computers, Virtual Event, September 21-25, 2021*, A. Doryab, Q. Lv, and M. Beigl, Eds. ACM, 2021, pp. 67–68. [Online]. Available: https://doi.org/10.1145/3460418.3479277

[10] M. Arsov, E. Zdravevski, P. Lameski, R. Corizzo, N. Koteli, K. Mitreski, and V. Trajkovik, "Short-term air pollution forecasting based on environmental factors and deep learning models," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 21, 2020, pp. 15–22. [Online]. Available: https://doi.org/10.15439/2020F211

[11] B. Cengiz *et al.*, "A survey on data fusion approaches in IoT-based smart cities: Smart applications, taxonomies, challenges, and future research directions," *Information Fusion*, vol. 121, 2025.

[12] R. Sinnott and S. Zhong, "Real-time route planning to reduce pedestrian pollution exposure in urban settings," in *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing*, 2023, pp. 1–10. [Online]. Available: https://doi.org/10.1145/3632366.3632381

[13] F. Bistaffa and P. C. Oliveira, "Green routes in barcelona: A pedestrian routing prototype to reduce air pollution exposure," https://filippobistaffa.github.io/papers/2025greenroutes.pdf, 2025, accessed on May 27, 2025.

[14] C. Bachechi, F. Desimoni, L. Po, and D. M. Casas, "Visual analytics for spatio-temporal air quality data," in *24th International Conference on Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*, E. B. et. al., Ed. IEEE, 2020, pp. 460–466. [Online]. Available: https://doi.org/10.1109/IV51561.2020.00080

[15] M. Casari and E. Montorsi, "Martinacasari/airqualitydatasetsrepository: Aqdr - v1.0.0 (v1.0.0)," *Zenodo* https://doi.org/10.5281/zenodo.13982208, 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13982208

[16] M. Casari, P. A. Kowalski, and L. Po, "Optimisation of the adaptive neuro-fuzzy inference system for adjusting low-cost sensors pm concentrations," *Ecological Informatics*, vol. 83, p. 102781, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954124003236

[17] M. Casari, L. Po, and L. Zini, "Airmlp: A multilayer perceptron neural network for temporal correction of pm2.5 values in turin," *Sensors*, vol. 23, no. 23, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/23/9446

[18] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," 2024. [Online]. Available: https://arxiv.org/abs/2310.10688

[19] V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. H. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam, "Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 74 147–74 181. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/874a4d89f2d04b4bcf9a2c19545cf040-Paper-Conference.pdf

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 375–380
DOI: 10.15439/2025F9016
ISSN 2300-5963 ACSIS, Vol. 43

Metaheuristics for Rolling Stock Cyclic Job Scheduling Problems with Maintenance

Radosław Rudek
General Tadeusz Kościuszko Military University of Land Forces
Czajkowskiego 109, 51-147 Wrocław, Poland
Email:rudek.radoslaw@gmail.com

***Abstract*—In this paper, we analyse rolling stock cyclic job scheduling problems with maintenance to maximize the total availability of vehicles or to minimize the total maintenance cost. The considered issues can be formulated as cyclic job scheduling problems on parallel machines with time and usage based deteriorating effects and maintenance. Since these problems are strongly NP-hard, we propose a dedicated double crossover genetic algorithm and compare its efficiency with other metaheuristics: particle swarm optimization, simulated annealing, and genetic algorithm (using a single crossover). The computational experiments reveal that our approach is robust, optimizes various instances, and overwhelms the other evaluated algorithms for the analysed scenarios correlated with real-life cases. Thus, our new method is applicable for the industrial practice to maximized availability of vehicles as well as to minimize maintenance costs.**

***Index Terms*—Scheduling, Rolling stock, Deteriorating, Maintenance, Metaheuristic**

I. Introduction

SCHEDULING problems with deteriorating effects and/or maintenance activities have attracted particular attention in the research society (e.g., [1], [2], [3]). It follows from the practical meaning of the related research, since most of the industrial systems undergo degradation, and to avoid unexpected breakdowns preventive maintenance are performed. It is especially important in a railway domain, where the related rules concerning maintenance of vehicles are strictly defined. Usually, for locomotives five levels of activities are distinguished: inspections (Level 1), periodic inspections (Level 2), extended periodic inspections (Level 3), overhauls (Level 4), overhauls and upgrades (Level 5) [4]. However, they are not only time consuming and making vehicles unavailable [5], but they are also expensive (see [6]) especially for higher levels. On the other hand, the maintenance dates can be controlled within a certain range by scheduling taking into account required mileage of transport tasks and additional constraints following from maintenance policies defining time-based maintenance (TBM) and mileage based-maintenance (MBM) cycles. Such approach was for the first time analysed, developed and implemented by the author for a locomotive rental company, and his results including the dedicated mathematical (simulation) model and the optimization algorithms were described in details in the industrial report [7].

In this paper, we will follow that approach on rolling stock cyclic job scheduling problems with maintenance to maximize the total availability of vehicles or to minimize of the total maintenance cost. To solve them, we will propose a dedicated double crossover genetic algorithm and compare its efficiency with other metaheuristics known from the literature such as our previous implementations of particle swarm optimization [8], simulated annealing ([7], also presented in [5]) and a genetic algorithm with a single partially mapped crossover PMX [5] adjusted to solve the required criteria.

II. Problem formulation

In this section, we will describe the analysed rolling stock cyclic job scheduling problems to maximize the total availability of vehicles (TA) and to minimize the total maintenance cost (TC) as a cyclic job scheduling problems on identical parallel machines with deteriorating effects and maintenance. It is worth highlighting that these problems are covered by the mathematical (simulation) model, which was formulated solely by the author (R. Rudek) and precisely described in his industrial report [7]. Its fundamental part is the discrete event simulator (DES) including the model parameters and their relations describing vehicles, maintenance policies, jobs, schedules and rules allowing for determining dates and durations of maintenance activities and calculating various criteria. These main assumptions in a simplified form were presented to a broader audience *inter alia* in [9] and [5] to optimize the total availability criteria, the total maintenance cost [8] or the total income [10].

Let us recall that the discussed problems can be formulated as cyclic job scheduling problems on parallel machines with time and usage based deteriorating effects and maintenance under the these criteria. The scheduling problem (model) parameters and constraints are given as follows. There is a time horizon $\mathcal{T} = \{1, \ldots, t, \ldots, T\}$ is divided into T periods (in practice referring to succeeding months). A set $\mathcal{V} = \{1, \ldots, v, \ldots, n\}$ of n parallel machines (representing locomotives) that have to process a set $\mathcal{J} = \{1, \ldots, j, \ldots, n\}$ of n cyclic jobs (identified with transport tasks) during all periods T. A job can be assigned exactly to one and the same machine for a given period t, and a machine can process only one job at a time, but the reassignment in the next periods is allowed. Moreover each job $j \in \mathcal{J}$ is characterized by the average milage per day (workload) w_j that defines daily deteriorating of a machine (increasing its accumulative mileage) during all days in period t, when that machine is

operational (available) and processes job j. On the other hand machines, due to their deteriorating, must to undergo periodic preventive maintenance activities MAs, and each of them belongs to exactly one family (type) $f \in \mathcal{F} = \{1, \ldots, F\}$. Machines of a given type f share the same maintenance policy (MP), which defines time-based maintenance (TBM) and mileage (distance) based-maintenance (MBM) cycles divided into K^f levels of complexity. Each level $k = 1, \ldots, K^f$ of MP is described by the related parameters TBM_k^f and MBM_k^f, which are the predetermined threshold values triggering MA according to TBM or MBM, respectively. A maintenance activity of level k for a machine of type f is described by the expected duration MD_k^f and the cost MC_k^f. The maintenance activities of lower levels are covered by the higher levels and the following relations hold: $TBM_k^f < TBM_{k+1}^f$, $MBM_k^f < MBM_{k+1}^f$, $MD_k^f < MD_{k+1}^f$ and $MC_k^f < MC_{k+1}^f$ for all $k = 1, \ldots, K^f - 1$. Each machine $v \in \mathcal{V}$ is characterized by its deteriorating (condition) parameters $VT_{v,k}$ and $VM_{v,k}$ denoting the period and the accumulated mileage since the last maintenance of level k, respectively. The maintenance activity MA of level k for machine v of type f is started, when the first of the following deteriorating rules does not hold: $VT_{v,k} \leq TBM_k^f$ or $VM_{v,k} \leq MBM_k$ for all $k = 1, \ldots, K^f$ (referring to time and usage deteriorating). If a machine is out of order due to MA, then each job assigned to that machine during related unavailability periods is serviced by the third party. After MA of level k is completed (meaning also the completion of lower levers) the related deteriorating values (counters/conditions) are set to zero $VT_{v,i} = 0$ and $VM_{v,i} = 0$ for $i = 1, \ldots, k$ and a machine can process an assigned job (returned from the third party).

The considered problem is to find a schedule S that is an assignment of all jobs $j \in \mathcal{J}$ to all machines $v \in \mathcal{V}$ during all periods $t \in \mathcal{T}$, which optimizes the required criterion value Q, i.e., the total availability of machines (vehicles) TA or the total maintenance cost TC. In other words, the schedule $S = \langle S_1, \ldots, S_t, \ldots, S_T \rangle$ is the sequence of the assignments of jobs to machines (vehicles) $S_t = (S_t(1), \ldots, S_t(v), \ldots, S_t(n))$ in succeeding periods $t = 1, \ldots, T$, where $S_t(v) \in \mathcal{J}$ is the index of a job assigned to machine (vehicle) v at period t. Note that S can be represented unambiguously by the sequence of permutations S_t. Furthermore, a schedule determines the deteriorating of machines in particular periods (usually months) caused by processing of jobs, which together with model parameters and constraints (such as the maintenance policies) imply the maintenance plan (dates of maintenance activities) that is precisely described by our mathematical model RRR including the discrete event simulator [7]. Thus, the calculation of the criterion value $Q(S)$ for a given schedule S is based on the maintenance plan provided by our model, and it requires $O(MKn)$ steps, where M is the number of performed maintenance activities and $K = \max_{f \in \mathcal{F}}\{K^f\}$ is their maximum level for all machines. Therefore, the objective is to find a schedule S^* that optimizes the criterion value $Q(S)$, i.e., the maximization of the total availability of machines (vehicles) TA: $S^* = \arg\max_{S \in \mathbb{S}}\{Q(S)\} = \arg\max_{S \in \mathbb{S}}\{TA(S)\}$, the minimization of the total maintenance cost TC: $S^* = \arg\min_{S \in \mathbb{S}}\{Q(S)\} = \operatorname{argmin}_{S \in \mathbb{S}}\{TC(S)\}$, where $\mathbb{S}$ is the set of all feasible schedules.

III. Computational complexity

The rolling stock cyclic job scheduling problem to maximize the total availability of vehicles is strongly NP-hard [5]. Here, we will present a sketch of the strong NP-hardness proof of the analysed problem to minimize the total maintenance cost, since its full version is beyond the scope of this paper.

Theorem 1: The rolling stock cyclic job scheduling problem to minimize the total maintenance cost is strongly NP-hard.

Proof. The strong NP-hardness follows from the proof of the strong NP-hardness of the problem with the minimization of the total time on maintenance activities (see [5]). It can be done by extending the pseudopolynomial time transformation such that the parameter *maintenance cost* $MC = 1$ is added and the criterion $MA(S, T)$ is replaced by correlated the total cost of maintenance activities $CMA(S, T)$. Thus, it can be observed that $CMA(S, T)$ is equal to $y = 0$ only if $MA(S, T) = y = 0$ and $CMA(S, T) > y$ only if $MA(S, T) > y = 0$. □

Thus, it is highly unlikely (until $P = NP$) to construct an exact polynomial time algorithm for the considered problem. On the other hand, the solution space $\mathbb{S}$ is significantly greater than for typical scheduling problems, since it is spread by the sequence of permutations with the cardinality $|\mathbb{S}| = O((n!)^T)$. Therefore, we will proposed a dedicated double crossover genetic algorithm (GARDX) and compare its efficiency with other methods already published (e.g., [5], [8]).

IV. Genetic algorithm

In this section, we will present our genetic algorithm railway double crossover (GARDX), which is based on the general concept of the genetic algorithm metaheuristic [11]. In particular, it follows the construction for one permutation [12], later on extended to the sequence of permutations being our earlier basic implementation GA (see [5]) evaluated for the total availability criterion TA. Let us bring closer the idea behind GARDX, which is describe by Algorithm 1. It uses the mathematical model developed and implemented solely by the author (R. Rudek) [7] and included in his dll library `RudRobustRailway` (RRR). Since GARDX is designed to optimize arbitrary criteria Q that are based on the maintenance plan resulting from the schedule, the proper handler has to be chosen (e.g., TA or TC) and also the values of its meta parameters have to be given, i.e., $PopulationSize$, $OffspringSize$, $MixSize$, $StopConition$ (see step 1). Each individual in the population is expressed as the pair (S, Q), where $S = \langle S_1, \ldots, S_t, \ldots S_T \rangle$ is the schedule represented by the sequence of permutations S_t and Q is the corresponding criterion value (e.g., calculated according to TA or TC). Thus, we will refer to the schedule S and to the criterion value Q of the idxth element from a required set (e.g., $Population$) as follows: $Population[idx].S$ and $Population[idx].Q$ (step 2).

Algorithm 1 Genetic Algorithm Railway Double Crossover GARDX

```
 1: Input: parameters and criterion type Q(·);
 2: Refer to the schedule S and the criterion Q of
    the element idx from the set SetName by
    SetName[idx].S, SetName[idx].Q;
 3: Population[1] = provided by FA;
 4: Population[2] = provided by BH;
 5: Population[3] = provided by PSO (StopCondition=10s);
 6: Population[4] = provided by SA (StopCondition=10s);
 7: for p = 5,...,⌊PopulationSize/2⌋
 8:   S' = a sequence of random permutations;
 9:   Calculate the criterion Q' for S': Q' = Q(S');
10:   Population[p] = S' and Population[p].Q = Q';
11: end for
12: for p = ⌊PopulationSize/2⌋ + 1,...,PopulationSize
13:   S' = Population[1].S;
14:   for t = 1,...,T
15:     for r = 1,...,⌊n/2⌋
16:       u = random vehicle index from {1,...,n};
17:       v = random vehicle index from {1,...,n};
18:       Swap jobs between u and v in S'_t
19:     end for
20:   end for
21:   Calculate the criterion Q' for S': Q' = Q(S');
22:   Population[p] = S' and Population[p].Q = Q';
23: end for
24: while (StopCondition ≠ true)
25:   ParentPool = {choose 2*OffspringSize random
      elements from Population};
26:   for idx = 1,...,OffspringSize
27:     ParentXX = ParentPool[2*idx − 1];
28:     ParentXY = ParentPool[2*idx];
29:     t_a = random period index from {1,...,T};
30:     t_b = random period index from {1,...,T};
31:     (S^X, S^Y) = TPX(ParentXX.S, ParentXY.S, t_a, t_b);
32:     S^C = ROX(S^X, S^Y, MixSize);
33:     Calculate the criterion value Q^C
        for the child schedule S^C: Q^C=Q(S^C);
34:     Offspring[idx].S=S^C and Offspring[index].Q=Q^C;
35:   end for
36:   Population = Choose PopulationSize best elements
      from the set {Population ∪ Offspring}
      ordered from the best to the worst Q;
37:   if Population[1].Q is better than Q* then
38:     Q* = Population[1].Q and S* = Population[1].S;
39:   end if
40: end while
41: Return the schedule S* = (S*_1,...,S*_T)
    and its criterion value Q*;
```

Algorithm 2 Two Point Crossover (TPX): $(S^X, S^Y) = TPX(S^{XX}, S^{XY}, t_a, t_b)$

```
 1: Input: parent schedules S^XX, S^XY and
    crossover parameters t_a, t_b;
 2: if t_a > t_b then exchange values t_a and t_b;
 3: for t = 1 to t_a − 1
 4:   S^X_t = S^XX_t;
 5:   S^Y_t = S^XY_t;
 6: end for
 7: for t = t_a to t_b
 8:   S^X_t = S^XY_t;
 9:   S^Y_t = S^XX_t;
10: end for
11: for t = t_b + 1 to T
12:   S^X_t = S^XX_t;
13:   S^Y_t = S^XY_t;
14: end for
15: Return resulting schedules S^X and S^Y;
```

Algorithm 3 Random Order Crossover (ROX): $S^C = ROX(S^X, S^Y, MixSize)$

```
 1: Input: parent schedules S^X, S^Y and
    parameter MixSize;
 2: Initialize the child schedule:
    S^C = ⟨S^C_1,...,S^C_t,...S^C_T⟩ = S^X;
 3: for q = 1,...,MixSize
 4:   t = random period index from {1,...,T};
 5:   for v = 1 to n
 6:     Get job index j assigned to vehicle v
        at period t for parent X: j = S^X_t(v);
 7:     Mark vehicle v in child schedule S^C_t as
        not assigned: S^C_t(v) = −1;
 8:     Mark job j as not assigned to the child:
        isJobAssigned[j] = false;
 9:     rv = random value from {0.0,...,0.9};
10:     if rv < 0.5 then
11:       Assign job j to vehicle v in S^C_t: S^C_t(i) = j;
12:       Mark j as assigned: isJobAssigned[j] = true;
13:     end if
14:   end for
15:   u = 1;
16:   for v = 1 to n
17:     Get job index j assigned to vehicle v
        at period t for parent Y: j = S^Y_t(v);
18:     if isJobAllocated[j] == false then
19:       while (S^C_t(u) > 0)
20:         u = u + 1;
21:       end while
22:       Assign job j to vehicle u in S^C_t: S^C_t(u) = j;
23:     end if
24:   end for
25: end for
26: Return the child schedule S^C = (S^C_1,...,S^C_T);
```

The first four individuals in the initial population (steps 3–6) are constructed by the following methods FA, BH, PSO and SA. Namely, Fixed Assignment (FA) is the constant assignment of jobs to vehicles during all months (no job reassignment between vehicles), i.e., a sequence of natural permutations (introduced as a default reference in [7], later on used in [5]). Balance Heuristic (BH) was developed and introduced in [7], later on used and shown to a broader audience in [9], whereas Particle Swarm Optimization (PSO) was presented in [8], and Simulated Annealing (SA) was introduced in [7], later on used in [5]. The applied PSO and SA optimize the criterion type chosen for GARDX and their meta parameter values are given in Section V, whereas their stop condition in GARDX is 10 seconds. Next $p = 5,\ldots,\lfloor PopulationSize/2 \rfloor$ individuals are the sequences of random permutations (steps 7-11). The rest of elements of the population $p=\lfloor PopulationSize/2 \rfloor+1,\ldots, PopulationSize$ are constructed on the basis of FA, where for each S_t jobs are swapped between two random vehicles u and v that are drawn $\lfloor n/2 \rfloor$ times (steps 12-23) for every $t=1,\ldots,T$.

In the each iteration of the main loop (steps 24-40), the algorithm chooses $OffspringSize$ random pairs from the population (step 25) that are the intermediate parents $ParentXX$ and $ParentXY$ (steps 27-28). Following them and for two random periods t_a and t_b (steps 29-30) the target parents are constructed represented by schedules S^X and S^Y (step 31). It is done by using the two point crossover TPX (Algorithm 2), which exchanges permutations of schedules

$ParentXX.S \equiv S^{XX}$ and $ParentXY.S \equiv S^{XY}$ in ranges $t \in \{t_a, \ldots, t_b\}$ (where $t_a < t_b$) returning the new schedules S^X and S^Y. On their basis, a new child schedule S^C is created following the random order crossover ROX and parameter $MixSize$ (step 32). The related procedure ROX described by Algorithm 3 initially clones the first parent schedule S^X to the child schedule S^C. Next, it draws $MixSize$ times random period indices $t \in \mathcal{T}$. For each t, the particular assignments of jobs to the vehicles from the permutation of the first parent S^X_t are copied with probability 0.5 to the related permutation of the child schedule S^C_t, whereas not chosen jobs are marked as unassigned. Subsequently, these missing jobs are assigned to unoccupied vehicles in S^C_t according to the order determined by the permutation of the second parent S^Y_t.

After the complete set $Offspring$ of $OffspringSize$ new schedules is generated, the new population is constructed. It is done by choosing $PopulationSize$ best solutions from the temporary set consisting the previous population and the new offspring, i.e., $\{Population \cup Offspring\}$ (step 36). If the new population contains a solution that is better than the best already found Q^*, then it is updated (steps 37-39). Note that *better* (step 37) for TA means that the new criterion value is greater than Q^*, whereas for TC means that it has to be smaller to be updated. Nevertheless, if the algorithm is implemented such that the availability TA is represented by positive and the cost TC by negative numbers, then we can unify the notation and focus only on the maximization, thereby the best found solution is updated only for greater values.

The computational complexity of a single iteration of GARDX (steps 24–40) is equal to $O(OffspringSize(MK+T)n)$ that is determined by TPX (step 31), ROX (step 32) and the calculation of $Q(\cdot)$ based on the maintenance plan (step 33), which are equal to $O(Tn)$, $O(Tn)$, $O(MKn)$, respectively.

V. Experiments

Let us now verify the accuracy of the proposed algorithm GARDX in reference to other methods implemented for similar problems. To provide reliable analysis of our algorithm, especially referring to the methods already described in the literature (particle swarm optimization PSO, simulated annealing SA, genetic algorithm GA), we will compare the scheduling algorithms for the instances constructed similarly as in the related works [5].[1] The set of all evaluated algorithms and their configurations are given as follows: Particle Swarm Optimization (PSO): $\omega = 0.6$, $c_1 = 0.5$, $c_2 = 0.5$, swarm size $m = 10$, bounds $X_{\min} = 0.0$, $X_{\max} = 4.0$, $\nu_{\min} = -4.0$, $\nu_{\max} = 4.0$ (following [8]); Simulated Annealing (SA): $Temp = 1000000$, $\alpha = 0.01$ (introduced in [7], later on used in [5]); Genetic Algorithm (GA): $PopulationSize = 200$, $OffspringSize = 50$, $MixSize = \lfloor 0.2T \rfloor$, (following [5]); Genetic Algorithm Railway Double Crossover (GARDX): $PopulationSize = 160$, $OffspringSize = 40$, $MixSize = \lfloor 0.2T \rfloor$. The initial solution (schedule) for each metaheuristic is determined by FA. To provide a fair comparison, due to nondeterministic nature of these methods, we run each of them 10 times and choose the best result (similarly as in [5]) as well as the stop condition for all metaheuristics is the same and set to 90s. Thus, we slightly extended the running times of PSO, SA and GA set in [8] and [5], giving them more time to traverse the search space.

[1]The algorithms were coded in C# (Microsoft Visual Studio 2022) and simulations were run under Windows 10 on PC, CPU Intel® Core™ i9-10885H 2.40GHz and 32GB RAM. All the corresponding software and algorithms, their previous as well as the current versions were coded solely by R. Rudek

TABLE I: Example preventive maintenance cycles, durations and costs of maintenance activities for Škoda 31E ($f = $ Š) and for E6ACTa Dragon 2 ($f = $ D) (see [10])

Type f	Level k	TBM^f_k	MBM^f_k [km]	MD^f_k [days]	MC^f_k [PLN*]
Š			Škoda 31E		
	1	14 days	2,500	0.125	1,000
	2	110 days	22,000	2	7,000
	3	32 months	220,000	21	50,000
	4	8 years	620,000	74	800,000
	5	16 years	1,200,000	84	1,000,000
D			E6ACTa Dragon 2		
	1	3 months	30,000	0.125	2,000
	2	1 year	150,000	2	12,000
	3	4 years	660,000	21	80,000
	4	8 years	1,220,000	74	1,000,000
	5	32 years	4,800,000	84	1,400,000

It is worth noticing that all the considered metaheuristics (PSO, SA, GA, GARDX) were developed and implemented solely by the author such that they can be applied for an arbitrary criterion based on his model developed in [7] and included in his dll library `RudRobustRailway` (RRR). Thereby, the only necessary action to optimize the required criterion by these algorithms is to choose and to apply a proper handler Q calculating the criterion value, e.g., referring to TA or TC; for Algorithm 1 see its step 33. Though our previous implementations SA, GA (e.g., [5]) and PSO [8] were originally used for different objective functions, they can be applied in the exactly same form to optimize also TA or TC (only by choosing a proper objective function handler).

Similarly as in [5] and [10], the computational experiments are provided for the instances constructed on the basis of two extreme types of vehicles representing one of the oldest (Škoda 31E series 181 produced till 1965) and one of the most modern locomotive (E6ACTa Dragon 2 produced since 2018) offered by rail vehicle rental companies in Europe. The maintenance cycles of these vehicles describing the time based TBM and the mileage based MBM maintenance policies, maintenance durations DM and costs MC are given in Table I. In the industrial practice, the maintenance activities of levels $k \geq 3$ start one day earlier than following from TBM^f_k [7]. Although the financial values are correlated with the market, they do not refer to any particular company. Moreover, if it is needed to analyse these values in US Dollars, the exchange rate to Polish Złoty (PLN) on 30 April 2025 was \$1 $\approx$ 3.77 PLN. The instances are based on $n \in \{10, 20\}$ vehicles of the same type

TABLE II: The total availability TA (in days), the improvement (additional days) δ_{TA} and the total loss (in days) Δ_{TA} to the best solution of the considered algorithms $A \in \{$FA, PSO, SA, GA, GARDX$\}$ for the analyzed instances

$T^{(y)}$ [years]	n	Measure [days]	FA	PSO	SA	GA	GARDX
Škoda 31E (series 181)							
5	10	TA	15,718	15,810	15,827	15,815	15,846
		δ_{TA}		92	109	97	128
		Δ_{TA}	128	36	19	31	0
	20	TA	31,448	31,546	31,568	31,585	31,624
		δ_{TA}		98	120	137	176
		Δ_{TA}	176	78	56	39	0
10	10	TA	31,060	31,268	31,330	31,356	31,367
		δ_{TA}		208	270	296	307
		Δ_{TA}	307	99	37	11	0
	20	TA	62,050	62,473	62,447	62,629	62,636
		δ_{TA}		423	397	579	586
		Δ_{TA}	586	163	189	7	0
15	10	TA	46,439	46,620	46,666	46,636	46,674
		δ_{TA}		181	227	197	235
		Δ_{TA}	235	54	8	38	0
	20	TA	92,877	93,026	93,168	93,061	93,183
		δ_{TA}		149	291	184	306
		Δ_{TA}	306	157	15	122	0
E6ACTa Dragon 2							
5	10	TA	17,820	17,911	17,914	17,914	17,915
		δ_{TA}		91	94	94	95
		Δ_{TA}	95	4	1	1	0
	20	TA	35,747	35,818	35,829	35,825	35,830
		δ_{TA}		71	82	78	83
		Δ_{TA}	83	12	1	5	0
10	10	TA	35,130	35,203	35,166	35,202	35,236
		δ_{TA}		73	36	72	106
		Δ_{TA}	106	33	70	34	0
	20	TA	70,360	70,407	70,370	70,381	70,458
		δ_{TA}		47	10	21	98
		Δ_{TA}	98	51	88	77	0
15	10	TA	52,519	52,602	52,555	52,519	52,625
		δ_{TA}		83	36	0	106
		Δ_{TA}	106	23	70	106	0
	20	TA	105,021	105,132	105,110	105,021	105,158
		δ_{TA}		111	89	0	137
		Δ_{TA}	137	26	48	137	0

TABLE III: The total maintenance cost TC, the financial savings δ_{TC} and the loss (in savings) Δ_{TC} to the best solution (all in thousand PLN, kPLN) of the considered algorithms $A \in \{$FA, PSO, SA, GA, GARDX$\}$ for the analyzed instances

$T^{(y)}$ [years]	n	Measure [kPLN]	FA	PSO	SA	GA	GARDX
Škoda 31E (series 181)							
5	10	TC	15,718	15,538	15,449	15,442	15,385
		δ_{TC}		180	269	276	333
		Δ_{TC}	333	153	64	57	0
	20	TC	31,906	31,518	31,243	31,311	31,200
		δ_{TC}		388	663	595	706
		Δ_{TC}	706	318	43	111	0
10	10	TC	36,482	33,563	36,132	33,269	33,181
		δ_{TC}		2,919	350	3,213	3,301
		Δ_{TC}	3,301	382	2,951	88	0
	20	TC	73,914	67,829	72,427	67,265	67,249
		δ_{TC}		6,085	1,487	6,649	6,665
		Δ_{TC}	6,665	580	5,178	16	0
15	10	TC	56,460	54,690	55,412	54,741	53,815
		δ_{TC}		1,770	1,048	1,719	2,645
		Δ_{TC}	2645	875	1,597	926	0
	20	TC	112,814	110,692	111,397	112,406	110,575
		δ_{TC}		2,122	1,417	408	2,239
		Δ_{TC}	2,239	117	822	1,831	0
E6ACTa Dragon 2							
5	10	TC	4,072	2,104	2,032	2,070	2,032
		δ_{TC}		1,968	2,040	2,002	2,040
		Δ_{TC}	2,040	72	0	38	0
	20	TC	6,152	4,210	4,074	4,150	4,074
		δ_{TC}		1,942	2,078	2,002	2,078
		Δ_{TC}	2,078	136	0	76	0
10	10	TC	15,784	13,660	13,730	13,664	13,604
		δ_{TC}		2,124	2,054	2,120	2,180
		Δ_{TC}	2,180	56	126	60	0
	20	TC	29,448	27,508	27,354	27,438	27,318
		δ_{TC}		1,940	2,094	2,010	2,130
		Δ_{TC}	2,130	190	36	120	0
15	10	TC	23,632	22,682	23,576	23,632	22,664
		δ_{TC}		950	56	0	968
		Δ_{TC}	968	18	912	968	0
	20	TC	47,424	46,354	47,114	47,424	46,272
		δ_{TC}		1,070	310	0	1,152
		Δ_{TC}	1,152	82	842	1,152	0

chosen from the considered locomotives. It is assumed that the simulations start at 01.01.2024 and at that day all vehicles are just after maintenance of the highest level K_f (overhaul) or they are just delivered from their producer, thereby $VT_{v,k}=0$ and $VD_{v,k}=0$ for $v=1,\ldots,n$ and $k=1,\ldots,K_f$ (identical parallel machines). The number of jobs is equal to the number of vehicles divided into $G=\lceil n/2 \rceil$ groups that share the same average mileage per day from range $w_j \in \{400,\ldots,700\}$ (see [5]). We consider periods referring to popular strategic planning time frame (from shorter to longer) $T^{(y)} \in \{5,10,15\}$ in years, i.e., $T \in \{60,120,180\}$ in months. The analysed algorithms $A \in \{$FA, PSO, SA, GA, GARDX$\}$ are evaluated for each instance I according to the following measures depending on the criterion $Q \in \{TA, TC\}$: the total availability of vehicles TA with $\delta_{TA}(A) \equiv \delta_{TA}(A,I) = TA(S^A) - TA(S^{FA})$ is the improvement (additional days) in reference to FA (higher values are better); $\Delta_{TA}(A) \equiv \Delta_{TA}(A,I) = TA^* - TA(S^A)$ is the total loss in days in reference to the best solution $TA^* = \max\{TA(S^A)\}$ for instance I (lower values are better); where $TA(S^A)$ is the total availability of vehicles (in days) for the schedule S^A provided by algorithm A; the total maintenance cost TC with $\delta_{TC}(A) \equiv \delta_{TC}(A,I) = TC(S^{FA}) - TC(S^A)$ is the improvement (financial savings) in reference to FA (higher values are better); $\Delta_{TC}(A) \equiv \Delta_{TC}(A,I) = TC(S^A) - TC^*$ is the total loss in savings in reference to the best solution $TC^* = \min\{TA(S^A)\}$ for instance I (lower values are better); where $TC(S^A)$ is the total maintenance cost for the schedule S^A provided by algorithm A. Note that Δ_{TA} and Δ_{TC} are redundant to δ_{TA} and δ_{TC}, respectively. Nevertheless, they are introduced for a more comprehensive comparison of the algorithms to clearly show differences between them and to easily recognize leading approaches (lower values are better).

The results of the computational experiments for instances based on the considered locomotive types, including criterion values TA (the total availability of vehicles in days) and TC (the total maintenance cost in thousand PLN), the improve-

ments to FA (δ_{TA}, δ_{TC}) and the losses to the best solutions (Δ_{TA}, Δ_{TC}) are shown in Tables II and III, respectively. It can be seen that all metaheuristics improved the results obtained by FA for both criteria. Our newly proposed GARDX overwhelms all other previous algorithms PSO, SA and GA for all considered instances under each of the separately optimized criteria: the total availability of vehicles TA and the total maintenance cost TC. It found schedules that offer additional measurable profits in reference to FA, e.g., from 128 additional days in 5 years and 10 locomotives to 586 days for 10 years and 20 locomotives (see Table II) and significant savings in the maintenance cost that exceed 2M PLN (c.a. \$0.5M) for $T^{(y)} = 5$ years and $n = 10$ locomotives or it can even reach 6M PLN (c.a. \$1.5M) for $T^{(y)} = 10$ and $n = 20$. The experiments revealed that GARDX is a robust method that well optimizes various instances, whereas other algorithms are instable. Therefore, GARDX can successfully replace each of the previous metaheuristics PSO, SA and GA for the optimization of TA or TA. Although the maximization of the total availability of vehicles TA is correlated with the minimization of the total maintenance cost TC, they are not equivalent. The optimization of TA does not need to lead to the identical improvement of TC, due to specifications and nonlinearity of these criteria. We have managed for Škoda instances to choose the costs of maintenance activities MC_k^f on particular levels (within the range of real values correlated with the market, see Table I) that allows us for the following analysis. Namely, for $T^{(y)} = 5$ and $n = 10$, FA obtained the maintenance plan that was characterized by TA equals to 15,718 days (Table II) and at the same time TC is equal to 15,718 thousand PLN (kPLN) (see Table III). Thus, we can see that the best result for TA is 15,846 days provided by GARDX and the ranking of algorithms is GARDX, SA, GA, PSO (see Table II), whereas the best value for TC is 15,385 kPLN obtained also by GARDX, but the ranking is different, i.e., GARDX, GA, SA, PSO (see Table III). GARDX optimizing TA for that instance was able to find the schedule S^{TA} resulting with a maintenance plan characterized by the best availability $TA(S^{TA}) = 15,846$ days (Table II) and the related total maintenance cost was $TC(S^{TA}) = 15,414$ kPLN. However, if GARDX optimized the criterion TC, then a different schedule S^{TC} was found corresponding to a maintenance plan characterized by the best total maintenance cost $TC(S^{TC}) = 15,385$ kPLN (Table III) and the related total availability in days was $TA(S^{TC}) = 15,826$. It can be seen that the total availability $TA(S^{TA}$ is better (greater) than $TA(S^{TC})$ if TA is optimized, and the total maintenance cost $TC(S^{TC})$ is better (smaller) than $TC(S^{TA})$ if the optimization objective is TC. Thereby, optimization of TA is not equivalent to optimization of TC.

VI. Conclusions

We analysed rolling stock cyclic job scheduling problems with maintenance to optimize the following objectives: the maximization of the total availability of vehicles TA and the minimization of the total maintenance cost TC. They were formulated as cyclic job scheduling problems on parallel machines with time and usage based deteriorating effects and maintenance. We constructed dedicated efficient genetic algorithm GARDX and we adjusted other metaheuristics PSO, SA, and GA known from the literature. The computational experiments revealed that GARDX overwhelms all compared methods and it proved its usefulness to the industrial practice to maximized TA and to minimize TC translating into convincing financial profits. For instance, it was able to significantly improve TA in reference to FA, e.g., from 128 additional days in 5 years and 10 locomotives to 586 days for 10 years and 20 locomotives. If it was applied to optimized TC, it found savings in reference to FA that exceeded 2M PLN (c.a. \$0.5M) for 5 years and 10 locomotives and 6M PLN (c.a. \$1.5M) for 10 years and 20 vehicles.

Acknowledgement

This work was supported by the Polish National Science Centre under grant no. DEC-2020/37/B/HS4/03235 and by General Kościuszko Military University of Land Forces (presentation and computational experiments). We would like to thank Anna Litwa-Janeczek and Tomasz Marchel for analysis concerning criteria and intermediate simulations of algorithms.

References

[1] M. Atsmony, B. Mor, and G. Mosheiov, "Minimizing tardiness scheduling measures with generalized due-dates and a maintenance activity," *Computers & Operations Research*, vol. 152, p. 106133, 2023.

[2] R. P. Nicolai and R. Dekker, "Optimal maintenance of multi-component systems: a review," in *Complex system maintenance handbook*. London: Springer Science & Business Media, 2008, pp. 263–286.

[3] X. Sun and X.-N. Geng, "Single-machine scheduling with deteriorating effects and machine maintenance," *International Journal of Production Research*, vol. 57, pp. 3186–3199, 2019.

[4] C. Zhang, Y. Gao, L. Yang, Z. Gao, and J. Qi, "Joint optimisation of train scheduling and maintenance planning in a railway network: A heuristic algorithm using Lagrangian relaxation," *Transportation Research Part B: Methodological*, vol. 134, pp. 64–92, 2020.

[5] R. Rudek and I. Rudek, "Models and algorithms for the preventive maintenance optimization of railway vehicles," *Expert Systems with Applications*, vol. 240, pp. 122 589.1–13, 2024.

[6] E. Eisenschmidt, S. Reimig, L. Schirmers, and S. Stern, "The rail sector's changing maintenance game," 2017, mcKinsey Global Institute, Report, https://tinyurl.com/ymbup3v5 [accessed: 2024-08-20].

[7] R. Rudek, "A decision support system for rolling stock management in a railway transport company," 2018, Industrial Report "Mozart Project" for Industrial Division Ltd., October 2017 – September 2018 (in Polish).

[8] I. Rudek and R. Rudek, "Optimizing maintenance cost of uniform rolling stock by scheduling algorithms," in *Lecture Notes in Networks and Systems*, L. Borzemski, H. Selvaraj, and J. Swiatek, Eds. Switzerland: Springer Nature, 2022, vol. 364, pp. 261–270.

[9] I. Heppner and R. Rudek, "Improve railway vehicle availability by scheduling under preventive maintenance policies," in *48th International Conference on Computers & Industrial Engineering 2018 (CIE48)*, X. Xun, M. I. Dessouky, and R. Y. Zhong, Eds. Computers and Industrial Engineering, 2018, pp. 2325–2332.

[10] R. Rudek, "Local search algorithms for a railway scheduling problem under maintenance and cost criteria," in *Models and Methods for Systems Engineering. Studies in Big Data*, G. Borowik, G. Chmaj, and R. Waszkowski, Eds. Cham: Springer, 2025, vol. 165, pp. 39–49, https://doi.org/10.1007/978-3-031-76440-0_4.

[11] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. New York: Springer-Verlag, 1996.

[12] R. Rudek, "A generic optimization framework for scheduling problems under machine deterioration and maintenance activities," *Computers & Industrial Engineering*, vol. 174, pp. 108 800.1–22, 2022.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 381–386
DOI: 10.15439/2025F5371
ISSN 2300-5963 ACSIS, Vol. 43

Symbolic vs Black-Box Explanations: A Model-Driven Approach Using Grammatical Evolution

Dominik Sepioło, Antoni Ligęza
0000-0001-7746-3781, 0000-0002-6573-4246
AGH University of Krakow,
Department of Applied Computer Science,
al. Adama Mickiewicza 30, 30-059 Kraków, Poland
Email: {sepiolo, ligeza}@agh.edu.pl

***Abstract*—Black-box explainability tools like LIME and SHAP are widely used to interpret machine learning models. However, their post-hoc, local nature often results in inconsistent and semantically opaque explanations. This paper presents a model-driven explainability approach using grammatical evolution (GE), enabling the discovery of symbolic, human-readable models. We compare black-box explanations to symbolic GE-generated models on two benchmark tasks: a quadratic equation classification problem and the Iris dataset. GE produces interpretable, consistent, and semantically meaningful expressions consistent with domain knowledge, offering a more trustworthy foundation for explainable AI. The use of Meaningful Intermediate Variables (MIVs) further improves the clarity and expressiveness of the symbolic models.**

***Keywords*—Explainable Artificial Intelligence (XAI); Grammatical Evolution; Symbolic Regression; Model-Driven XAI; Black-Box Explanations; Post-Hoc Explainability; Transparent AI Models**

I. Introduction

THE INCREASING deployment of machine learning (ML) systems in critical domains has renewed attention to the need for transparency, interpretability, and trust in artificial intelligence (AI) models. In practical applications, such as biomedical diagnostics and regulatory environments, it is not sufficient for a model to produce a correct prediction; it must also be able to explain *how and why* it provided a decision in a way understandable to human users [1], [2].

Explainable Artificial Intelligence (XAI) aims to make AI systems more interpretable and trustworthy [1], [3]. XAI methods include inherently transparent models (e.g. decision trees) and post hoc techniques such as LIME [4] and SHAP [5], which approximate black-box models locally. However, post-hoc explanations are often inconsistent, sensitive to sampling, and lack semantic grounding in the domain [6], [7].

These challenges have led to a growing interest in *Model-Driven* approaches to XAI, which construct interpretable models directly from data using symbolic techniques [8], [9]. Among them, Grammatical Evolution (GE) is a promising evolutionary algorithm that evolves human-readable expressions guided by context-free grammars [10]. GE has shown effectiveness in symbolic regression, functional modeling, and equation discovery - especially when domain knowledge is embedded through grammar design or Meaningful Intermediate Variables[1] (MIV) [11], [13].

This paper presents a comparative exploration of symbolic model-driven explanations using GE and black-box post-hoc tools. We argue that symbolic models offer greater semantic clarity, stability, and alignment with human reasoning. Using two benchmarks, a synthetic quadratic classification task and the Iris dataset, we show how GE produces simple, meaningful models that either recover known functional relationships (e.g. $d = b^2 - 4ac$) or leverage interpretable MIVs (e.g. petal area) to achieve high accuracy with full transparency.

Our main contributions are as follows:

- We contrast symbolic and black-box explanations on two classification tasks, highlighting the strengths and weaknesses of each approach.
- We show that GE provides clear, interpretable models aligned with domain knowledge, unlike the often unstable outputs of LIME and SHAP.
- We demonstrate the use of MIVs to enhance interpretability and simplify the model structure.
- We propose a methodology for integrating GE into a transparent MD-XAI pipeline.

The remainder of the paper is organized as follows. Section 2 outlines the background and methods. Section 3 presents experiments. Section 4 discusses the results and implications, and Section 5 concludes with insights and future work.

II. Related Work and Methods

This section outlines the two primary paradigms of explainability considered in this study: post-hoc explanation methods for black-box models, and model-driven symbolic approaches based on Grammatical Evolution. We discuss their principles, strengths, and known limitations, focusing on their impact on semantic clarity, model transparency, and alignment with domain knowledge.

[1] A Meaningful Intermediate Variable (MIV) has clear semantics, is derived from simple functions, and depends on data or other MIVs.

Topical area: Advanced Artificial Intelligence in Applications

A. Black-Box Explanation Techniques

Post-hoc explainability interprets black-box models (e.g. ensembles, neural networks) by analyzing the output without accessing internal structure. Two widely used methods are LIME [4] and SHAP [5]:

- **LIME** fits a local linear surrogate model by perturbing input data to explain predictions.
- **SHAP** uses game theory to calculate feature contributions by calculating Shapley values.

Although popular for their visual intuitiveness, LIME and SHAP face intrinsic key limitations [7], [3], [6]:

- *Instability:* Results vary with random seeds or sampling.
- *Locality:* Explanations are instance-specific and are not generalizable.
- *Poor semantics:* Attributions lack domain or causal grounding.
- *Explainer opacity:* The logic behind them is often difficult to validate.

These weaknesses have motivated the development of model-driven alternatives that prioritize transparency and interpretability by discovering and designing structures [15], [7], [8].

B. Symbolic Regression with Grammatical Evolution

Symbolic regression seeks interpretable mathematical expressions that fit the data accurately. Unlike classical regression with fixed structures (e.g. linear, polynomial), it explores a broad space of formulas composed of variables, constants, and operators, producing models that are directly interpretable and verifiable by humans.

Model-driven explainability seeks to construct models that are inherently interpretable. Among symbolic machine learning techniques, GE provides a flexible and powerful approach to discovering human-readable models [10].

GE is a genetic programming technique that uses context-free grammars to define valid symbolic expressions. Solutions are encoded as genomes that are decoded into mathematical formulas or rules based on the grammar.

- **Expressive Power:** GE supports arithmetic, logic, and domain-specific operations via customizable grammars.
- **Interpretability:** The output models are human-readable symbolic expressions.
- **Incorporation of Domain Knowledge:** Grammar rules or MIVs embed prior knowledge into the search space.

This approach has been successfully applied in various domains, including bioinformatics [9], symbolic regression [10], [12], and model-driven XAI [11], [8], [13]. GE discovers interpretable functions that reflect the data structure while allowing integration of constraints and semantic guidance.

Using user-defined grammars, GE can enforce domain-specific constraints or promote particular structural patterns. For instance, intermediate variables or known transformations can be explicitly included in the grammar, guiding the search process toward semantically valid and explainable models. This makes GE especially suitable for applications that require compliance with expert knowledge or regulatory compliance.

Our prior work [13] showed that GE can rediscover symbolic relationships (e.g. the quadratic discriminant) and generate compact and transparent classifiers. For example, using the hand-made variable $PLmPW = \text{petal.length} \cdot \text{petal.width}$, GE accurately classified the Iris species with high accuracy and full transparency.

Recent research explores hybrid methods that combine GE with neural networks or extract symbolic rules from black-box models [14].

C. Comparison Criteria

In this work, we evaluate post-hoc and symbolic explanation methods based on the following key dimensions:

- **Fidelity:** Accuracy of the explanation relative to the model or the true function.
- **Simplicity:** Complexity of the resulting explanation (e.g. expression length, tree depth).
- **Semantic Alignment:** Degree to which the explanation uses meaningful or domain-relevant variables.
- **Stability:** Consistency of the explanation across multiple runs or perturbed inputs.

D. Incorporating GE into XAI Pipelines

To formalize the use of GE as an alternative to post-hoc explainers, we propose a transparent pipeline for integrating it into explainability workflows. This replaces black-box interpretation with directly evolved, interpretable models:

1) **Feature Preprocessing:** Prepare input data; optionally derive MIVs.
2) **Grammar Specification:** Define a context-free grammar with domain semantics.
3) **GE Model Training:** Apply GE to evolve models mapping features/MIVs to outputs.
4) **Symbolic Rule Extraction:** Select models according to fitness, simplicity, and semantic alignment.
5) **Validation and Interpretation:** Assess accuracy and interpret model structure.

Figure 1 illustrates the structure of this methodology.

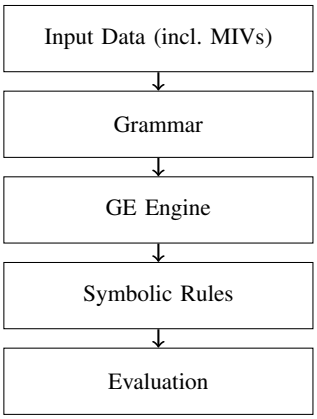


Fig. 1. Pipeline for symbolic model construction using GE

III. Experimental Comparison

To compare black-box explanation methods with symbolic GE models, we performed experiments on two tasks: a synthetic quadratic discriminant classification and the Iris dataset, a standard benchmark for interpretability studies.

A. Quadratic Classification Task

This task involves a simple symbolic function: the discriminant of a quadratic equation $d = b^2 - 4ac$ to classify input into two ($d > 0$), one ($d = 0$) or no real solutions ($d < 0$).

The dataset contains synthetic (a, b, c) tuples labeled with the corresponding value of d.

In the initial phase of the experiment, we trained models to classify the number of real roots using only (a, b, c), without providing $d = b^2 - 4ac$. This tested whether standard methods, linear regression, decision trees, and random forest (RF), could learn the decision boundary from data alone. The performance of the model was assessed using two metrics: RMSE (Root Mean Squared Error), defined as $\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, and its normalized form, NRMSE, $NRMSE = \frac{RMSE}{\bar{y}}$.

The prediction accuracy is presented in Table I. Although random forest had the lowest test error, none of the models produced interpretable rules or captured the underlying structure. Their generalization, especially near $d = 0$, was unstable.

TABLE I
RMSE and NRMSE of selected machine learning models trained on (a, b, c) without access to the discriminant

Model	RMSE (train)	NRMSE (train)	RMSE (test)	NRMSE (test)
Linear Regression	0.8267	0.8026	0.8239	0.8368
Decision Tree	0.5431	0.5273	0.8142	0.8224
Random Forest	0.2284	0.2218	0.5883	0.5942

This motivated the introduction of a domain-informed variable representing the discriminant to evaluate the benefits of symbolic modeling and explainability.

a) LIME and SHAP Results: We applied both explanation tools to the RF classifier and observed notable inconsistencies in predictions:

- LIME frequently attributed great importance to the variable b, with fewer or inconsistent roles for a and c.
- SHAP produced attributions in which c often appeared dominant, sometimes contradicting LIME.

Neither method revealed the discriminant formula or suggested that a quadratic combination of a, b, and c was relevant.

As shown in Fig. 2, LIME and SHAP sometimes disagreed on both the importance and direction of the feature. Similar instability was observed in [13], where shallow explanations lacked semantic consistency.

b) GE Results: In contrast, the GE model directly rediscovered the symbolic form of the discriminant: $d = b^2 - 4ac$ GE then generated an interpretable rule-based classifier:

$$\texttt{ifelse}(d > 0, \text{"two"}, \texttt{ifelse}(d == 0, \text{"one"}, \text{"zero"}))$$

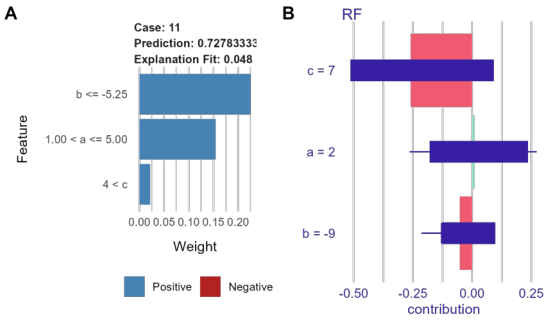


Fig. 2. (**A**) LIME and (**B**) SHAP explanations for a given prediction

This compact and interpretable structure is exactly consistent with the known semantics of the problem. Figure 3 illustrates the classification process: calculate the discriminant d, then assign the class based on its value.

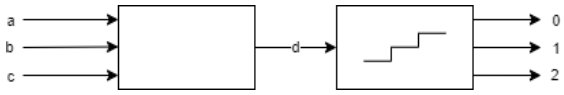


Fig. 3. GE-evolved symbolic rule with intermediate variable d representing the discriminant. The rule exactly matches the theoretical solution structure.

The symbolic rule for d was discovered by evolutionary search on a grammar encoding arithmetic operations and constants relevant to the discriminant:

```
<expr> ::= <var> | <op>(<n>, <var>)
    | <op>(<var>, <n>) | <op>(<expr>, <expr>)
<op>   ::= "-" | "*" | "^"
<var>  ::= a | b | c
<n>    ::= 1 | 2 | 3 | 4
```

The grammar for the final classification, which assigns the number of solutions of a quadratic equation, is listed below:

```
<result>     ::= ifelse(<expr>, "one", "other")
<expr>       ::= (<expr> & <sub_expr>)
        | (<expr> | <sub_expr>) | <sub_expr>
<sub_expr>   ::= <comparison>(<var>, <func_var>)
<comparison> ::= > | < | ==
<func_var>   ::= <num> | <var> | <func>(<var>)
<func>       ::= mean | max | min | sd
<var>        ::= a | b | c | d
<num>        ::= 0 | 1
```

The GE model achieved 100% accuracy on the training and test sets due to perfect alignment with the true decision boundary. In contrast, the random forest misclassified 4–5% near $d = 0$ and failed to generalize. This case demonstrates that GE provides strong performance and formal transparency, providing stable, reproducible, and verifiable expressions, unlike post-hoc black-box explainers. This experiment highlights the value of MIVs, like d, which capture complex feature relationships. MIVs reflect domain knowledge and decompose learning into interpretable parts. As shown in [13], they help GE efficiently discover accurate and human-readable rules, which post-hoc tools such as LIME or SHAP cannot achieve due to their lack of structural abstraction.

B. Iris Dataset

The Iris dataset includes 150 samples labeled: Setosa, Versicolor, or Virginica, each with four features: sepal length/width and petal length/width. It is commonly used to assess the interpretability of the model.

a) LIME and SHAP Results: RF trained on the Iris dataset was explained with LIME and SHAP. Compared to the quadratic task, the results were more consistent. Both tools often highlighted *Petal.Length* and *Petal.Width*, though discrepancies remained. As shown in Fig. 4, LIME negatively emphasized all features, while SHAP prioritized petal features over sepal ones.

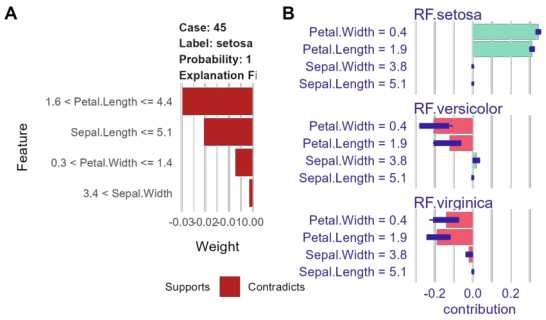


Fig. 4. (**A**) LIME and (**B**) SHAP explanations for the same Iris sample

Despite improved consistency compared to the delta experiment, neither method offered a compact or semantically meaningful rule that could generalize throughout the dataset.

b) GE Results: GE was applied to the full feature set to minimize misclassifications, running for 1000 iterations with default settings and the grammar shown below:

```
<expr>       ::= (<expr> & <sub_expr>)
       | <expr> | <sub_expr>) | <sub_expr>
<sub_expr>  ::= <comparison>(<var>, <func_var>)
<comparison> ::= > | < | == | >= | <=
<func_var>   ::= <num> | <var> | <func>(<var>)
<func>       ::= mean | max | min | sd
<var>        ::= Sepal.Length | Sepal.Width
               | Petal.Length | Petal.Width
<num>        ::= 1 | 1.5 | 2 | 2.5 | 3 | 4 | 5
```

Representative symbolic models and their corresponding plots are shown below, highlighting the effective separation of the *setosa* and *versicolor* classes.

```
ifelse(Petal.Length <= Sepal.Width, "setosa",
"other")

ifelse((Petal.Width <= 1.5)&(Petal.Length >= 2.5),
"versicolor", "other")
```

Although the model identified all *setosa* samples, its classification of *versicolor* at 94.66% accuracy indicates potential for further improvement.

To improve the classification, we extended the feature set by introducing a Meaningful Intermediate Variable that approximates the petal area:

$$PLmPW = petal.length * petal.width \quad (1)$$

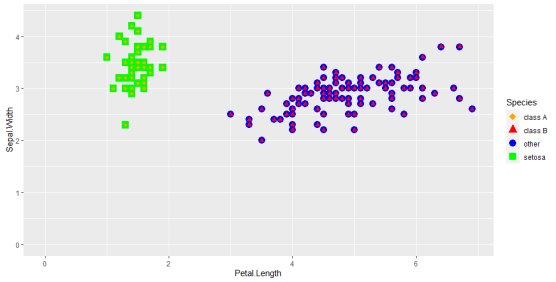


Fig. 5. *Setosa* classification with sepal width and petal length

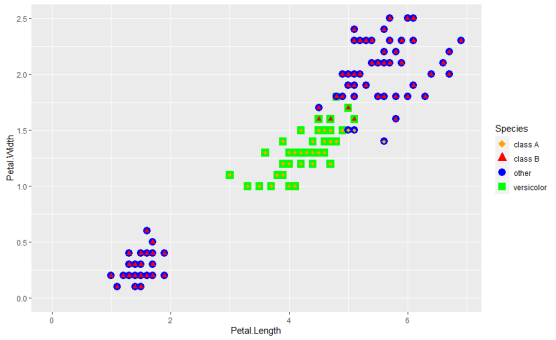


Fig. 6. *Versicolor* classification with petal width and petal length

Using this MIV, GE developed a simple and accurate rule-based model:

- If $PLmPW \leq 2$: Setosa.
- If $2 < PLmPW \leq 8$: Versicolor.
- If $PLmPW > 8$: Virginica.

This rule achieved approximately 95% precision, with a decision structure that is semantically meaningful and easy to understand.

Figure 7 presents the functional model obtained by combining the individual rules.

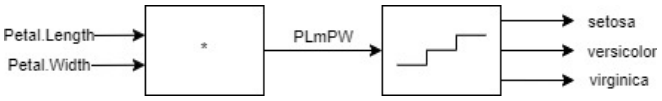


Fig. 7. Functional model for Iris species classification with MIV

The results of the classification process performed with the functional model provided are depicted in Figure 8.

We also introduced additional MIVs including the sepal area:

$$SLmSW = sepal.length * sepal.width, \quad (2)$$

petal and sepal perimeters, and various length-to-width ratios to improve model accuracy.

MIVs used in the model may be derived not only from the original input features but also from other previously defined MIVs. One such example is the petal-to-sepal area ratio, proposed using the MIVs defined in Equations 1 and 2:

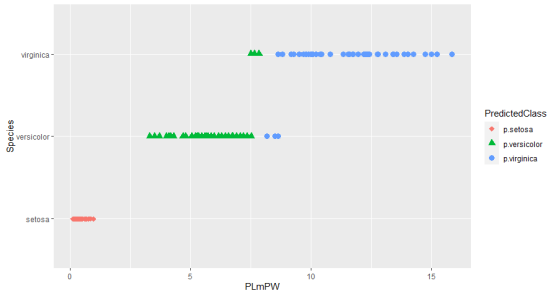


Fig. 8. Iris species classification with $MIV = petal.length * petal.width$

$$PdS = \frac{PLmPW}{SLmSW} \tag{3}$$

Using the GE approach, the following classification rules were generated on the basis of the derived PdS value:

```
ifelse(PdS <= 0.2, "setosa", "other")
ifelse((PdS <= 0.43) & (PdS >= 0.22),
"versicolor", "other")
ifelse(PdS >= 0.43, "virginica", "other")
```

These rules define a functional model, presented in Figure 9.

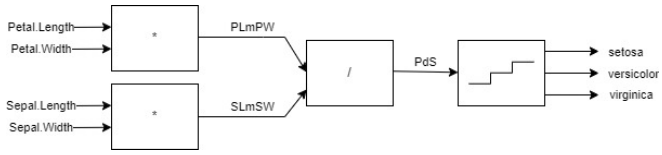


Fig. 9. Functional model for Iris classification with multiple MIVs

The classification results obtained using the functional model based on multiple MIVs are presented in Figure 10.

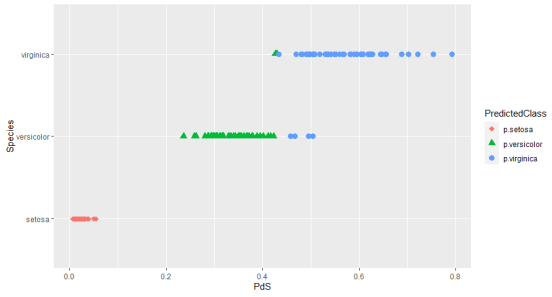


Fig. 10. Iris species classification with multiple MIVs

The model demonstrates an improvement over the previous one, achieving an accuracy of 96%.

IV. Analysis and Discussion

The experimental results highlight differences between post-hoc explainability techniques and model-driven symbolic methods. Our comparative analysis focuses on four critical dimensions: stability, interpretability, semantic alignment, and predictive performance.

A. Stability and Reproducibility

An intrinsic limitation of post-hoc methods is the *instability*. Attribution values vary across runs or small input changes, especially in the quadratic task. For example, LIME overemphasized b, while SHAP inconsistently prioritized c, reducing reliability in high-stake settings.

In contrast, GE consistently produced models with stable structure and clear semantics across runs, offering strong support for reproducibility, system auditing, and model understanding.

B. Interpretability and Semantic Coherence

GE produces symbolic models based on mathematical or logical rules that are directly interpretable. In the quadratic task, it rediscovered $d = b^2 - 4ac$. In the Iris dataset, MIV-based rules (e.g. petal area, petal-sepal ratios) reflect domain structure and improved transparency.

LIME and SHAP offer intuitive local insights but lack structural integration. Their explanations are not reusable and generalizable, limiting their value for model critique and design.

C. Role and Value of Intermediate Variables

MIVs significantly enhanced interpretability and performance. In the Iris task, handcrafted variables like `PLmPW` and `PdS` helped GE derive simple domain-aligned rules by embedding biological knowledge into the search process. We also showed that MIVs can be composed hierarchically—for example, `PdS` combines two others, supporting scalable symbolic modeling in complex domains with important but implicit constructs.

D. Prediction Accuracy

Although symbolic modeling prioritizes interpretability, GE models also achieved high precision. In the quadratic task, GE accurately captured the decision boundary. For the Iris dataset, MIV-based models reached 96% accuracy—matching black-box classifiers, but with much greater transparency. In particular, some gains came from structured variables, emphasizing the value of hybrid approaches that combine domain knowledge with symbolic learning.

E. Limitations and Practical Considerations

Despite its advantages, GE has practical constraints. Grammar design requires domain expertise, and poor design either limits the search space excessively or fails to guide the model toward meaningful solutions. As a stochastic method, GE may require multiple runs or tuning. However, in domains with expert input or bounded problems, GE offers a controllable and verifiable alternative to opaque black-box explanations.

The comparative evaluation described in Section II-C established four key criteria: reproducibility, interpretability, semantic consistency, and structural clarity. These dimensions provided a principled framework for analyzing the strengths and limitations of each method. The experimental findings validate the relevance of this framework: post-hoc methods

like LIME and SHAP showed weaknesses in reproducibility and semantic alignment, often generating inconsistent or context-insensitive attributions. In contrast, the symbolic models developed through GE satisfied all four criteria. They were reproducible across runs, structurally coherent, interpretable by domain experts, and semantically aligned with known functional relationships. The use of this criteria-driven perspective was essential in highlighting that precision alone is insufficient when evaluating explainable models, and structural and semantic properties must also be considered.

To consolidate the results and observations discussed in the experimental sections, Table II provides a qualitative comparison of the three approaches evaluated. LIME, SHAP, and Grammatical Evolution. The comparison is organized around the four criteria defined in Section II-C: fidelity, simplicity, semantic alignment, and stability.

TABLE II
QUALITATIVE COMPARISON OF EXPLANATION METHODS

Criterion	LIME	SHAP	GE
Fidelity	Low–Medium	Low-Medium	High (in known domains)
Simplicity	Medium	Medium	High (compact rules)
Semantic Alignment	Low	Low	High (uses MIVs)
Stability	Medium	Medium	High

As shown in the table, LIME and SHAP provide partial fidelity and limited semantic coherence, particularly in tasks where the decision boundary depends on specific functional structures. Their attributions are also sensitive to sampling variability and lack cross-instance consistency. GE, by contrast, produces symbolic models that offer high fidelity when the target rule is recoverable, and do so using semantically meaningful constructs such as intermediate variables. These symbolic rules are compact and structurally transparent, providing an interpretable basis for classification decisions. This distinction is particularly evident in the quadratic task, where GE recovered the exact mathematical discriminant, while LIME and SHAP failed to detect the underlying structure.

V. CONCLUSION

This study presents a comparative evaluation of black-box post-hoc explanation methods (LIME and SHAP) and symbolic, model-driven approaches using Grammatical Evolution. In two classification tasks, GE consistently demonstrated superior semantic clarity, interpretability, and reproducibility.

The core findings are as follows:

- **Interpretability and Semantic Alignment:** GE-generated symbolic models are inherently interpretable, expressed as explicit formulas aligned with domain semantics (e.g., discriminants or feature ratios). In contrast, LIME and SHAP offer fragmented, instance-specific attributions that are often semantically opaque and inconsistent across similar inputs.
- **Structured Abstraction through MIVs:** GE incorporates structured domain knowledge via Meaningful Intermediate Variables, such as petal area or petal-to-sepal ratios. These enhance model clarity, simplify decision boundaries, and support composability by enabling hierarchical MIVs built from previously defined variables.
- **Accuracy and Trustworthiness:** While GE emphasizes interpretability, it also achieves competitive accuracy—up to 96% on the Iris dataset, generating compact and transparent rules. This combination improves trust and auditability in high-stakes or regulated applications.

These results suggest that symbolic regression techniques, and particularly GE, provide a viable alternative to post-hoc explainers in contexts where explanation fidelity and transparency are critical.

Future work will focus on intelligent and automated grammar generation, discovering intermediate variables, and improving convergence strategies for grammar-guided evolution.

REFERENCES

[1] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, Elsevier (2019), https://doi.org/10.1016/j.inffus.2019.12.012.

[2] Guidotti, R., Monreale, A., Ruggieri, S., et al.: A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* **51**(5), 1–42 (2019), https://doi.org/10.1145/3236009.

[3] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**(1), 18 (2021), https://doi.org/10.3390/e23010018.

[4] Ribeiro, M. T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *KDD 2016*, pp. 1135–1144. ACM (2016), https://doi.org/10.1145/2939672.2939778.

[5] Lundberg, S. M., Lee, S. I.: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* **30**, pp. 4765–4774 (2017), https://dl.acm.org/doi/10.5555/3295222.3295230.

[6] Sepioło, D., Ligęza, A.: Towards Explainability of Tree-Based Ensemble Models: A Critical Overview. In: *New Advances in Dependability of Networks and Systems*, pp. 287–296. Springer (2022), https://doi.org/10.1007/978-3-031-06746-4_28.

[7] Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019), https://doi.org/10.1038/s42256-019-0048-x.

[8] Ligęza, A., et al.: Explainable Artificial Intelligence. Model Discovery with Constraint Programming. In: *ISMIS 2020*, Springer, pp. 171–191 (2020), https://doi.org/10.1007/978-3-030-67148-8_13.

[9] Hu, T.: Can Genetic Programming Perform Explainable Machine Learning for Bioinformatics? In: *Genetic and Evolutionary Computation*, Springer, pp. 63–77 (2020), https://doi.org/10.1007/978-3-030-39958-0_4.

[10] Ryan, C., O'Neill, M., Collins, J. J. (eds.): *Handbook of Grammatical Evolution*. Springer (2018), https://doi.org/10.1007/978-3-319-78717-6.

[11] Sepioło, D., Ligęza, A.: Towards Model-Driven Explainable Artificial Intelligence. An Experiment with Shallow Methods Versus Grammatical Evolution. In: *ECAI 2023 Workshops*, Springer, pp. 360–365 (2024), https://doi.org/10.1007/978-3-031-50485-3_36.

[12] Orzechowski, P., La Cava, W., Moore, J.H.: Where Are We Now? A Large Benchmark Study of Recent Symbolic Regression Methods. IN: *GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference*, Association for Computing Machinery, pp. 1183–1190 (2018), https://doi.org/10.1145/3205455.3205539.

[13] Sepioło, D., Ligęza, A.: Towards Model-Driven Explainable Artificial Intelligence: Function Identification with Grammatical Evolution. IN: *Applied Sciences* 14, 5950, (2024), https://doi.org/10.3390/app14135950.

[14] Tsoulos, I. G., Tzallas, A., Karvounis, E.: Using Optimization Techniques in Grammatical Evolution. In: *Future Internet*, 16, 172, (2024), https://doi.org/10.3390/fi16050172.

[15] Ligęza, A.: An experiment in causal structure discovery. A constraint programming approach. In: In: *ISMIS 2017*, Springer, pp. 261-268 (2017), https://doi.org/10.1007/978-3-319-60438-1_26.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 387–392
DOI: 10.15439/2025F3421
ISSN 2300-5963 ACSIS, Vol. 43

Comparative Analysis of AI Software and Traditional Software GitHub Repositories Using Process Mining

Oguzhan Tasci, Tugba Gurgen Erdogan
0009-0009-2549-758X, 0000-0003-1491-8739
Hacettepe University, Department of Computer Engineering
Beytepe Campus, 06800 Ankara/Türkiye
Email: oguzhantasci5561@gmail.com, tugba@cs.hacettepe.edu.tr

***Abstract*—This study conducts a comparative analysis of Artificial Intelligence (AI) software and traditional GitHub repositories, focusing on workflow efficiency and sentiment dynamics. Using process mining and sentiment analysis techniques, we examine repositories from eight prominent projects, encompassing diverse datasets of issues and pull requests filtered for relevance and consistency. Our findings reveal that AI software repositories exhibit different workflow patterns and sentiment dynamics compared to traditional repositories. Sentiment analysis uncovers that contributors to AI software repositories experience more positive sentiment dynamics, likely reflecting structured workflows and collaborative tools. Conversely, traditional repositories exhibit longer resolution times and more fluctuating sentiment patterns, which may indicate higher complexity or less automation. These insights provide valuable recommendations for optimizing repository management, fostering contributor satisfaction, and improving collaborative software development environments.**

I. Introduction

MANAGING software development processes that are multi-disciplinary, human-intensive, and prone to change requires effective tools. Git, a free and open-source distributed version control system, is widely used to manage projects of all sizes with speed and efficiency. GitHub, an open-source hosting platform, extends Git's capabilities by offering cloud-based version control and collaboration features. As one of the leading data sources for software development research [1], GitHub enables developers to manage and store their code while providing a wealth of information for analyzing software development processes, particularly in collaborative software projects.

As artificial intelligence (AI) has become integral to modern software systems, understanding the development processes of AI software systems has emerged as a significant area of research in software engineering [2]. GitHub provides rich data, including information on issues, commits, projects, users, comments, and the defined GitHub development flow.

Process mining is a data science technique that extracts process insights from event logs stored in information systems [3]. GitHub data, generated during the software development can be converted to event logs to enable process discovery and analysis. By identifying elements of the actual software development process, such as issues, pull requests (PR), and commits, along with their descriptive attributes, process deviations and inefficiencies can be systematically investigated.

Mining Software Repositories (MSR) techniques focus on extracting and analyzing data from software repositories to uncover interesting, useful, and actionable insights about software systems. MSR is considered one of the fastest-growing and most compelling areas in the field of software engineering [4]. The literature contains numerous studies focused on mining GitHub repositories to analyze software development processes for specific purposes. These include detecting anomalies in PR processes [5], understanding and improving students' project workflows [6], [7], [8], managing issues in AI software repositories [9], and analyzing the effects of continuous integration and continuous deployment (CI/CD) practices on repositories [10]. Several tools such as GitHubDataViz [11], GitLab Analyser [6], and Git Truck [12] have been introduced to support the analysis of software development data by visualizing Git issue and commit data.

In this study, we present a GitHub-based data collection framework for process mining and an empirical analysis comparing actual software development processes using process mining techniques and sentiment dynamics in AI software and traditional software repositories. Specifically, this research explores whether AI software projects exhibit distinct patterns in issue resolution, PR management, and contributor sentiment compared to traditional repositories.

II. Related Work and Background

Poncin et al. [13] demonstrated the application of process mining in software repositories, facilitating the analysis of developer roles and bug report handling. Expanding on this, Macak et al. [7] used process mining techniques for the Git log of students' projects to understand their development processes.

Sentiment analysis has become an important tool in understanding the emotional undertones in software development processes. Comments of repository artifacts, such as issues, PRs, and code, contain unstructured and valuable data to understand the sentiment dynamics that impact workflow efficiency and collaboration.

Jurado and Rodriguez [14] conducted a study that explored the use of sentiment analysis to monitor distributed teams, highlighting how emotional analysis of developer-written issues and tickets can provide a richer understanding of the development environment. Yang et al. [15] carried out an empirical study specifically focused on sentiment analysis in the GitHub repositories, revealing that the emotional tone of comments significantly affects the speed of bug fixing. Guzman et al. [16] identified patterns indicating higher negative sentiment in Java-based projects and more positive sentiment in geographically distributed teams. Research into developer behavior in open-source projects indicates a strong correlation between developer activities and sentiment shifts, which can serve as predictors of project success [17]. Robinson et al. [17] provided insights into how these factors influence repository health and productivity. Sinha et al. [18] also analyzed the sentiment in commit logs, exploring the emotional undertones in commits and their impact on project outcomes. While these studies offer valuable perspectives on developer sentiment, our research differentiates itself by investigating how the development processes of AI software products might influence these sentiments.

Zhou Yang et al. [9]'s empirical study on issue management in AI software repositories highlighted challenges such as runtime errors and unclear instructions.

Our study extends previous studies to a broader range of repository artifacts, including issues, code comments, and commit messages, to provide a more holistic view of sentiment dynamics in software development processes, which are discovered through process mining techniques.

III. Methodology

A. Repository Selection and Data Collection

Data was collected from publicly available GitHub repositories using the GitHub REST API. AI software repositories (TensorFlow, PyTorch, scikit-learn, Keras) focus on machine learning algorithms and frameworks. Traditional repositories (Angular, React, Bootstrap, Node.js) emphasize conventional software technologies. The selected repositories are listed in Table I.

TABLE I: Selected Repositories for Analysis

Repository	Type	Stars	Issues	PRs
TensorFlow [19]	AI software	187k	37k	31.1k
PyTorch [20]	AI software	86.1k	31.1k	85.7k
Scikit-Learn [21]	AI software	60.8k	9.9k	16.9k
Keras [22]	AI software	62.4k	11.9k	7.3k
Node.js [23]	Traditional	109k	16.1k	33.4k
Bootstrap [24]	Traditional	171k	23.9k	15.1k
Facebook-React [25]	Traditional	232k	10.4k	15.5k
Angular [26]	Traditional	96.7k	27.9k	25.2k

Selection criteria included high repository activity, star count indicating popularity, contributor diversity, and clear domain classification.

B. Data Preparation

Issues, PRs, and comments from the repositories were extracted. For issues, the title, body, state (open/closed), timestamps (created, updated, closed), and associated labels were collected. For PRs, the title, body, state, timestamps, and merge details were gathered. Additionally, comments associated with both issues and PRs were extracted and used for sentiment analysis.

The data collected from the repositories underwent several data preprocessing stages to ensure its quality and suitability for analysis. These steps included data cleaning, data filtering, data formatting, data balancing, and data validation techniques.

Records with incomplete or missing critical information were either supplemented with placeholder text or discarded. Duplicate entries were identified and removed to ensure the uniqueness of each record.

Text processing techniques such as tokenization, stopword removal, and lemmatization were applied to the comment data. A custom `clean_text` function was used to standardize text data by converting to lowercase and removing unnecessary characters such as punctuation marks and numbers.

Several filters were applied to ensure the dataset remained focused and relevant. Records with fewer than five comments were removed to focus on meaningful discussions. Issues or PRs that had not been updated in the last year or were created before 2020 were removed to ensure current relevance. Only merged or closed PRs and issues were retained for further analysis. Extreme outliers with unusually high or low comment counts were removed to prevent skewing results.

Dates and times were standardized to ISO 8601 format. Additional features such as time-to-close for issues and PRs were created. Categorical variables were encoded numerically for machine learning models.

To ensure fair representation of both AI software and traditional repositories, the dataset was carefully balanced using a stratified approach. Datasets from each category were combined in proportional measures, maintaining equal representation of issues and PRs across both repository types.

Data validation steps included spot checks using functions like `sample()` and `head()`, statistical summaries of key attributes, and missing data analysis using the k-nearest neighbors (KNN) imputation for numerical values and mode imputation for categorical values.

The dataset was enhanced by integrating process mining data, structured into event logs containing key attributes such as case ID, activity, timestamp, and resource attributes. These logs were validated using process mining tools like Disco [27] and PM4Py [28] to ensure they met the necessary quality standards for analysis.

C. Data Modelling

The data model for this research integrates various data entities, each representing key elements of the GitHub repository workflow. Figure 1 illustrates the interconnections among users, repositories, issues, PRs, commits, and comments.

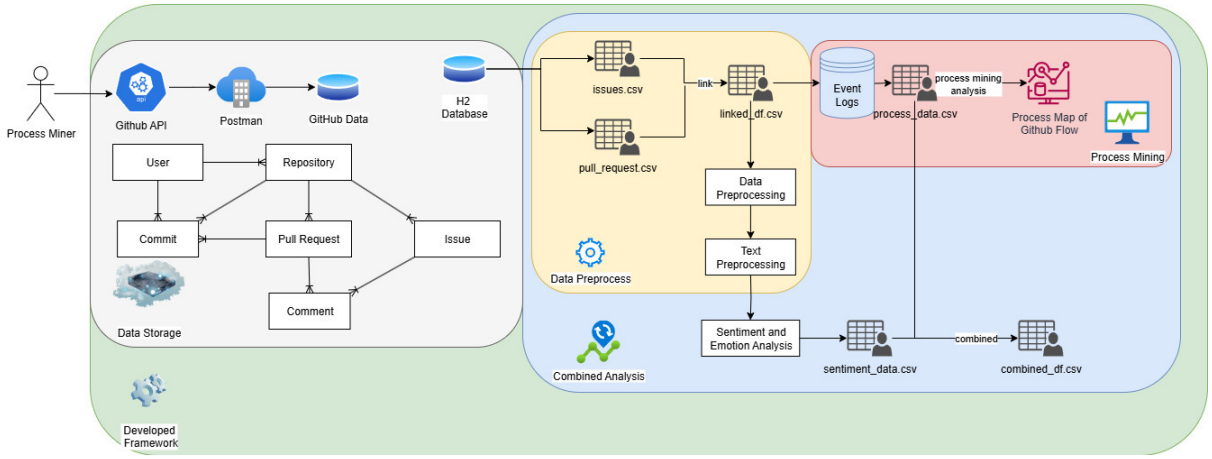


Fig. 1: Overview of the Developed Framework for GitHub Workflow Analysis

The core entities in the data model include users, repositories, issues, PRs, commits, and comments. A user can own or contribute to multiple repositories and can create or be assigned to various issues. Similarly, users can create or review multiple PRs, as well as produce numerous commits and comments.

Each repository in the model can contain multiple issues, PRs, and commits. Issues are typically associated with multiple comments, forming a one-to-many relationship. Similarly, PRs can encompass multiple commits and comments, with each commit associated with a specific PR.

Moreover, comments are crucial for communication during the development process. They are linked to either an issue or PR, serving as feedback or further discussion on a specific task. These interactions enable developers to refine code, resolve issues, and collaborate effectively.

The data is collected using the GitHub API, with retrieval performed via Postman. The collected data is stored in an H2 Database [29]. Subsequently, the raw data is transformed into two files: issues.csv and pullrequest.csv. These files are integrated into a third file, linkeddf.csv, which links issues to their corresponding PRs for analysis. This integration is crucial for understanding the relationship between issues and PRs and enables the analysis of their lifecycle within the repository.

The dataset is subjected to extensive data cleaning and text processing to prepare it for sentiment and emotion analysis. Additionally, the processed data is analyzed using process mining techniques to generate visual process maps, helping to trace the flow of actions such as issue or PR creation, updates, closures, and their relationship with sentiment.

D. Predictive Process Modeling

The structured dataset formed the basis for training various machine learning models aimed at analyzing the workflow in GitHub repositories. The predictive models developed in this study focused on two primary tasks: classifying issues and PRs, and predicting sentiment and emotional tone from text data. This section outlines the steps followed in the modeling process, which include feature engineering, model selection, and model training.

1) Feature Engineering: The first step in predictive modeling was feature engineering, where key attributes from the 'issues.csv' and 'pull_request.csv' files were extracted. The features considered for classification include text-based attributes such as the length of issue descriptions, the titles of PRs, and the presence of specific keywords that may indicate the nature of the task (e.g., "bug", "feature", and "documentation"). These features are crucial for distinguishing between different types of issues and PRs.

Additionally, sentiment-related features were generated from the text data. Sentiment scores (positive, negative, neutral) were extracted using sentiment analysis tools, which helped gauge the emotional tone of the text.

2) Predictive Model Selection and Training: A variety of machine learning models were considered to classify issues and PRs into categories such as "bug", "feature request", or "documentation update". Models such as Random Forest, Support Vector Machines (SVM), XGBoost, Naive Bayes, and Logistic Regression were selected for their proven effectiveness in text classification tasks.

Sentiment analysis was performed on comments within issues and PRs using the Valence Aware Dictionary and sEntiment Reasoner (VADER) [30] model in this study. VADER is well-suited for evaluating sentiment patterns, as it is specifically designed to detect the intensity of sentiments expressed in social media-like content. The analysis categorized comments as positive, neutral, or negative, with aggregated results providing insights into sentiment trends throughout the issue resolution and PR processes.

To complement sentiment analysis, emotion analysis was conducted using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [31] model, which had been fine-tuned for emotion classification. BERT was chosen due to its ability to capture subtle contextual nuances in text, making it particularly effective for analyzing the emotional tone of developer comments. The model identified several

emotions, including neutral, sadness, surprise, anger, joy, fear, and disgust.

VADER and BERT have limitations with domain-specific technical language and context length variations common in software development discussions.

These emotional insights were then correlated with key metrics such as issue resolution times and overall repository activities, aiming to explore the potential impact of developer emotions on workflow efficiency.

The final step in the modeling phase involved deploying the best-performing models in a real-time environment. These were integrated into an automated pipeline that classifies incoming issues and PRs, predicts their sentiment, and performs process mining analysis.

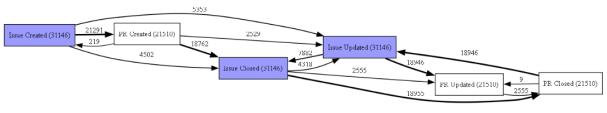
IV. Results and Discussion

A. Process Discovery of Software Repositories

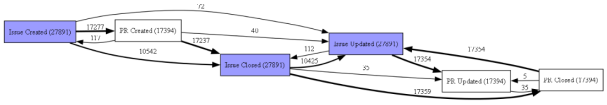
Event logs of repositories are used to discover the process map of the repositories using DFG Miner of the PM4Py library [28]. Figure 2 illustrates the process maps of PyTorch (AI software repository) and Angular (traditional software repository), showing common states including issue created, issue updated, PR created, issue closed, PR updated, and PR closed.

The flow of issues typically follows a consistent pattern: creation, updates, and eventual closure. However, certain issues are directly closed without further updates, while others may lead to the creation of corresponding PRs before they are closed. Similarly, PRs generally undergo creation, updates, and closure, although some are updated shortly before closure.

Notably, Angular exhibits a higher frequency of issues and PR interactions, indicating a more rapid and iterative development approach, which may be typical for traditional software development practices. In contrast, PyTorch demonstrates fewer PR updates and fewer instances where PRs are generated following issues, suggesting a more cautious and deliberate workflow, potentially attributed to the complexity and specificity of AI-related code. These findings emphasize key differences in workflow efficiency and developer behavior between AI software and traditional repositories.



(a) PyTorch Process Map



(b) Angular Process Map

Fig. 2: Process Maps for AI software and Traditional Repositories

B. Sentiment Analysis of the Repositories

Figure 3 depicts the distribution of sentiments across varying time-to-close durations for issues. AI software repositories exhibit a greater density of issues with neutral to positive sentiment and shorter resolution times. In contrast, traditional repositories exhibit a wider spread and with more cases of longer time-to-close, regardless of sentiment.

The data shows that issues with positive or neutral sentiment are generally resolved more faster in both types of repositories, possibly due to their less critical or complex nature. However, traditional repositories have a more distributed model, suggesting variability in the handling and resolution of issues, which may reflect different workflows or resource allocations.

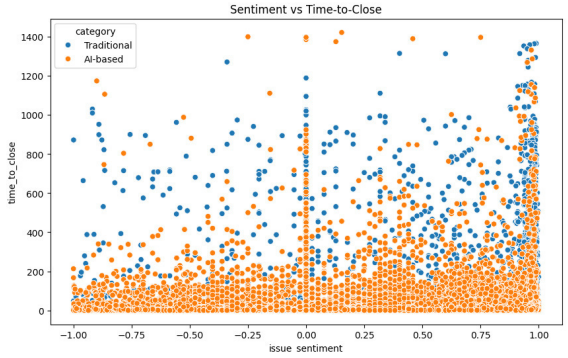


Fig. 3: Comparison of Issue Resolution Times

Figure 4 illustrates the distribution of sentiment scores. We see that AI software repositories have a significant peak around the neutral sentiment (0), with a lower positive sentiment peak, suggesting that issues in AI repositories often maintain a balanced or positive view, even with their complexity. In traditional repositories, there is a larger spread towards positive sentiments, potentially reflecting trust in established development practices and community support.

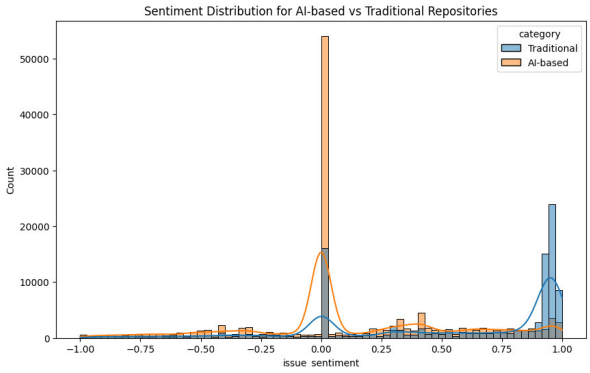


Fig. 4: Sentiment Polarity Distribution

C. Emotion Analysis of the Repositories

Figure 5 provides insight into the common emotional responses in AI software and traditional repositories. In particular, "sadness" and "fear" emotions are high in both categories.

However, AI software repositories show a significantly higher frequency of "sadness" compared to traditional repositories. Also, "anger" and "surprise" are observed more in AI software repositories, while traditional repositories show slightly higher levels of "fear".

This distribution suggests that problems in AI software repositories may trigger stronger negative emotions. This is probably due to the complexity and rapid evolution characteristic of AI projects. Contributors may experience frustration and even hesitation when working with advanced models and algorithms, which may increase emotional reactions. Traditional repositories, in contrast, seem to reflect a more balanced emotional distribution, possibly because they are dealing with well-established technologies and often face problems with familiar, predictable solutions.

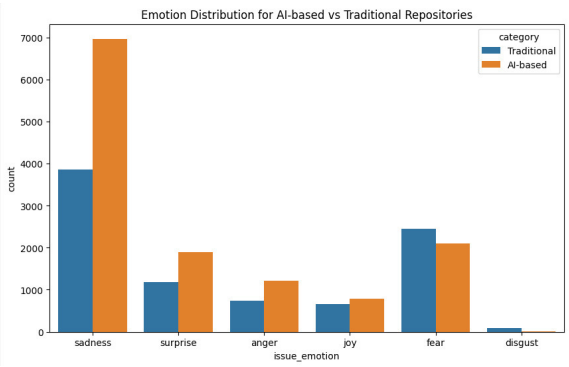


Fig. 5: Emotion Distribution for AI software vs. Traditional Repositories

D. Comparison of Issue Labels the Repositories

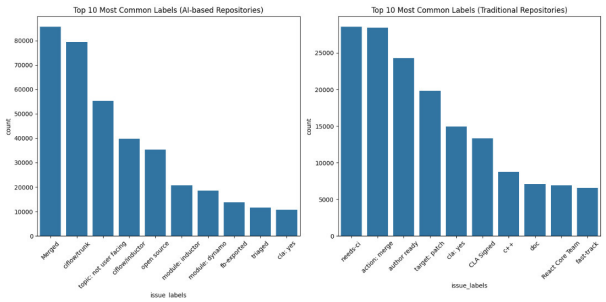
Figure 6a displays the top 10 most common issue labels in and Figure 6b also displays the top 10 labels with the highest average sentiment for both AI software repositories (left) and traditional repositories (right).

Figure 6 compares issue labels. AI software repositories frequently use "Merged," "ciflow/trunk," emphasizing automated testing crucial in ML projects. Traditional repositories show "Needs-ci," "action: merge," emphasizing formal CI processes.

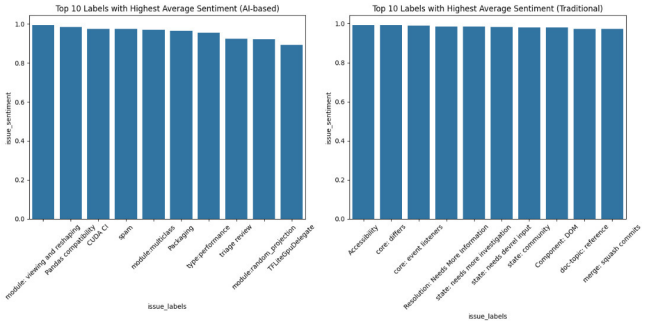
The comparison reveals that while both AI software and traditional repositories focus on merging contributions and continuous integration, AI software repositories have a greater emphasis on modularity and rapid integration of changes, while traditional repositories follow more structured, incremental workflows. These distinctions align with the experimental nature of AI development versus the stability and predictability required in traditional software engineering.

High-sentiment labels differ: AI software repositories focus on performance optimization ("CUDA CI," "type-performance"), while traditional repositories emphasize core functionality ("Accessibility," "core: event listeners").

The comparison reveals that while both AI software and traditional repositories foster positive sentiment toward issue resolution, their focus areas differ. AI software repositories emphasize modularity, performance optimization, and compatibility across frameworks, which are crucial in cutting-edge AI/ML development. In contrast, traditional repositories concentrate on core stability, accessibility, and detailed issue investigation, which aligns with the needs of more mature, stable software projects.



(a) Most Common Labels



(b) Highest Sentiment Labels

Fig. 6: Label Analysis in AI software and Traditional Repositories

E. Resolution Time Comparison

We compare the resolution time of issues and PRs for each repository type and list them in Table II and III. AI software repositories take longer to close issues (55.04 vs 43.63 days) but close PRs faster (36.68 vs 40.43 days). Both show similar median times (2-3 days), indicating most items are resolved quickly regardless of domain.

TABLE II: Time to Close Comparison for Issues

Category	Mean (days)	Median (days)
AI software	55.04	2
Traditional	43.63	2

TABLE III: Time to Close Comparison for Pull Requests

Category	Mean (days)	Median (days)
AI software	36.68	3
Traditional	40.43	3

V. Conclusion

This study provides a comparative analysis of AI software and traditional GitHub repositories, focusing on workflow efficiency and sentiment dynamics.

A comparison between AI software repositories, such as PyTorch, and traditional repositories, such as Angular, highlighted significant differences in development practices. The process maps indicated that workflows in AI software repositories are often more deliberate and cautious, with fewer PR updates, while traditional repositories tend to exhibit more iterative processes with frequent PR updates. These findings are instrumental for assessing workflow efficiency and serve as a foundation for exploring how sentiment and emotion analysis might impact productivity and developer behavior within these repositories.

These findings demonstrate how software domain characteristics influence development patterns, workflow efficiency, and contributor sentiment.

Despite these promising results, this study has several limitations. The dataset is limited to eight repositories due to GitHub REST API rate limits (5,000 requests/hour), potentially restricting the generalizability of the findings. Additionally, the analysis does not account for code coverage of AI code of the repositories and variations in team size, project complexity, or contributor demographics, which may influence repository performance.

Future research could address these limitations by exploring a broader range of repositories and incorporating additional contextual factors. Furthermore, integrating explainable AI methods to assess issue complexity dynamically could provide deeper insights into the decision-making processes of contributors and administrators.

VI. Acknowledgment

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

References

[1] João Caldeira, Fernando Brito e Abreu, Jorge Cardoso, Rachel Simões, Toacy Oliveira, and José Pereira dos Reis. Software development analytics in practice: A systematic literature review. *Archives of Computational Methods in Engineering*, 30(3):2041–2080, 2023.

[2] Tugba Gurgen Erdogan, Haluk Altunel, and Ayça Kolukısa Tarhan. A process model for ai-enabled software development: A synthesis from validation studies in white literature. *Journal of Software: Evolution and Process*, page e2743, 2024.

[3] Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

[4] Davide Spadini, Maurício Aniche, and Alberto Bacchelli. Pydriller: Python framework for mining software repositories. In *Proceedings of the 2018 26th ACM Joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 908–911, 2018.

[5] Bohan Liu, He Zhang, Weigang Ma, Hongyu Kuang, Yi Yang, Jinwei Xu, Shan Gao, and Jian Gao. Mining pull requests to detect process anomalies in open source software development. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13, 2024.

[6] Bjoern M Eskofier. Exploration of process mining opportunities in educational software engineering-the gitlab analyser. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pages 601–604, 2020.

[7] Martin Macak, Daniela Kruzelova, Stanislav Chren, and Barbora Buhnova. Using process mining for git log analysis of projects in a software development course. *Education and information technologies*, 26(5):5939–5969, 2021.

[8] Saimir Bala Thanh Nguyen and Jan Mendling. Multi-dimensional process analysis of software development projects. pages 179 – 186, 2024.

[9] Zhou Yang, Chenyu Wang, Jieke Shi, Thong Hoang, Pavneet Kochhar, Qinghua Lu, Zhenchang Xing, and David Lo. What do users ask in open-source ai repositories? an empirical study of github issues. In *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, pages 79–91. IEEE, 2023.

[10] Jeffrey Fairbanks, Akshharaa Tharigonda, and Nasir U Eisty. Analyzing the effects of ci/cd on open source repositories in github and gitlab. In *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 176–181. IEEE, 2023.

[11] Rifat Ara Proma and Paul Rosen. Visual analysis of github issues to gain insights. *arXiv preprint arXiv:2407.20900*, 2024.

[12] K Højelse, T Kilbak, J Røssum, E Jäpelt, Leonel Merino, and M Lungu. Git-truck: Hierarchy-oriented visualization of git repository evolution. In *2022 Working Conference on Software Visualization (VISSOFT)*, pages 131–140. IEEE, 2022.

[13] Wouter Poncin, Alexander Serebrenik, and Mark van den Brand. Process mining software repositories. In *2011 15th European Conference on Software Maintenance and Reengineering*, pages 5–14. IEEE, 2011.

[14] Francisco Jurado and Pablo Rodriguez. Sentiment analysis in monitoring software development processes: An exploratory case study on github's project issues. *Journal of Systems and Software*, 104:82–89, Jun. 2015.

[15] Bo Yang, Xinjie Wei, and Chao Liu. Sentiments analysis in github repositories: An empirical study. In *2016 Asia-Pacific Software Engineering Conference Workshops (APSECW)*, pages 67–74. IEEE, 2016.

[16] Eleni Guzman, David Azócar, and Yijun Li. Sentiment analysis of commit comments in github: An empirical study. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR)*, pages 352–355, 2014.

[17] William N. Robinson, Tianjie Deng, and Zirun Qi. Developer behavior and sentiment from data mining open source repositories. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 5386–5395. IEEE, 2016.

[18] Vinayak Sinha, Alina Lazar, and Bonita Sharif. Analyzing developer sentiment in commit logs. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 520–523. IEEE, 2016.

[19] TensorFlow Repository. An open source machine learning framework for everyone, 2025. Last accessed 1 January 2025.

[20] PyTorch. Tensors and dynamic neural networks in python with strong gpu acceleration, 2025. Last accessed 1 January 2025.

[21] Scikit-Learn. scikit-learn: machine learning in python, 2025. Last accessed 1 January 2025.

[22] Keras Team. Deep learning for humans, 2025. Last accessed 1 January 2025.

[23] Node.js. Node.js is an open-source, cross-platform javascript runtime environment, 2025. Last accessed 1 January 2025.

[24] Bootstrap. The most popular html, css, and javascript framework for developing responsive, mobile first projects on the web, 2025. Last accessed 1 January 2025.

[25] Facebook React. The library for web and native user interfaces, 2025. Last accessed 1 January 2025.

[26] Angular. Deliver web apps with confidence, 2025. Last accessed 1 January 2025.

[27] Christian W Günther and Anne Rozinat. Disco: Discover your processes. In *Demonstration Track of the 10th International Conference on Business Process Management, BPM Demos 2012*, pages 40–44. CEUR-WS. org, 2012.

[28] Alessandro Berti, Sebastiaan van Zelst, and Daniel Schuster. Pm4py: A process mining library for python. *Software Impacts*, 17:100556, 2023.

[29] Thomas Mueller. *H2 Database Engine*, 2023. Version 2.2.222, last accessed 27 May 2025.

[30] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225. AAAI, 2014.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. Fine-tuning methodology described within the original BERT paper.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 393–398
DOI: 10.15439/2025F2661
ISSN 2300-5963 ACSIS, Vol. 43

Challenges in Evaluating OSS Quality: Results from SLR on Quality Evaluation Tools

M. Aslı Taşgetiren
0009-0001-5369-4080
Hacettepe University
Üniversiteler Mah., 06800 Çankaya/Ankara/Türkiye
Email: aslitasgetiren@cs.hacettepe.edu.tr

Ayça Kolukısa-Tarhan
0000-0003-1466-9605
Hacettepe University
Üniversiteler Mah., 06800 Çankaya/Ankara/Türkiye
Email: atarhan@cs.hacettepe.edu.tr

***Abstract*—Open Source Software (OSS) quality evaluation is essential for ensuring the adoption and effective use of OSS projects across various domains. Most works on OSS quality evaluation have focused on the development of models or frameworks rather than providing practical implementations. These proposals can be challenging to use for inexperienced users, which highlights the need for user-friendly tools. This paper presents preliminary findings from a Systematic Literature Review (SLR) that investigates the characteristics, limitations, and gaps of current Quality Evaluation Tools (QETs) for OSS. Our analysis reveals the diversity of quality models underlying these tools and the absence of standardization, which impedes the comparability and reliability of evaluation results. This work also informs the next steps in improving OSS quality evaluation practices.**

I. Introduction

DEVELOPMENT and the use of open source software (OSS) has become widespread, with more than 100 million public projects available on GitHub alone. However, the abundance of options makes selecting the right OSS challenging. Unique characteristics of OSS, such as full availability of source code, version history, and community-driven and decentralized development, further complicate the selection process. These traits limit the effectiveness of standard software quality evaluation tools, which are typically designed for closed-source software and do not account for OSS-specific characteristics. As a result, tools specifically designed or adapted for OSS are needed to support more accurate, consistent, and relevant evaluations.

Specialized quality evaluation tools are also critical, as quality evaluation directly impacts OSS adoption, maintenance, and contribution decisions. Developers, organizations, and users rely on these evaluations to determine whether an OSS project is reliable, secure, and maintainable enough to integrate into their systems. Without appropriate tools, OSS quality evaluations have a risk of becoming inconsistent, overly subjective, and difficult to compare across projects.

To allow for ease in OSS selection and comparison, various quality models and frameworks have been proposed. A systematic literature review published in 2020 [1] identified 35 primary studies that proposed an OSS quality evaluation model or framework. Another published in 2022 [2] identified 36 primary studies with the same purpose. However, manually applying these models can be challenging for the average user, creating a barrier to their practical adoption. To facilitate the use of these quality models, an easy-to-use and preferably automated method for adopting these quality models is necessary. To this end, some quality evaluation tools have been proposed along with studies about quality models.

In [2] it was discovered that out of 36 primary studies, 16 of them (44%) were supported by a tool. Additionally, [1] revealed that among the 35 primary studies analyzed, 22 referenced the availability of toolsets.

Nevertheless, proposing tools is not enough; the tools must be published and maintained for long-term usage. Only two tools mentioned in the reviewed papers in [1] had active, up-to-date web pages. Only one tool remains accessible in 2025.

In addition to tool availability, aspects such as usability, clear evaluation criteria, and compatibility with evolving OSS practices are essential for real-world adoption. Tools that are difficult to install, bound to outdated standards, or lacking documentation further reduce their utility to practitioners and researchers alike. These challenges highlight the need for a structured investigation into the current state of QETs.

Previous study [2] investigated literature up to 2020 and focused on Quality Evaluation Models or Frameworks (QEMoF) for OSS. It provided valuable insights into OSS quality assessment approaches by investigating various aspects of QEMoF, including their goals, structure, and evaluation criteria. However, it only briefly examined the existence of supporting tools as a sub-question, and did not have a dedicated focus on Quality Evaluation Tools (QETs).

Given the strong connection between QEMoFs and QETs, we build upon this prior work by conducting a new Systematic Literature Review (SLR) that explicitly focuses on QETs for OSS. Our work expands on [2] by:

1) Introducing new research questions specifically targeting QET characteristics.
2) Extending the investigation coverage beyond 2020 to include more recent developments.
3) Modifying the search strategy to specifically identify studies that propose practical, user-accessible tools rather than only conceptual models.

This study was supported by Scientific and Technological Research Council of Turkey (TUBITAK) under the Grant Number 746360. The authors thank to TUBITAK for their supports.

To that end, in comparison to 13,146 studies initially retrieved in [2], our updated search retrieved 15,036 additional works. After applying inclusion and exclusion criteria, we selected 19 primary studies on QETs, which are published between 2006 and November 2024, for detailed analysis.

This paper presents initial findings from this SLR on QETs for OSS, highlighting key challenges, limitations, and gaps in existing tools. We discuss the lack of maintenance, accessibility, and standardization among the QETs and outline a vision for future research towards developing standardized, user-friendly tools to improve OSS quality evaluation practices.

II. Research Design

Goal of this research was to comprehensively examine and characterize existing quality evaluation tools for OSS by their content, structure and practical use. The research design followed the SLR methodology [3] and took advantage of PICOC (population, intervention, comparison, outcomes and Context) template [4].

Population: OSS quality evaluation

Intervention: OSS quality evaluation tools

Comparison: Characteristics of OSS quality evaluation tools

Outcomes: Basic characteristics and structure, degree of guidance provided, basic characteristics in evaluation, development method, and validation method of the proposals.

Context: Academia (scientific literature)

A. Research Questions

Given the strong connection between QEMoFs and QETs, some questions were taken from [2] and were adapted to be more in line with our focus on tools. Aim of this adaptation was to ensure that our study builds upon prior research while addressing the gap in practical tool implementations.

B. Search String

To make search results comparable with those of the SLR done for QEMoFs in 2020, the search string in [2] was taken as a basis. In order to focus the search on QETs in this SLR, the phrase *(tool OR framework OR application)* was added to the search string in [2] to obtain the final search string. This adjustment ensured that the search captured studies proposing implemented and potentially usable tools, rather than solely conceptual models. The final search string was:

> **(tool OR framework OR application)** AND
> ("quality") AND
> ("evaluation model" OR "assessment model" OR "measurement model" OR "evaluation framework" OR "assessment framework" OR "measurement framework") AND
> ("OSS" OR "FOSS" OR "FLOSS" OR "open source software")

This query was executed across seven major scientific databases: IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, Springer, Web of Science, and Google Scholar.

C. Inclusion and Exclusion Criteria

As a result of the updated search string, our SLR retrieved 15,036 additional studies that were published between 2020 and 2024, beyond those analyzed in [2]. While the study count is relatively high for a four year period, most of the results came from Google Scholar which is prone to result duplication.

1) Exclusion Criteria (EC): The following criteria were applied to remove irrelevant or duplicate studies:

EC1: Duplicate studies

EC2: Studies that were not published in English

EC3: Studies that are not accessible in full text

EC4: Secondary studies (e.g., SLR, meta-analyses)

EC5: Books and theses

EC6: Studies that discuss OSS quality evaluation without proposing, applying, or describing a specific QET

EC7: Studies on OSS quality evaluation that do not reference a known quality model or standard

EC6 aims to exclude purely conceptual or theoretical discussions that lack reference to any practical tool or implementation. EC7 was introduced to distinguish dedicated OSS quality evaluation tools from generic tools that only incidentally evaluate software quality (e.g., security scanners, static analysis tools).

2) Inclusion Criteria (IC): A study was included if:

IC1: It explicitly focused on OSS quality evaluation.

IC2: It proposed a tool specifically designed for evaluating OSS quality.

D. Data Extraction

After screening, 19 primary studies on QETs were selected, these primary studies are listed in a supplementary material (see Data Availability section). To ensure a systematic review process, data extraction was guided by the predefined research questions (RQs) which are introduced and answered in Section III. A structured spreadsheet was used to record the extracted data, with each sub-question represented as a column. For each study, relevant information was manually extracted from the full text.

The data extraction was performed by the first author, and ambiguous cases were reviewed and discussed with the second author to promote consistency and reduce bias. The RQs were designed to capture key characteristics of each QET, including:

- Basic characteristics (RQ-1)
- Structural attributes (RQ-2)
- Guidance and usability (RQ-3)
- Evaluation methods (RQ-4)
- Development details (RQ-5)
- Current accessibility and maintenance status (RQ-6)

This structured approach ensured that findings are comparable across different QETs and can be synthesized effectively.

III. Results

This section presents the preliminary findings of our study. The results are structured based on the defined RQs and cover aspects that were specified in the previous section.

The following subsections systematically address each RQ, summarizing the key insights extracted from the reviewed studies. An overview of the results is provided in table format, along with full references to primary studies, on Zenodo (see Data Availability section).

The findings highlight recurring issues such as lack of standardization, limited usability, and sustainability concerns; reinforcing the need for more consistent, well-maintained, and user-friendly QETs.

A. RQ-1: What are the basic characteristics of the QET proposed for OSS?

Data were extracted from each primary study to determine the main purpose of the QET and assess the use of AI techniques in its evaluation logic. Understanding these is essential for distinguishing between broadly applicable tools and those tailored to specific evaluation goals or contexts.

1) RQ-1.1: What is the main purpose of the QET in the study?: Each QET was categorized based on its primary evaluation focus. Tools were labeled as either:

- G: General-purpose i.e., designed to evaluate overall OSS quality, or
- S: Domain-specific i.e., targeting one or more focused quality attributes such as maintainability or security.

The findings indicate that out of the 19 tools analyzed, 13 (68%) of the QETs were designed for general-purpose evaluation, while 6 (32%) focused on a specific quality aspect. This distribution suggests a trend toward designing general-purpose evaluation approaches, although domain-specific tools remain relevant for targeted quality evaluations.

2) RQ-1.2: Does the QET utilize AI?: For AI usage, we checked if the tool explicitly described the integration of AI-based techniques (e.g., machine learning, natural language processing (NLP)) in its implementation. Identifying the potential use of AI integration with this sub-question helps to asses how modern and scalable these tools are, and highlights opportunities for future research.

Out of 19 QETs, only one utilized AI methods. The work used NLP to extract data from community channels. While there are other works that talked about incorporating AI to their tools in the future, they have not released a version that contains it yet. This suggests that although the application of AI in OSS quality evaluation is recognized as promising, its actual implementation in the published QETs remains rare.

B. RQ-2: How is the QET for OSS structured?

1) RQ-2.1: Which quality model is the QET based on?: We analyzed if the QETs referred to formal software quality models—such as ISO/IEC 25000, ISO/IEC 9126, or Dromey—or used custom or OSS-specific models such as OpenBRR, QSOS, or OSMM.

This sub-question was included to assess if the QETs are grounded in standardized definitions of software quality, which supports consistency and comparability across different tools and evaluation contexts. When tools rely on outdated or informal models, their results may not align with current quality standards or be usable beyond their original context. Thus, this analysis helps determine how closely existing QETs align with widely accepted software quality frameworks and whether there is a need for standardization or modernization.

Results indicate that 47% of the QETs used ISO/IEC 9126 [5] as at least one of their base models; however, ISO/IEC 9126 has been replaced with ISO/IEC 25010 [6] and new models should take this standard as the basis. Only 16% of the tools were based on the newer ISO/IEC 25010 standard. Several tools referenced OSS-specific quality models like QSOS, OpenBRR, and OSMM, which were designed to address OSS evaluation challenges but may no longer align with current international standards. This suggests a lag in the adoption of up-to-date quality models within the QETs.

2) RQ-2.2: From what aspects can OSS be evaluated by the QET?: The QETs were classified based on the quality aspects they evaluated, including product quality, quality-in-use, community-related factors, a combination of these, or a single specific attribute.

This sub-question was included to evaluate how comprehensively each tool assesses OSS quality. OSS differs from proprietary software not only in its technical traits, but also in how it is developed, used, and maintained by its community. As such, a meaningful OSS quality evaluation should ideally include technical, user-facing, and community aspects. By analyzing which aspects each QET covers, we can evaluate if a tool is too narrowly focused or provides a more comprehensive assessment aligned with the unique nature of OSS.

The analysis revealed that 84% of the QETs addressed product quality (e.g., maintainability, reliability), 58% considered quality-in-use (e.g., usability, user satisfaction) and 74% incorporates community-related metrics (e.g., maintenance capacity, sustainability). Seven tools specifically used a mix of each of these quality aspects and two tools focused only on a single attribute such as security. These results indicate that while most tools emphasize code-level characteristics, many also acknowledge the importance of user experience and community health, which are both critical to the success and adoption of OSS projects.

C. RQ-3: What degree of guidance is provided for the evaluation of OSS by the QET?

This question examined how clearly the QETs explain their evaluation procedures and whether they demonstrated the tool's application through practical examples. Clear guidance is essential for consistent use and broader adoption; without it, even technically sound tools may remain unused.

1) RQ-3.1: Is the evaluation procedure of the QET adequately described in the study?: To assess the transparency of the evaluation process, each study was analyzed to determine whether it provided an adequate (yes), partial, or no description of how the QET could be used to assess OSS. A well-documented evaluation process increases the likelihood that a tool can be adopted by practitioners or extended by other researchers.

Less than half (47%) of the QETs were determined to include a detailed evaluation procedure, while 32% provided partial descriptions, and 21% lacked a guideline. These findings indicate that the majority of tools do not offer sufficient instructions, which can make them harder to apply in practice and reduce the consistency and reproducibility of their results.

2) RQ-3.2: Is a demonstration of the evaluation using the QET provided in the study?: Demonstration of the tool in practice is critical for illustrating how the tool functions, validating its usefulness, and helping users understand expected outcomes.

The analysis revealed that 47% of the studies included some practical demonstration. The remaining 53% lacked detailed examples of how the tool was applied, which limits users' ability to assess its effectiveness and real-world applicability.

D. RQ-4: What are the basic characteristics of the QET for evaluating OSS?

1) RQ-4.1: Is the license type of the OSS product used as an evaluation criterion?: Licensing type is a very important characteristic in OSS, as it can potentially restrict or influence the licensing of other software that incorporates it. Seven (37%) of the QETs explicitly incorporated license type as a factor in quality evaluation, while the rest did not consider licensing constraints in the assessment process.

2) RQ-4.2: Does the QET support subjective or objective evaluation?: The studies were analyzed to determine whether the QETs relied on subjective assessments (e.g., taking user opinions into consideration) or objective assessments (e.g., metric-based analysis). Objective assessments tend to support consistency and automation, while subjective ones can capture nuanced user perspectives that are difficult to quantify.

The results indicate that 53% of the QETs combined subjective and objective approaches, while 16% supported entirely objective evaluation and 32% relied on subjective assessments. These results suggest that while many tools aim for a balanced approach, a significant portion still depends heavily on subjective input. Subjective input is risky as it can limit standardization or automation in large-scale OSS evaluations; however, some tools knowingly allow subjective input to include diverse user perspectives.

3) RQ-4.3: Does the QET support qualitative or quantitative evaluation?: This sub-question assessed the QET's reliance on quantitative methods (e.g., numeric metrics, automated analysis) or incorporation of qualitative insights (e.g., user interpretations, expert reviews). The evaluation approach influences how scalable, replicable, and interpretable the tool is in practice.

The analysis reveals that 89% of the QETs primarily relied on quantitative evaluation, using methods such as automated code analysis and structured data collection. Remaining two tools combined both approaches for a more comprehensive evaluation. The predominance of quantitative evaluation aligns with the goal of automation but may limit the capture of context-dependent or user-centered insights.

4) RQ-4.4: What quality attributes are used for evaluation by the QET?: The inclusion of different attributes reflects how the tool defines "quality" and which aspects of OSS it prioritizes. The most frequently addressed quality attributes were maintainability (68% of the tools), reliability (58%), efficiency (47%), and security (42%).

These results suggest that most QETs focused on code-centric attributes commonly emphasized in traditional software quality models. This focus was likely due to relative ease of automatically extracting code-related metrics. However, for tools specialized in evaluating OSS, it is important to consider OSS-specific attributes. RQ-4.5 will show that while there are tools that consider community activity metrics, which are unique to OSS, code related metrics still play a more prominent role.

5) RQ-4.5: What software metrics are used for evaluation by the QET?: To further understand how the QETs measure and apply quality criteria, we investigated the specific software metrics used in evaluation. These metrics serve as the foundation for generating evidence-based quality assessments. Across the reviewed studies, commonly used metrics included:

- Code quality indicators (e.g., complexity, code smells),
- Community activity metrics (e.g., issue tracking, pull request volume),
- Bug tracking statistics,
- Support responsiveness and service quality.

Details of these distributions can be seen more clearly in the summary table (see Data Availability section).

6) RQ-4.6: What aggregation methods are used for evaluation by the QET?: Metric values gathered using various techniques tend to be heterogeneous in nature, making combining and comparing them difficult. To combine different metric values into a single score, aggregation methods are used. Weighted arithmetic mean was the most used aggregation method among the studies (63%), with some using other techniques such as overall sum or other mathematical equations.

The widespread use of weighted mean indicates a preference for simple and interpretable methods of combining quality metrics. At the same time, the presence of alternative aggregation methods, such as summation or custom formulas, reflects a lack of standardization in how overall quality is calculated. This variation can reduce the transparency and comparability of results across different QETs.

7) RQ-4.7: Does the QET support evaluation at a single point in time or throughout the evolution of OSS?: Evaluating quality over the course of a project's development is particularly relevant for OSS, where code and community activity evolve continuously. Tools that support analysis throughout evolution can help detect trends, regressions, or improvements, and are valuable for maintainers making long-term decisions.

The results indicate that 79% of the QETs performed single-point evaluations, while 3 tools (16%) allowed tracking quality over time, enabling historical analysis of OSS quality.

A study made on OSS project size growth patterns [7] observed that many OSS repositories reach approximately 75% of their final code size within the first 25% of commits, which

might suggest that a snapshot taken after this period could capture much of the system's core. However, there is no guarantee that the snapshot will be taken at the right time, potentially causing significant early changes to be missed. Additionally, one-time evaluations risk overlooking later community-driven refinements or regressions.

8) RQ-4.8: How are data collected for OSS evaluation in the QET?: Data collection methods were categorized as manual (e.g., form entries, expert-provided ratings etc.) and automated (e.g., metric extraction from repositories, codes, or APIs). The method of data collection directly affects a tool's usability, scalability, and potential for continuous evaluation. Automated tools typically require less user effort and support more consistent and repeatable assessments. However, not all the data can be retrieved automatically, which might limit the data to be taken into consideration in fully automated tools.

The findings showed that 47% of the QETs relied on manual input, 47% used fully automated data retrieval, and one did not specify it. This even split suggests that while many tools aim to automate evaluation, a significant number still depends on user-provided data; which can limit scalability and introduce subjectivity, but also allow for the inclusion of qualitative or context-specific information that automated methods may not be able to access.

9) RQ-4.9: What is the required skill level of users who evaluate OSS using the QET?: Usability and accessibility are critical for encouraging widespread adoption of the tools, especially in OSS communities, where users may include non-experts.

The analysis showed that 58% of the QETs required a medium to high level of expertise, such as familiarity with software quality models, metric interpretation, or manual configuration. This suggests that many tools are geared toward expert users and may pose a barrier for broader adoption, particularly among smaller or less formal OSS projects.

It should be noted that in cases where the required skill level was not explicitly stated, it was inferred based on the tool's interface, configuration requirements, technical documentation and results from RQ-3.

10) RQ-4.10: How are the results of the evaluation provided to the user in the QET?: The format of the results significantly influences the clarity, practical usefulness, and availability of quality assessments—especially for non-expert users or decision-makers.

The analysis showed that the QETs presented their results using the formats listed below. Some used more than one format to show their results; these were counted in the percentages:

- Indexed values in range (e.g., 0-1, 0-100) in 53%,
- Ordinal scales (e.g., ranked scores) in 42%,
- Nominal scales (e.g., "good," "fair," "poor") in 11%.

These findings indicate a tendency toward numeric outputs, which are well-suited for integration into automated systems or dashboards but may reduce usefulness for stakeholders needing more descriptive evaluations.

E. RQ-5: What programming tools and frameworks were used while developing the QET?

The technologies used to develop the QETs were analyzed, including programming languages, frameworks, and external libraries. This information is important for assessing the maintainability, accessibility, and extensibility of the tools. It can also provide guidance for future QET developers.

Most papers did not specify the technologies used. To address this, the source code of the tools—when available—was examined to manually extract the necessary information.

Most commonly used technologies were found to be Python, Java and JavaScript. However, due to limited reporting and missing source code links, a complete analysis could not be performed for all the QETs.

F. RQ-6: What is the current status of the QETs for OSS?

Evaluating the current state (i.e., maintained, discontinued, or still in progress) of each tool is essential to understand its sustainability, potential for long-term use, and relevance to practitioners.

A five-year threshold was used to assess maintenance status, i.e., the tools that had no updates within the last five years were classified as no longer actively maintained. According to this criterion, only six of the tools (32%) were found to be actively maintained. Out of the six active tools, only one was published before 2019. Additionally, out of these, only four tools have had updates within the last five years.

For tools that were discontinued, an archived version was searched on platforms such as Wayback Machine and GitHub. This process aimed to determine whether older versions of the tools were still accessible for reference, analysis or use. The results indicate that 63% of the discontinued tools had an archived version available, while an archived version could not be found for the remaining 37%.

These results highlight a significant sustainability issue: Many QETs are not maintained over time, and in several cases, their disappearance makes it difficult for others to study, reuse, or build on them. This poses a barrier to both reproducibility and practical adoption of the QETs for OSS evaluation.

IV. Discussion

This SLR on Quality Evaluation Tools for OSS reveals several practical and methodological challenges that limit the tools' effectiveness, adoption, and long-term utility.

A key issue is the short lifespan and limited maintenance of many QETs; several tools are either no longer accessible or have not been updated in recent years. This result might be influenced by our focus on academic literature, where prototypes might not be sustained beyond a single study. However, to support reliable and replicable evaluation in real-world OSS contexts, QETs must remain accessible over time. Even if they are not maintained, at a minimum, tools should be archived and publicly available to enable reproducibility and research. Ideally, QETs should also be hosted in open-source repositories, like the OSS projects they assess. Without this,

both practitioners and researchers face significant barriers in applying existing tools or advancing the field.

A second major concern is the lack of standardization across the tools. The QETs are based on a wide range of quality models, ranging from outdated standards (e.g., ISO/IEC 9126) to OSS-specific frameworks such as QSOS or OpenBRR. This diversity, while reflective of differing needs, leads to inconsistent and non-comparable evaluation results. The absence of a shared framework or terminology makes it difficult to compare evaluations across tools or projects, limiting the potential for creating universal benchmarks or guidelines. This puts users in a position where they might need to evaluate the QETs before they can use the QET to evaluate OSS quality.

Finally, usability remains an issue. While some tools were aimed to be user-friendly, many require medium to high levels of technical expertise to use, limiting accessibility for non-expert users. This usability gap limits the widespread adoption of the QETs. Future tools should prioritize intuitive interfaces, quality model templates, clear documentation, and automated data collection to lower the barrier for inexperienced users.

Taken together, these findings highlight the need for more sustainable, standardized, and accessible Quality Evaluation Tools for OSS. Addressing these issues would not only improve the reliability and comparability of OSS evaluations, but would also promote widespread adoption and continuous quality improvement in OSS ecosystems.

V. Threats to Validity

This study, like any other SLR, is subject to several potential threats to validity. One potential threat is the completeness of our search strategy. While we carefully designed our search string to capture relevant QETs for OSS, some studies might have been missed due to variations in terminology or indexing issues in databases. Additionally, the digital libraries we used may not cover all relevant papers, particularly those from lesser-known conferences or non-indexed journals.

The scope of the dataset also presents limitations. Our findings are based on a finite set of primary studies, which may not fully represent all existing QETs for OSS. Some tools may exist only as industry applications without academic documentation, leading to a possible bias toward research-focused tools rather than those used in practice. Additionally; EC7, as stated in Section II.C, further limits the scope. Per EC7, this review focuses on QETs that reference known quality models or standards. That is, tools that evaluate OSS quality using non-standard methods (including some that might be widely used in practice) may not be represented. This choice was made to prioritize comparability and structure, but limits generalization of the results to all tools.

It should also be noted that some aspects of the analysis involved interpretive judgment, especially when tool characteristics (e.g., required skill level in RQ-4.9 or maintenance status in RQ-6) were not explicitly stated in the studies. In such cases, we inferred answers based on indicators like documentation detail, user interface design, or repository activity. For example, the five-year threshold in RQ-6 is an arbitrary cutoff and may not fully reflect a tool's usability, especially if it has remained functional without frequent updates.

Lastly, the study does not assess the real-world effectiveness of the identified QETs through hands-on user testing. While we analyzed their documentation and reported functionality, actual usability, reliability, adoption, and performance of the tools remain unverified. Future studies could address this by running case studies or user surveys to validate our findings.

VI. Conclusion

This study provides preliminary insights into the states of Quality Evaluation Tools for Open Source Software, identifying key challenges in tool sustainability, standardization, and usability. While OSS quality evaluation models are well-studied, practical implementation is still an issue due to outdated tools, inconsistencies, and complexity.

Moving forward, future research should focus on improving the longevity and accessibility of the QETs, ensuring that tools remain accessible even if they are not actively maintained. Additionally, establishing standardized frameworks and meta-models could help make evaluation results more comparable across different tools. Finally, incorporating automation and AI-driven approaches, which are not yet widely implemented in the QETs for OSS, has the potential to improve scalability and reduce the manual effort required for quality assessment.

Addressing these gaps will help create more sustainable, standardized, and user-friendly QETs, ultimately contributing to more effective evaluation and selection of OSS products, greater trust in OSS systems, and broader adoption of quality assessment practices.

Data Availability

The complete summary table of RQ results and the full list of analyzed primary studies are openly available on Zenodo at https://doi.org/10.5281/zenodo.15869799

References

[1] V. Lenarduzzi, D. Taibi, D. Tosi, L. Lavazza, and S. Morasca, "Open Source Software Evaluation, Selection, and Adoption: a Systematic Literature Review," in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Aug. 2020. doi: 10.1109/SEAA51224.2020.00076 pp. 437–444.

[2] N. Yılmaz and A. Kolukısa Tarhan, "Quality evaluation models or frameworks for open source software: A systematic literature review," *Journal of Software: Evolution and Process*, vol. 34, no. 6, p. e2458, Jun. 2022. doi: 10.1002/smr.2458 Publisher: John Wiley & Sons, Ltd.

[3] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," Technical report, ver. 2.3 ebse technical report. ebse, Tech. Rep., 2007.

[4] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén *et al.*, *Experimentation in software engineering*. Springer, 2012, vol. 236.

[5] "ISO/IEC 9126-1:2001." [Online]. Available: https://www.iso.org/standard/22749.html

[6] "ISO 25010." [Online]. Available: https://iso25000.com/index.php/en/iso-25000-standards/iso-25010

[7] K. Szymański and M. Ochodek, "On the applicability of the pareto principle to source-code growth in open source projects," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F5221 pp. 781–789.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 399–404
DOI: 10.15439/2025F3412
ISSN 2300-5963 ACSIS, Vol. 43

Anomaly Detection for Unmanned Surface Vehicles Based on a Multi-Modal Bayesian Generative Model

Micheale H. Tekulu*, Ali Krayani*, Pamela Zontone*, Lucio Marcenaro*,
Francesco Caprile‡, Antonino Masaracchia†, and Carlo Regazzoni*
*DITEN, University of Genoa, Italy
†School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
‡Leonardo Sistemi Integrati, Italy

***Abstract*—In this paper, we propose a novel method for abnormality detection in Unmanned Surface Vehicles (USVs) based on a Multi-Modal Bayesian generative model to enhance safety and monitoring. During the training phase, we use a Null Force Filter and an unsupervised clustering algorithm on multimodal data collected from Global Positioning System (GPS) and motor current sensors. In the testing phase, we use a Coupled Modified Markov Jump Particle Filter (CM-MJPF) to infer the GPS position and motor current of the USV, as well as to detect abnormalities in both modalities. Due to the coupled methodology, the system is able to learn the statistical similarity between the evolving GPS and motor current data. As a result, the causality of defects is inherently captured within the dynamical inference, making the proposed approach explainable.**

***Index Terms*—Bayesian generative model, anomaly detection, unmanned surface vehicle, unsupervised clustering, explainable AI.**

I. Introduction

IN Intelligent Transportation Systems (ITS), the goal of improving transport safety has led to groundbreaking developments such as autonomous surface vehicles. Since unmanned surface vehicles are autonomous, they are prone to mechanical or electrical faults [1]. If such faults are not detected accurately and in a timely manner, they may not only cause malfunction of the agent but also cause critical hazard. Therefore, to minimize potential risks and ensure robust functionality, the agent might relay on its exteroceptive and proprioceptive sensors, like motor current and GPS, to monitor its internal states while interacting with the external environment.

An anomaly is an unexpected event or behavior that could indicate a problem or fault (deviation from normality) [2], [3]. Recently, as part of machine learning, Bayesian Networks have proven well suited for anomaly detection because they support both discrete and continuous variables, model uncertainty, and allow for representation of time series data [4]. Sensors mounted on the agent help it sense and perceive the environment, and the signal measured by the sensors can reflect the agent condition. Relaying only on one sensor for the detection of operational failures can lead to inaccuracies due to inadequate information. Similarly, considering multiple sensors independently also provides limited information. Therefore, it is important to have a comprehensive system that integrates multiple interacting sensors [5].

In this work, we propose an interactive Multi-Modal Bayesian generative model based on time series data, which can detect USV operational faults. Unlike studies in literature, our approach utilizes the interaction of multiple sensors using the Bayesian learning mechanism. Additionally, our method is able to differentiate the main source of the defect by exploiting the causal behavior of Bayesian networks. The main contributions of this paper are the following.

- A novel data-driven Bayesian generative model for USVs is proposed.
- The proposed method uses an interactive multi-modal approach, where we combine both GPS and motor current for learning the model.
- The proposed framework can trace the cause of the abnormality in an explainable way.
- The proposed method can be generalized to any kind of multi-modal sensors.

The remainder of the paper is organized as follows: Section 2 presents a review of related work on probabilistic inference approaches for abnormality detection in autonomous vehicles. Section 3 describes the proposed method in detail. After this, the simulation and experimental results obtained from the proposed approach, along with a discussion, are presented in Section 4. Finally, Section 5 concludes the paper and outlines directions for future work.

II. Related Work

Model-based fault diagnosis uses physics-based or mathematical models to detect and identify system faults. These models describe the expected behavior of vehicle components under normal conditions, allowing for comparison with real-time sensor data. In [6], the authors propose a steering wheel fault diagnosis method for autonomous vehicles, using a model-based residual generator with an SVM (Support Vector Machine) classifier. Similarly, in [7], they use Position Estimation Models, along with an Extended Kalman Filter (EKF), to detect faults by computing residuals (Mahalanobis distance) between estimated positions from sensor pairs. Abci

Topical area: Advanced Artificial Intelligence in Applications

B. et al. [8] also formulate a model-based fault detection and isolation (FDI) system for autonomous mobile robots that can detect and isolate sensor and actuator faults using an informational framework. Although model-based fault diagnosis applications are useful, they have some drawbacks, such as inaccuracies in modeling complex scenarios.

Due to the improvement of machine learning and computational techniques, there has been growing interest in data-driven fault diagnosis. This paper, therefore, takes advantage of data-based abnormality detection methods focused on autonomous vehicles. While only a few studies in the literature directly address the operational fault diagnosis for USVs, as examined in this paper, several related works share similar characteristics and offer relevant insights. In [10], the authors use a data-driven Neural network to detect and isolate faults, based on the differences between the measured and estimated outputs (residuals). Furthermore, in [9], Gaussian process regression to generate the residuals is used.

The methods mentioned above focus on a single data source. However, as systems become more complex, relying on data from just one source is not sufficient for effective monitoring. To address this limitation, fault diagnosis systems now employ multi-sensor technology to capture a wider range of characteristic data. In [13], the authors propose convolutional neural network (CNN) architectures for the fault diagnosis of an induction motor, using different combinations of multi-sensory signals such as vibration, current, voltage, and speed. Similarly, Ö. Gültekin et al. [5] utilize time–frequency representations of time series data obtained from both vibration and sound sensors, and use a CNN model for the classification process to diagnose the operational faults of autonomous transfer vehicles. Shaoxuan Xia et al. [14] applied a hierarchical attention-based multi-source data fusion method for fault diagnosis in Autonomous Underwater Vehicles (AUVs), addressing challenges like multi-source data heterogeneity and strong coupling. They used a Bidirectional Long Short-Term Memory (BiLSTM) network for feature extraction and a Multilayer perceptron (MLP) for fault detection. Although multisensory fusion provides more information compared to using a single type of sensor, previous studies in the literature often consider each sensor independently. Moreover, they did not explain the cause of a defect. Unlike most of the papers available in the literature, our method is an interactive and multi-modal approach, and is able to explain the cause of a defect.

III. Methodology

A. Experimental Setup

We use a telemetry system to collect the necessary data for our experiment, as shown in fig. 1. Data from the inverter are gathered via the Controller Area Network bus (CANbus) communication protocol. This is achieved by sending a request to the inverter for a specific parameter reading, after which the inverter responds with the corresponding measurement. Moreover, the GPS module communicate via universal asynchronous receiver / transmitter (UART). Data provided by

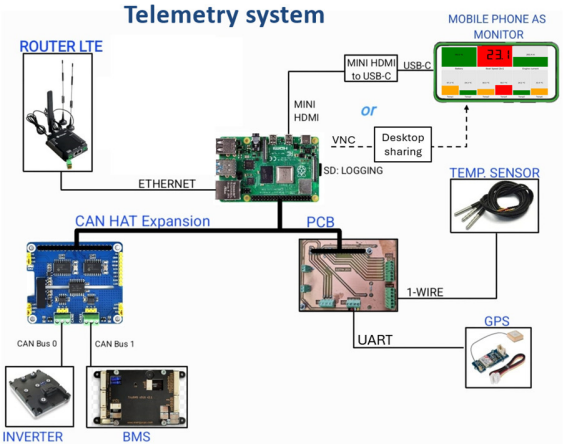


Fig. 1: Experimental setup-Telemetry system

GPS are encoded according to the NMEA (National Marine Electronics Association) standard.

B. Preliminaries

- Generative models: There are two types of models in machine learning, i.e., discriminative and generative. Discriminative models focus on learning the probability of a specific class label given an input, which is the approach commonly used in classification tasks. Conversely, generative models learn the underlying data distribution, enabling them to produce new data samples that follow the same pattern.
- Bayesian Networks (BNs): They are a type of generative model that allow us to model causal relationship between random variables.
- Dynamic Bayesian networks (DBNs): They are a particular type of BN which can describe dynamic processes that evolve over time.
  The main advantage of a DBN is its capability to formulate causal relationships between semantic information (high level) and sensory data (low level) in a hierarchical level. DBNs with hierarchical structure are called Hierarchical Dynamic Bayesian networks (HDBNs). Fig. 2 shows a DBN representation for three hierarchical variable levels. The nodes represent the random variables for a continuous level state (X_t and Z_t) or a discrete level state (S_t). The links represent conditional dependencies between nodes.

C. Proposed DBN for Learning Interaction

1) Training phase: Initially, the USV perceives its surroundings under the static assumption that the environmental state does not change. Therefore, the USV predicts its position (GPS) and internal state (motor current) using a null force filter with the following model:

$$X_t^n = X_{t-1}^n + W_t^n, \tag{1}$$

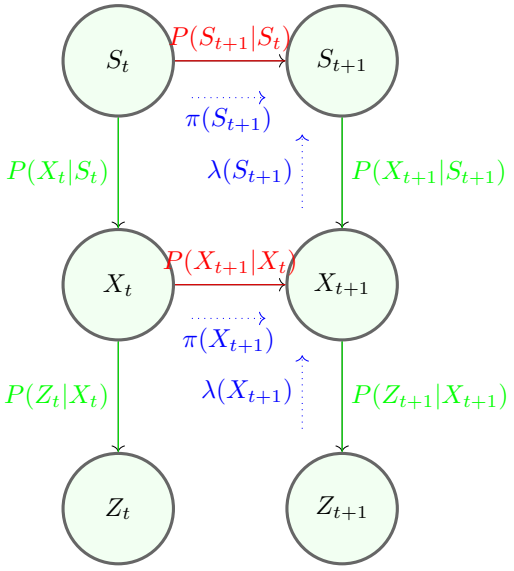


Fig. 2: Three level DBN. The links highlighted in red are inter-slice links, and the ones in green are intra-slice links at different times.

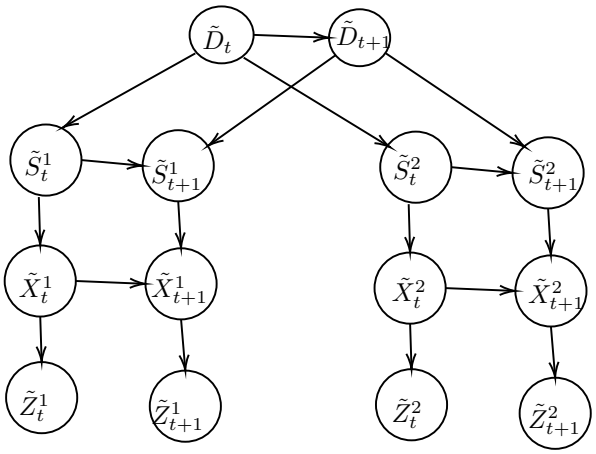


Fig. 3: Learning interactions among GDBN Models for both modalities (1 for motor current and 2 is for GPS), where $\tilde{D}$ is the Interaction Discrete level, $\tilde{S}$ is the Superstate Discrete level, $\tilde{X}$ is the continuous level and $\tilde{Z}$ is the measurement level

where W^n_t represents the noise accounting for the measurement uncertainty of the n-th sensor (with n=1 for motor current and n=2 for GPS) at time t.
In order to explore the dynamic rule through time in the environment, a new rule should be created by exploiting the generalized error, which is the difference between prediction and observation. Generalized errors $\tilde{X}_t$ are expressed as:

$$\tilde{X}_t = [X_t, \dot{X}_t]^T \tag{2}$$

where X_t is calculated using equation (1), and $\dot{X}_t$ can be considered as the innovation of the null force filter calculated

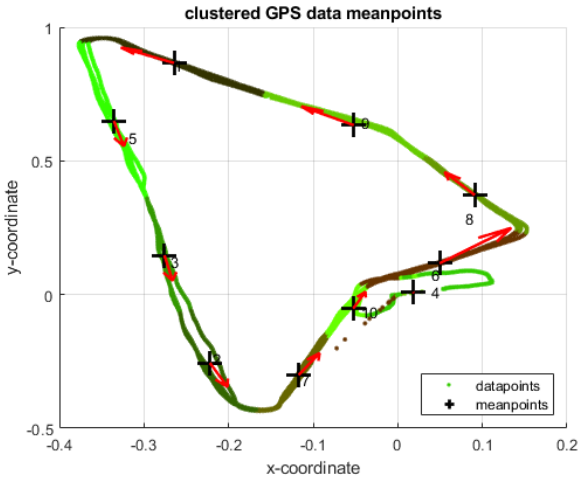


Fig. 4: Clustered GPS data

using this equation:

$$\dot{X}_t = H^{-1} Z_t - X_t \tag{3}$$

where Z_t is the measurement acquired from the sensors at time t and H is matrix that maps the observation to the state X.

Consequently, the Generalized errors collected from previous experience are used as input to an unsupervised clustering algorithm (i.e., Growing Neural Gas (GNG)), which encodes them to discrete variables or neurons ($\tilde{S}^n$) called super states (letters):

$$\tilde{S}^n = \tilde{S}^n_1, \tilde{S}^n_2, ..., \tilde{S}^n_M \tag{4}$$

where M is the total number of superstates.

Each generalized superstate $\tilde{S}^n_t (\tilde{S}^n_t \in \tilde{S}^n)$ is associated with its corresponding statistical properties mean $(M^{\tilde{S}^n_t})$ and covariance matrix$(Q^{\tilde{S}^n_t})$, as shown in figure 4.

Following the clustering of each sensor's generalized errors, a word or meta cluster ($\tilde{D}_t$) is formed, containing the superstate clusters (letters) of the two sensors, motor current and GPS ($\tilde{S}_t{}^1$ and $\tilde{S}_t{}^2$), at time t.
After meta-clustering, a $P \times P$ word transition matrix (Π) is defined as:

$$\Pi = \begin{bmatrix} \pi(\tilde{D}_t = \tilde{D}_1) \\ \vdots \\ \pi(\tilde{D}_t = \tilde{D}_p) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1P} \\ \vdots & \ddots & \vdots \\ \pi_{P1} & \cdots & \pi_{pp} \end{bmatrix} \tag{5}$$

where P is the number of unique words in the dictionary. This matrix is learned by estimating the transition probabilities $\pi_{ij} = P(\tilde{D}_t = j | \tilde{D}_{t-1} = i), j, i \in \tilde{D}$ over time.

2) Testing phase:

- *Real time Joint Position and Motor current inference:* In order to perform inference at different hierarchical levels, an interactive modified Markov Jump Particle Filter (IM-MJPF) has been employed, as illustrated in figure 3, to extract the knowledge embodied in the

coupled GDBN (C-GDBN). The MJPF is a switching model that uses a combination of a particle filter (PF), to predict discrete superstates, and Kalman Filters (KF) for continuous states prediction and estimation. Such a switching behavior, between the dynamic transitions at discrete/continuous levels and observations, enables the updating of the belief about the motor current state and its corresponding position by passing local messages in simultaneous inference modes, namely the predictive or causal inference (top-down) and the diagnostic inference (bottom-up). The top down messages from $\tilde{D}$ to $\tilde{X}$ depend on the clustered values, including super states with their corresponding mean and covariance statistics. The temporal predictive messages $(\pi(\tilde{D}_t) = P(\tilde{D}_t|\tilde{D}_{t-1}))$ and $(\pi(\tilde{S_t}^n) = P(\tilde{S_t}^n|\tilde{S}^n_{t-1}), \pi(\tilde{X_t}^n) = P(\tilde{X_t}^n|\tilde{X}^n_{t-1}))$ for both modalities depend on the dynamic rule stored in the model (intra-slice messages) as shown in figure 2.
The particle filter is employed to predict the discrete values (or words) relied in the word transition matrix encoded in the dynamic model as a proposal distribution. Once the word is predicted, each superstate of the two modalities is inherently contained within the word. Then, for each particle, a Kalman filter is employed to predict continuous states (position and motor current) of the agent using this equation: (6).

$$\tilde{X}^n_t = A\tilde{X}^n_{t-1} + BU^{\tilde{S}^n_t} + W^n_t \tag{6}$$

where A and B are the dynamic model matrix and the control model matrix, respectively. And $U^{\tilde{S}^n_t}$ is a control vector which represents the dynamic rules of the signal temporal evolution encoded in the superstate.
Once the evidence is observed, a message is backward-propagated from the bottom level towards the higher levels (λ messages as shown in figure 2). This message (reasoning) is based on likelihood models consisting of messages $(\lambda(\tilde{D}_t), \lambda(\tilde{S_t}^n), \lambda(\tilde{X_t}^n))$, in which all continuous nodes are Gaussian-distributed variables and the two GDBN models are conditionally independent given the interaction node:

$$\lambda(\tilde{D}_t) = \prod_{i=1}^{n} \lambda(\tilde{S_t}^n) \tag{7}$$

$$\lambda(\tilde{S_t}^n) = \lambda(\tilde{X_t}^n) P(\tilde{X_t}^n|\tilde{S_t}^n) \tag{8}$$

$$\lambda(\tilde{X_t}^n) = P(\tilde{Z_t}^n|\tilde{X_t}^n) \approx N(\mu_{\tilde{Z_t}^n}, \Sigma_{\tilde{Z_t}^n}) \tag{9}$$

Consequently, each posterior distribution ($belief = \pi(.) * \lambda(.)$) at word, superstate, and state level is updated according to $\lambda(\tilde{D}_t), \lambda(\tilde{S_t}^n), \lambda(\tilde{X_t}^n)$, respectively. Furthermore, these beliefs will represent the initial states on the next time instance.

- *Multi-modal Abnormality Measurements:*
  During the journey, the USV checks whether it is in an optimal state or experiencing an abnormality by predicting the motor current and GPS signal. When new data arrive, the trained model uses the GDBN to make sense of it and spot different types or levels of unusual patterns. To measure the motor current or GPS deviation we use an abnormality indicator at superstate level defined as a distance between the predictive message $(\pi(\tilde{S_t}^1))$ and the diagnostic message $(\lambda(\tilde{S_t}^1))$ entering the node $\tilde{S_t}^1$.
  We use the symmetric Kullback-Leibler Divergence (KLDA) to measure the similarity between the two discrete probability distributions $(\pi(\tilde{S_t}^1)$ and $\lambda(\tilde{S_t}^1))$ defined by:

$$\begin{aligned} KLDA = D_{KL}\left(\pi(\tilde{S_t}^1) \parallel \lambda(\tilde{S_t}^1)\right) \\ + D_{KL}\left(\lambda(\tilde{S_t}^1) \parallel \pi(\tilde{S_t}^1)\right) \end{aligned} \tag{10}$$

  where $\parallel$ identifies the divergence.
  Similarly, we use the Bhattacharyya distance (BD) to measure the probabilistic distance between the two continous probability distributions $(\pi(\tilde{X_t}^1)$ and $\lambda(\tilde{X_t}^1))$ defined by:

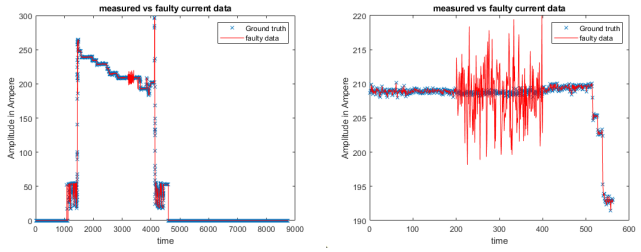
$$BD(p,q) = -ln\left(\sum_{i=1}^{k} \sqrt{p_i * q_i}\right) \tag{11}$$

  where $p = \pi(\tilde{X_t}^1)$, $q = \lambda(\tilde{X_t}^1)$, and k represents the number of elements considered in the computation. Furthermore, this can be calculated for all other possible sensors attached to the agent including GPS.
- *Evaluation matrices:*
  To assess the performance of our multi-modal abnormality detection approach, we use the Receiver Operating Characteristic (ROC) curve. The ROC curve visually shows how a model performs at different classification thresholds. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) for all possible threshold values. The AUC, or Area Under the Curve, summarizes this graph into a single number, indicating the model's ability to distinguish between classes. An AUC of 1.0 reflects perfect classification, whereas an AUC of 0.5 suggests the model is no better than making random guesses.

IV. Simulation and Experimental Results

A permanent magnet synchronous motor (PMSM) is a motor known for its outstanding dynamic performance and high reliability. It is extensively applied in areas such as electric vehicles, rail transportation, smart manufacturing, and more [11]. Our USV uses PMSM as its core engine.

An inter-turn short circuit (ITSC) fault in a stator typically happens when the insulation between the windings of the same phase breaks down. In this section, we perform the simulation of the ITSC fault to test our abnormality detection framework. ITSC in the stator results in a huge eddy current in the short circuit [12]. This will increase the RMS (Root Mean Square) current amplitude in the motor. This is simulated by adding Gaussian noise to the real motor current data, as shown in figure 5.



(a) Normal current data vs faulty (b) Chunked version-Visualization

Fig. 5: Normal vs Faulty RMS Current signal

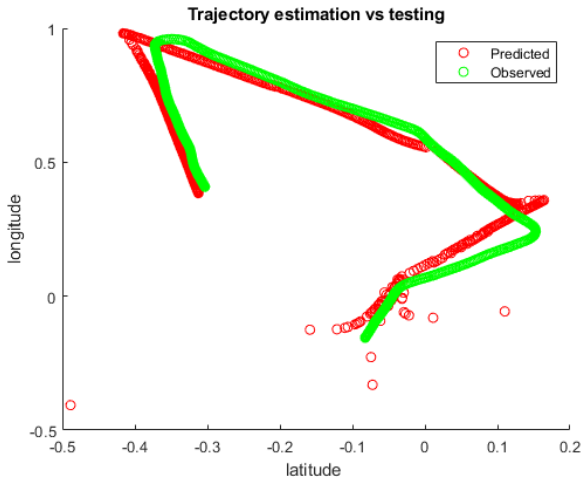


Fig. 6: Trajectory (chunk of the GPS data) estimation vs observed

The multi-sensor tracking capability of the USV is demonstrated by the spatio-temporal prediction of the trajectories, as in figure 6. Consequently, the motor current can be estimated based on the interaction between the two sensors as shown in figure 7.

To enable scalable, efficient, and stochastic learning of the posterior, we divided our dataset into 8 examples (mini-batches) during the training of the Bayesian generative model. During testing, we use one of the batches containing a simulated motor fault as the test set.

The decision making process of the autonomous agent is illustrated in figures 8 and 9, in which the abnormality is detected if it passes the predefined threshold (set during training). This demonstrates the effectiveness of our framework in detecting abnormalities. Moreover, the cause of the abnormality can be traced to either a GPS error or a motor current fault. This makes our framework explainable.

In addition, we carried out a performance test for the continuous state (motor current) abnormality detection. The ROC curves shown in both figures 10 and 11 indicate that our framework achieves good abnormality detection efficiency.

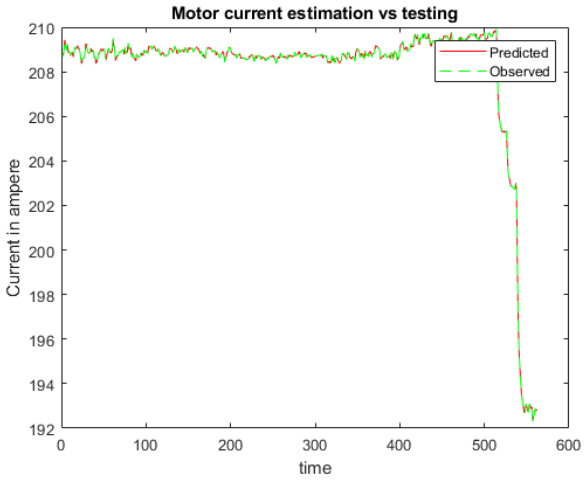


Fig. 7: Motor current estimation vs observed (chunk of motor current data cross ponding to GPS shown in figure 6)

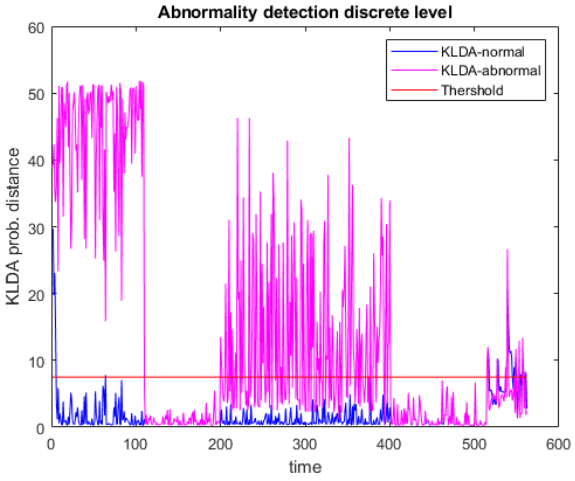


Fig. 8: Tracking abnormality at cluster level (Superstate level)

V. Conclusion and Future Work

This paper proposes a multi-modal Bayesian generative model for fault detection and probabilistic inference of system states in an Unmanned Surface Vehicle (USV). The proposed approach uses coupled Dynamic Bayesian Networks (DBNs) to assimilate different observations over time, enabling incremental learning and inherent explainability. The method predicts future states based on past observations, providing an interpretable explanation of the model's reasoning.

The performance of the proposed method is evaluated using real-world data and simulated faults that mimic actual motor malfunctions. The results indicate that the approach can predict and detect faults in a USV with high accuracy. Furthermore, it is capable of identifying the root cause of the defect.

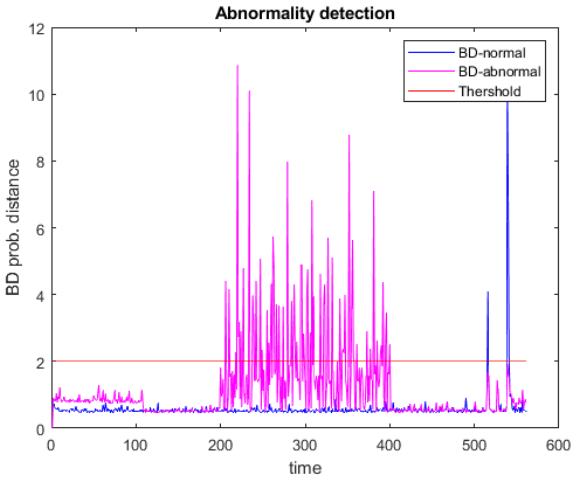


Fig. 9: Tracking abnormality at continuous level

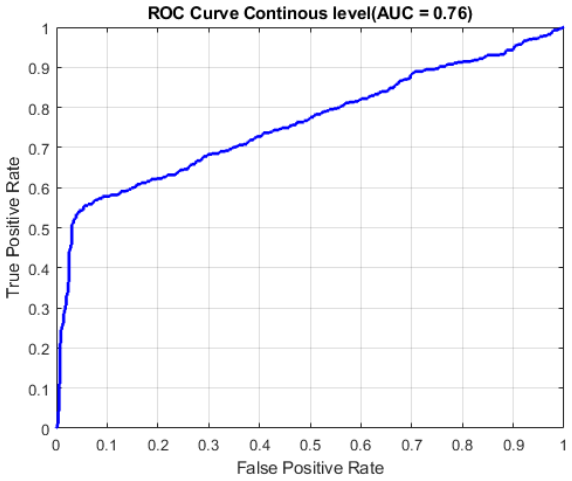


Fig. 11: ROC (BD) curve for motor abnormality detection

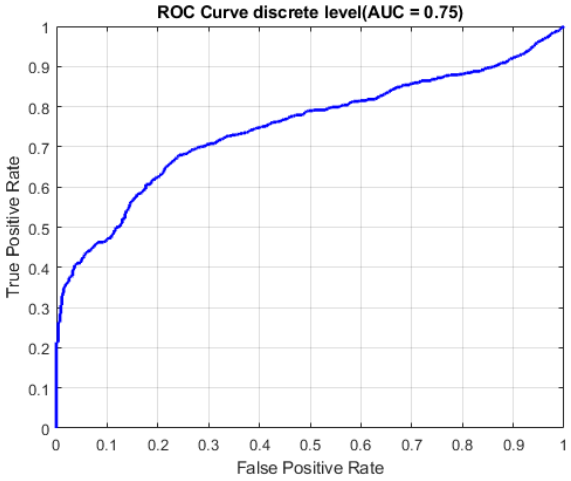


Fig. 10: ROC (KLDA) curve for motor abnormality detection

In our future work, we are planning to calculate meta-cluster level (word level) abnormality to detect possible anomalies on the agent and to learn the time evolution synchronization between the multi-modal sensors used in our approach.

REFERENCES

[1] S. Baccari, M. Hadded, H. Ghazzai, H. Touati and M. Elhadef, "Anomaly Detection in Connected and Autonomous Vehicles: A Survey, Analysis, and Research Challenges," in IEEE Access, vol. 12, pp. 19250-19276, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3361829.

[2] G. Slavic, M. Bracco, L. Marcenaro, D. M. Gómez, C. Regazzoni, and P. Zontone, "Joint Data-Driven Analysis of Visual-Odometric Anomaly Signals in Generative Ai-Based Agents," 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), Seoul, Korea, Republic of, 2024, pp. 264-268, doi: 10.1109/ICASSPW62465.2024.10626634.

[3] S. Memon, A. Krayani, P. Zontone, L. Marcenaro, D. M. Gomez, and C. Regazzoni, "Learning 3D LiDAR Perception Models for Self-Aware Autonomous Systems," 2024 27th International Conference on Information Fusion (FUSION), Venice, Italy, 2024, pp. 1-8, doi: 10.23919/FUSION59988.2024.10706364.

[4] https://www.bayesserver.com/docs/techniques/anomaly-detection/

[5] Ö. Gültekin, E. Cinar, K. Özkan, and A. Yazıcı, "Multisensory data fusion-based deep learning approach for fault diagnosis of an industrial autonomous transfer vehicle," Expert Systems with Applications, vol. 200, p. 117055, Aug. 15, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117055.

[6] Q. Shi and H. Zhang, "Fault Diagnosis of an Autonomous Vehicle With an Improved SVM Algorithm Subject to Unbalanced Datasets," in IEEE Transactions on Industrial Electronics, vol. 68, no. 7, pp. 6248-6256, July 2021, doi: https://doi.org/10.1109/TIE.2020.2994868.

[7] Pratama, P.S., Gulakari, A.V., Setiawan, Y.D. et al. Trajectory tracking and fault detection algorithm for automatic guided vehicle based on multiple positioning modules. Int. J. Control Autom. Syst. 14, 400–410 (2016), doi: https://doi.org/10.1007/s12555-014-0294-y

[8] Abci, B., El Badaoui El Najjar, M., Cocquempot, V. et al. An informational approach for sensor and actuator fault diagnosis for autonomous mobile robots. J Intell Robot Syst 99, 387–406 (2020), doi: https://doi.org/10.1007/s10846-019-01099-7

[9] P. Castaldi, S. Farsoni, M. Menghini and S. Simani, "Data–Driven Fault Detection and Isolation of the Actuators of an Autonomous Underwater Vehicle," 2021 5th International Conference on Control and Fault-Tolerant Systems (SysTol), Saint-Raphael, France, 2021, pp. 139-144, doi: https://doi.org/10.1109/SysTol52990.2021.9595605.

[10] Changxin Zhang and etal (2022). A data-driven fault detection and diagnosis method via just-in-time learning for unmanned ground vehicles. Automatika, 64(2), 277–290, https://doi.org/10.1080/00051144.2022.2142924

[11] Chen, Z.; Liang, K.; Peng, T.; Wang, Y. Multi-Condition PMSM Fault Diagnosis Based on Convolutional Neural network Phase Tracker. Symmetry 2022, 14, 295, doi:https://doi.org/10.3390/sym14020295

[12] Chen, X.; Qin, P.; Chen, Y.; Zhao, J.; Li, W.; Mao, Y.; Zhao, T. Inter-Turn Short Circuit Fault Diagnosis of PMSM. Electronics 2022, 11, 1576, doi: https://doi.org/10.3390/electronics11101576

[13] Ayankoso, S., He, Y., Gu, F., Ball, A. (2024). Multisensory Data-Based Fault Diagnosis of Induction Motors Using 1D and 2D Convolutional Neural Networks. In: Ball, A.D., Ouyang, H., Sinha, J.K., Wang, Z. (eds) Proceedings of the UNIfied Conference of DAMAS, IncoME and TEPEN Conferences (UNIfied 2023). TEPEN IncoME-V DAMAS 2023 2023 2023. Mechanisms and Machine Science, vol 152. Springer, Cham, doi: https://doi.org/10.1007/978-3-031-49421-5_92

[14] Shaoxuan Xia, Xiaofeng Zhou, Haibo Shi, Shuai Li, Chunhui Xu, A fault diagnosis method with multi-source data fusion based on hierarchical attention for AUV, Ocean Engineering, Volume 266, Part 1,2022,112595, ISSN 0029-8018, doi: https://doi.org/10.1016/j.oceaneng.2022.112595.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 405–410
DOI: 10.15439/2025F9025
ISSN 2300-5963 ACSIS, Vol. 43

Unified Data-Driven Prediction of Photovoltaics Output from Weather and Geographic Data Across Diverse Systems

Joanna Wójcicka
Faculty of Electrical Engineering,
Automatics, Computer Science
and Biomedical Engineering,
AGH University of Krakow,
Al. Mickiewicza 30,
30-059 Krakow, Poland
Email: wojcickajoanna1@gmail.com,
joannawoj@student.agh.edu.pl

Tomasz Hachaj
ORCID: 0000-0003-1390-9021
Faculty of Electrical Engineering,
Automatics, Computer Science
and Biomedical Engineering,
AGH University of Krakow,
Al. Mickiewicza 30,
30-059 Krakow, Poland
Email: thachaj@agh.edu.pl

***Abstract*—As the adoption of renewable energy continues to rise, precise forecasting of solar power generation is essential for optimizing energy storage and distribution. This article explores the prediction of energy output in photovoltaic systems using machine learning models that leverage environmental and geographical factors. The study utilizes data from 9,182 private photovoltaic installations across Poland and publicly available weather records. Additionally, a data preprocessing method was introduced to filter out non-useful data, such as records indicating malfunctioning installations, ensuring that only relevant information is used for prediction. Key variables considered include temperature, cloud cover, wind speed, and solar panel efficiency. This paper studies the effectiveness of data-driven energy production forecasting methods, namely linear regression, polynomial regression, decision tree regression, random forest regression, and multilayer fully connected artificial neural network, designed to make predictions for various installations with different parameters and geographical locations, considering atmospheric conditions in contrast to frequently published articles in which predictions are fitted on data from a single photovoltaic installation. Due to this, our work has broad research value, explores the boundaries and limitations of such approaches, and can be considered a reference for energy engineers, computer scientists, and researchers.**

***Index Terms*—Photovoltaic Infrastructure, Machine Learning, Prediction, Feature Selection, Data-Driven**

I. Introduction

RENEWABLE energy has experienced substantial growth in recent decades, driven by the increasing global population and the recognition of the finite nature of fossil fuel resources. The transition to renewable energy sources is widely regarded as a viable strategy to reduce dependence on fossil fuels while mitigating the environmental impacts of their extraction and consumption. Technological advancements in renewable energy have been considerable, with numerous solutions already implemented in practical applications in private and industrial sectors. Additionally, reliable environmental and production data availability has improved significantly, caused by the growing adoption of monitoring systems in private photovoltaic installations.

Among renewable energy sources, solar and wind technologies have been recognized for their substantial potential to meet global energy demands. As noted by Gross et al. [1], these technologies were considered to have the most significant impact within the renewable energy sector. However, subsequent research by Olabi and Abdelkareem [2] emphasised that both solar and wind energy are highly dependent on environmental conditions, introducing variability and limiting their reliability as primary global energy sources.

One of the critical challenges in integrating renewable energy sources is the efficient management of energy storage and distribution to maintain a stable balance between generation and consumption. As discussed by Roberts et al. [3], the variability of renewable energy output, primarily influenced by environmental conditions, requires the implementation of adaptive and responsive strategies for storage and distribution. At a macro-level, particularly within centralised energy systems such as Nord Pool [4], this involves determining optimal periods for storing energy for future use versus distributing it to external sectors to meet immediate demand. In addition, these strategies should include dynamic pricing models that account for fluctuations in energy demand and the availability of generation resources, thereby enhancing the overall efficiency and reliability of the energy system.

Therefore, accurate forecasting of solar energy production has become a critical component in the efficient operation of modern energy systems. Reliable predictions enable better planning for energy storage, grid balancing, and market operations. Forecasting supports decision-making processes by providing information on expected energy availability, essential for maintaining supply stability, reducing reliance on backup fossil fuel generation, and optimising energy pricing and distribution strategies.

Two primary methodologies are commonly used to forecast

Topical area: Advanced Artificial Intelligence in Applications

the energy production of solar installations. The first is a deterministic approach, which involves the development of a physical model that represents the electrical behaviour of photovoltaic cells. This method simulates the underlying physical processes to estimate the energy output based on the known specifications and characteristics of the solar panel. A wide range of physical models can be applied to predict the performance of photovoltaic systems with varying configurations, as examined by Dolara et al. [5], [6].

An alternative approach to forecasting solar energy production involves using statistical and machine learning models, which are developed based on historical data from photovoltaic installations and relevant environmental variables that influence energy generation. These methods aim to identify and model the underlying relationships between features in the dataset to enable accurate prediction. Linear regression and artificial neural networks (ANNs) are the most commonly used models in this domain. Both approaches have been extensively applied in various research contexts with notable success. For example, Gratidi et al. [7], Barrera et al. [8], and Zazoum [9] each developed accurate prediction models tailored to different datasets, photovoltaic system architectures, and methodological frameworks. These studies highlight the flexibility and effectiveness of machine learning techniques in capturing the complexity inherent in solar energy production forecasting.

Despite significant advancements and promising outcomes in solar energy forecasting, considerable room for improvement remains. Studies by Gratidi et al. [7] and Barrera et al. [8] commonly relied on data collected from small-scale installations located in research centres, which, while controlled and consistent, may limit the model's ability to generalise to real-world scenarios. For predictive models to be applicable globally, it is crucial to incorporate diverse datasets drawn from various geographic locations and installation types. Additionally, many existing studies utilize well-structured solar radiation data complemented by ambient temperature measurements. Although this data can enhance model accuracy, it often depends on specialized infrastructure that may not be available in all regions. This reliance poses a challenge to scalability, particularly in areas lacking comprehensive meteorological monitoring systems. It highlights the need for models capable of operating effectively with more readily accessible environmental data.

A. Novelty of this paper

This work examines the effectiveness of statistical and machine learning models for predicting energy output in photovoltaic systems that leverage environmental and geographical factors. New to previous work dealing with similar topics, the study was conducted on a large dataset containing installations located almost all over a medium-sized European country. Our goal was to create a predictive model that can effectively predict energy production based on data such as the power of an installation, its geographic location, and weather data to improve forecast accuracy [10]. Thus, in this work, we study the effectiveness of data-driven energy production forecasting methods designed to make predictions for various installations with different parameters and geographical locations, taking into account atmospheric conditions, in contrast to frequently published articles in which predictions are fitted on data from a single photovoltaic installation [11]–[15]. Due to this fact, our work has broad research value, explores the boundaries and limitations of such an approach, and can be considered a reference for energy engineers, computer scientists, and researchers.

II. Materials and Methods

The foundation of this study is a dataset that accurately represents real-world conditions of photovoltaic installations. The data were collected from multiple private solar installations and supplemented with commonly observed meteorological variables to ensure a comprehensive analysis. The selected methodology prioritizes data that are accessible to the average user, eliminating the need for specialized professional measuring equipment. The dataset from photovoltaic installations covers the year 2020 and, due to licensing and legal restrictions, cannot be publicly shared with other researchers.

A. Data set

The initial database comprised 145,068,684 records from 9,182 photovoltaic installations in Poland through the PVmonitor monitoring system. The measurements taken from these installations were aggregated every three minutes, capturing various parameters such as energy production, efficiency over specific time intervals, the cumulative sum of energy generated since monitoring began, and the corresponding timestamps. Each installation was assigned a unique ID and included localization details and the manufacturer's power specifications for the photovoltaic panels.

The weather dataset was gathered to use only commonly available environmental information like temperature, cloudiness, or wind speed. To achieve this, the data was collected from a publicly available AccuWeather archive using a data scraping technique. This source was selected due to its frequent measurements and broad coverage of features relevant to this study. A notable characteristic of this dataset is the cloudiness variable, which is provided as a percentage. This numerical representation is more precise and requires less preprocessing than categorical cloud cover classifications.

B. Preprocessing

Since the installations dataset was collected from private users, it relied heavily on user input, with no control over the accuracy of the solar panel setup or the calibration of the monitoring devices. This resulted in a necessity to validate and prune data strictly. Several conditions must be satisfied to determine whether or not an installation is active and whether it should be considered.

- Power of Installation provided by a user cannot be equal to zero,

- Installation should at least once in a month reach over 40% efficiency,
- Installation's mean efficiency in winter (October through March) should be over 5%,
- Installation's mean efficiency in summer (April through September) should be over 10%,
- Installation has to be monitored throughout the entire year.

Applying these conditions reduced the number of installations to 32. Although this significantly decreased the amount of data available for energy production predictions, ensuring that the dataset contained only high-quality records was necessary. Despite the reduction, the remaining installations still contributed 181,306 measurement points, providing sufficient data for accurate predictions.

The next step was to combine both datasets into a consistent database. Since installation measurements were made every three minutes and weather data was aggregated hourly, it was necessary to aggregate the measurements to match this time frame. Aggregation was achieved by calculating the mean value of each efficiency of each Installation and the sum of power (P) and power increment (DPV). Finally, the data were aligned based on geographical location.

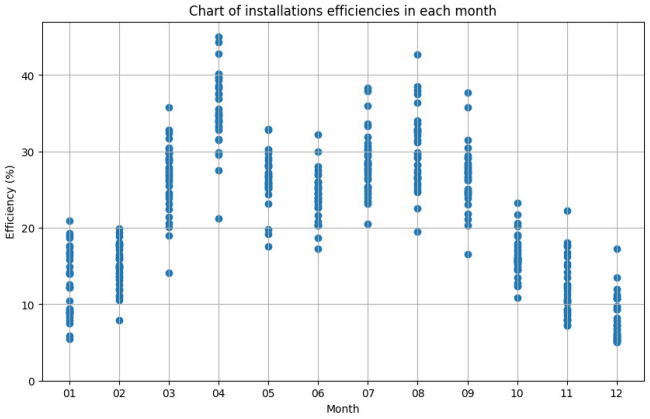


Fig. 1. Chart of mean installations' efficiencies aggregated monthly

After proper data filtering, it was possible to distinguish a desired trend in efficiency throughout the year presented in Figure 1. This trend is a direct result of yearly changes in the solar azimuth angle because photovoltaic installations are more efficient when the radiation reaches solar panels at around 90° angle [16], [17] (as a combination of azimuth and tilt angles) that occurs at the spring and autumnal equinox in Poland.

C. Experimental setup

The dataset was divided into training and testing sets in an 80%-20% ratio based on installation indexes, ensuring that instances from the same installation did not appear in both sets, thus maintaining a clear separation for evaluation. Due to the data being inconsistent regarding values range, the action taken to ensure good performance of models was scaling both training and test data using the Equation 1.

$$Z = \frac{x - \rho}{\sigma} \tag{1}$$

where:
x is the value of sample,
ρ is a mean values of all samples and
σ is the standard derivation.

The linear regression model was proposed as a baseline model for comparison, as it is a simple machine learning model that does not require additional hyperparameters tuning. Another models we also evaluated were polynomial regression, K-nearest neighbour regression (KNN), decision tree regression, random Forest regression (RFR) and artificial neural network (ANN). The architecture of implemented ANN is detailed in Table I.

TABLE I
ARCHITECTURE OF IMPLEMENTED ARTIFICIAL NEURAL NETWORK

Layer	Layer Type	Number of Neurons	Activation Function
Input	Input	7 (number of features)	
Hidden n°1	Dense	10	ReLU
Hidden n°2	Dense	10	ReLU
Hidden n°3	Dense	10	ReLU
Output	Dense	1	Linear

D. Metrics used to evaluate models' performance

Evaluating machine learning models using standardised metrics is essential to ensure comparability. The performance of the models was assessed on the following metrics:

- Normalized Root Mean Squared Error

$$NRMSE_{minmax} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}},\ y_{max} \neq y_{min} \tag{2}$$

- Mean Squared Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

- R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \exists i, j \text{ such that } y_i \neq y_j \text{ for } i \neq j, \tag{4}$$

where:
N is a number of testing samples,
y_i are the predicted values,
$\hat{y}_i$ are the actual values,
$y_{\max}$ and $y_{\min}$ are the maximum and minimum values in the dataset,
$\bar{y}$ is mean of the observed values.

III. RESULTS

Table II presents the evaluation metrics calculated for the training and test datasets. The NRMSE$_{minmax}$ metric is the primary indicator of predictive accuracy, as it assesses the performance of the models relative to the data scale. In addition,

TABLE II
EVALUATION OF MODELS ON TRAINING AND TEST DATA

Test Set			
Model	**NRMSE$_{minmax}$**	**MAE**	**R^2**
Linear Regression	0.1809	17.4593	0.2390
Regression with Polynomial Features	0.1210	11.1261	0.6591
K-Nearest Neighbors (KNN)	0.1304	10.5962	0.6045
Decision Tree Regressor (DTR)	0.1237	9.8932	0.6445
Random Forest Regressor (RFR)	**0.1113**	**9.1079**	**0.7122**
Artificial Neural Network (ANN)	**0.1137**	**9.4136**	**0.6994**
Training Set			
Model	**NRMSE$_{minmax}$**	**MAE**	**R^2**
Linear Regression	0.2019	17.4593	0.2290
Regression with Polynomial Features	0.1260	10.1974	0.6996
K-Nearest Neighbors (KNN)	**0.0637**	**4.0101**	**0.9231**
Decision Tree Regressor (DTR)	0.1179	7.9926	0.7228
Random Forest Regressor (RFR)	**0.0254**	**1.5648**	**0.9877**
Artificial Neural Network (ANN)	0.1189	8.7776	0.7322

MSE and R^2 serve as complementary metrics, offering further insight into the overall effectiveness of the models.

In the test set, the NRMSE$_{minmax}$ values indicate that the best-performing models are RFR and ANN, with RFR showing a slight advantage. This suggests that RFR is particularly effective at capturing the dataset's complexity while maintaining high predictive accuracy. In contrast, models such as DTR and polynomial regression obtained satisfactory results but did not achieve the same level of accuracy as RFR and ANN.

Comparison with training data provides insight into the generalisability of the models. Although RFR and KNN achieve exceptionally high performance on the training data, their performance on the test dataset is noticeably lower. Among the evaluated models, ANN exhibits the most consistent performance in both training and test datasets, indicating a strong balance between predictive accuracy and generalisability.

There is a notable improvement between the baseline model (linear regression) and the more advanced models, emphasizing the effectiveness of non-linear approaches. Among the models tested, RFR demonstrates the best overall performance, closely followed by ANN. Additionally, based solely on the evaluation metrics, regression with polynomial features and KNN exhibit strong performance for this dataset. Further insights into the models can be obtained by analysing Figures 2 and 3, which compares the real and predicted values for each model across training and test datasets. These visualizations offer a clear representation of the models' generalization capabilities and enable a direct comparison of their performance on training versus test data.

To objectively evaluate the performance of the models, it is

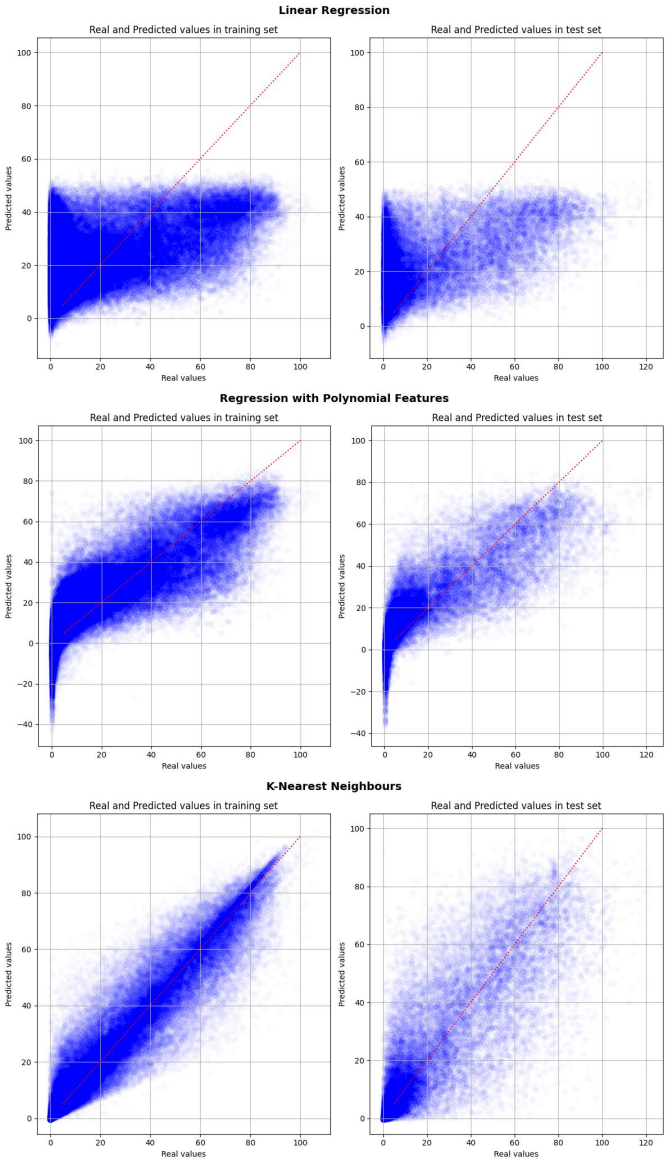


Fig. 2. Comparison of training and test predicted values vs actual values - linear regression, regression with polynomial features and K-nearest neighbour

crucial to compare the results with those reported in other studies. Table III presents a performance summary of the selected models in various data sets and methodologies. The comparison is based on NRMSE values, with linear regression serving as the baseline, normalised to 100% using the min-max method in this study and different normalisation approaches reported in the studies [7], [18], [19]. The relative performance of other models is then expressed as a percentage improvement or decline relative to this baseline. This approach ensures the comparability of the metric between different methodologies.

IV. DISCUSSION

The comparison between the training and test results (Table II) provides insights into the generalisation capabilities of the

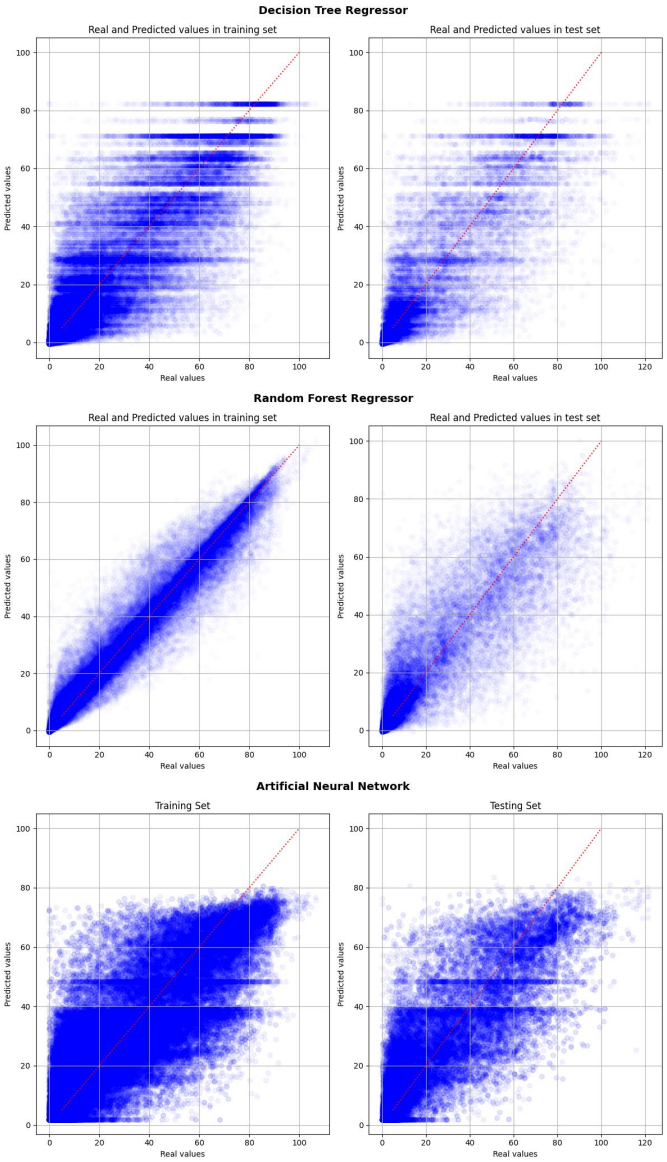


Fig. 3. Comparison of training and test predicted values vs actual values - decision tree regressor, random forest regression and artificial neural network

models. Although RFR and KNN achieve high performance on the training set, their test performance declines, indicating overfitting. This suggests that these models capture patterns specific to the training data but struggle to generalise to unseen data. In contrast, ANN exhibits the most consistent performance across both datasets, highlighting its robust generalisation ability. This suggests that ANN effectively learns underlying relationships in the data without overfitting, making it a more reliable model for photovoltaic energy prediction.

The observed differences in model performance may be attributed to the nonlinear nature of photovoltaic energy generation, which is influenced by multiple environmental factors [20], [21]. Although decision tree-based models like RFR effectively handle complex interactions, their sensitivity to training data can lead to overfitting. On the other hand, the ANN model, with its ability to capture nonlinear dependencies, demonstrates a better balance between flexibility and generalisation.

TABLE III
COMPARISON OF MODEL PERFORMANCE BASED ON NRMSE VALUES, WITH LINEAR REGRESSION SERVING AS THE BASELINE (100%), AND THE PERFORMANCE OF OTHER MODELS EXPRESSED AS A PERCENTAGE IMPROVEMENT OR DECLINE RELATIVE TO THIS BASELINE

Model	relative performance (in %)			
	S	S_1 [7]	S_2 [18]	S_3 [19]
LR	100	100	100	100
KNN	138.7			173.33
RFR	162.53		160.8	288.89
ANN	158.10	71.38	350.33	236.36

In Figures 2–3, there is a noticeable increase in point density for lower efficiency values in all models. This can be attributed to the dataset's characteristics, where higher efficiency values are less frequent than lower ones. Power generation is typically most efficient during specific periods, particularly in summer and around midday, if weather conditions are favourable. This distribution reflects the natural variability of photovoltaic systems, which is strongly influenced by seasonal and temporal factors. The effect is even more pronounced in the training set, probably because of the smaller number of samples available. The uneven distribution of training samples in the dataset is an unfavorable phenomenon for machine learning. It could have impacted the training and final performance of the models considered.

Table III compares the findings of this study with related works. Graditi et al. [7] analysed a single photovoltaic installation under controlled conditions, where linear regression performed well due to simple variables. Jebli et al. [18] used historical weather and radiation data and found ANN to be the best performer, reinforcing the advantage of nonlinear models. Similarly, Chaaban and Alfadl [19] studied solar power plant data, with RFR and KNN outperforming linear regression, aligning closely with the findings of this research.

The presented metric provides valuable information on the dataset's characteristics and underscores the performance improvements of RFR and ANN models in this specific problem setting. Compared to other studies, this research's outcomes are satisfactory, as the observed trends across different models align well with the findings reported in the literature.

V. CONCLUSION

The aim of this study was to create machine learning models to predict energy production in photovoltaic infrastructures with different parameters, locations, and under the influence of variable weather conditions. It is an area of study of growing importance, given the increasing adoption of renewable energy as a primary energy source across a large number of sectors. Its significance is particularly evident for energy management systems, which depend on accurate energy production predictions to adjust distribution strategies accordingly to resource availability.

We have shown that a single data-driven predictive model can accurately predict the output of many different installations. Among the evaluated models, the ANN showed the most consistent performance across both training and test datasets, demonstrating strong generalisability. This indicates that ANN-based approaches are particularly suitable for solar energy forecasting. In contrast, the Random Forest and K-Nearest Neighbors models exhibited indications of overfitting, as evidenced by a notable difference in performance between training and test sets. However, the RFR remains a highly accurate model in general, showcasing strong predictive capabilities.

There is still room for improvement in models' performance that was not addressed in this study. For future work, it would be beneficial to incorporate data from installations that span a broader geographical area. This would enhance the generalisation capability of the models, enabling them to make accurate predictions across diverse contexts and conditions. In their current form, the models developed in this project are not well-suited for deployment in locations beyond Poland and nearby areas.

REFERENCES

[1] R. Gross, M. Leach, and A. Bauen, "Progress in renewable energy," *Environment international*, vol. 29, no. 1, pp. 105–122, 2003. doi: https://doi.org/10.1016/S0160-4120(02)00130-7

[2] A. Olabi and M. A. Abdelkareem, "Renewable energy and climate change," *Renewable and Sustainable Energy Reviews*, vol. 158, p. 112111, 2022. doi: https://doi.org/10.1016/j.rser.2022.112111

[3] B. P. Roberts and C. Sandberg, "The role of energy storage in development of smart grids," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 1139–1144, 2011. doi: 10.1109/JPROC.2011.2116752

[4] E. Serban, M. Ordonez, C. Pondiche, K. Feng, M. Anun, and P. Servati, "Power management control strategy in photovoltaic and energy storage for off-grid power systems," in *2016 IEEE 7th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, 2016. doi: 10.1109/PEDG.2016.7527070 pp. 1–8.

[5] A. Dolara, S. Leva, and G. Manzolini, "Comparison of different physical models for pv power output prediction," *Solar energy*, vol. 119, pp. 83–99, 2015. doi: https://doi.org/10.1016/j.solener.2015.06.017

[6] A. Bączkiewicz and J. Wątróbski, "Selection of floating photovoltaic system considering strong sustainability paradigm using ssp-copras method," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS*, 09 2023. doi: 10.15439/2023F492 pp. 901–905.

[7] G. Graditi, S. Ferlito, and G. Adinolfi, "Comparison of photovoltaic plant power production prediction methods using a large measured dataset," *Renewable energy*, vol. 90, pp. 513–519, 2016. doi: https://doi.org/10.1016/j.renene.2016.01.027

[8] J. M. Barrera, A. Reina, A. Maté, and J. C. Trujillo, "Solar energy prediction model based on artificial neural networks and open data," *Sustainability*, vol. 12, no. 17, p. 6915, 2020. doi: https://doi.org/10.3390/su12176915

[9] B. Zazoum, "Solar photovoltaic power prediction using different machine learning methods," *Energy Reports*, vol. 8, pp. 19–25, 2022. doi: https://doi.org/10.1016/j.egyr.2021.11.183

[10] F. Koutenský, J. Pihrt, M. Čepek, V. Rybar, P. Šimánek, M. Kepka, K. Jedlička, and K. Charvat, "Combining local and global weather data to improve forecast accuracy for agriculture," in *Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS*, 11 2024. doi: 10.15439/2024F5990 pp. 77–82.

[11] M. Ceci, R. Corizzo, F. Fumarola, D. Malerba, and A. Rashkovska, "Predictive modeling of pv energy production: How to set up the learning task for a better prediction?" *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 956–966, 2017. doi: 10.1109/TII.2016.2604758

[12] Y. Ledmaoui, A. El Maghraoui, M. El Aroussi, R. Saadane, A. Chebak, and A. Chehri, "Forecasting solar energy production: A comparative study of machine learning algorithms," *Energy Reports*, vol. 10, pp. 1004–1012, 2023. doi: https://doi.org/10.1016/j.egyr.2023.07.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484723011228

[13] N. Dimitropoulos, N. Sofias, P. Kapsalis, Z. Mylona, V. Marinakis, N. Primo, and H. Doukas, "Forecasting of short-term pv production in energy communities through machine learning and deep learning algorithms," in *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2021. doi: 10.1109/IISA52424.2021.9555544 pp. 1–6.

[14] M. Pikus and J. Wąs, "Predictive modeling of renewable energy purchase prices using deep learning based on polish power grid data for small hybrid pv microinstallations," *Energies*, vol. 17, no. 3, 2024. doi: 10.3390/en17030628. [Online]. Available: https://www.mdpi.com/1996-1073/17/3/628

[15] M. Pikus and J. Was, "Using deep neural network methods for forecasting energy productivity based on comparison of simulation and dnn results for central poland—swietokrzyskie voivodeship," *Energies*, vol. 16, no. 18, 2023. doi: 10.3390/en16186632. [Online]. Available: https://www.mdpi.com/1996-1073/16/18/6632

[16] X. Chen, Y. Li, B. Zhao, and R. Wang, "Are the optimum angles of photovoltaic systems so important?" *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109791, 2020. doi: https://doi.org/10.1016/j.rser.2020.109791

[17] M. Z. Jacobson and V. Jadhav, "World estimates of pv optimal tilt angles and ratios of sunlight incident upon tilted and tracked pv panels relative to horizontal panels," *Solar energy*, vol. 169, pp. 55–66, 2018. doi: https://doi.org/10.1016/j.solener.2018.04.030

[18] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," *Energy*, vol. 224, p. 120109, 2021. doi: https://doi.org/10.1016/j.energy.2021.120109

[19] A. K. Chaaban and N. Alfadl, "A comparative study of machine learning approaches for an accurate predictive modeling of solar energy generation," *Energy Reports*, vol. 12, pp. 1293–1302, 2024. doi: https://doi.org/10.1016/j.egyr.2024.07.010

[20] M. A. Hassan, N. Bailek, K. Bouchouicha, A. Ibrahim, B. Jamil, A. Kuriqi, S. C. Nwokolo, and E.-S. M. El-kenawy, "Evaluation of energy extraction of pv systems affected by environmental factors under real outdoor conditions," *Theoretical and Applied Climatology*, vol. 150, no. 1, pp. 715–729, 2022. doi: https://doi.org/10.1007/s00704-022-04166-6

[21] T. Hai, M. Aksoy, and K. Nishihara, "Optimized mppt model for different environmental conditions to improve efficacy of a photovoltaic system," *Soft Computing*, vol. 28, no. 3, pp. 2161–2179, 2024. doi: https://doi.org/10.1007/s00500-023-09195-5

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 411–416
DOI: 10.15439/2025F3970
ISSN 2300-5963 ACSIS, Vol. 43

Evaluating Effectiveness of Nonlinear Dimensionality Reduction in Hedge Funds' Returns Forecasting

Milica Zukanović, Aleksa Radosavčević
0000-0003-3650-8327, 0009-0006-0069-2117
University of Belgrade – Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: milica.zukanovic@fon.bg.ac.rs, ar20235065@student.fon.bg.ac.rs

Ana Poledica, Pavle Milošević, Ivan Luković
0000-0002-5929-6446, 0000-0002-5943-6023, 0000-0003-1319-488X
University of Belgrade – Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: {ana.poledica; pavle.milosevic; ivan.lukovic}@fon.bg.ac.rs

***Abstract*—Hedge funds (HF) are actively managed investment vehicles employing diverse and often complex strategies. Accurate returns forecasting is essential for optimizing their performance and managing risk. This paper investigates the application of nonlinear dimensionality reduction (DR) methods in forecasting HF strategy performance, building upon prior work in financial time series analysis. We evaluate the effectiveness of Kernel Principal Component Analysis (KPCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and autoencoders on predictive performance of machine learning models. The extracted features are fed into several forecasting models, Support Vector Machine (SVM) with linear and nonlinear kernels, Neural Network (NN), and Extreme Gradient Boosting (XGB), to predict returns of five diverse HF investment strategies: Commodity Trading Advisors, Equity Long Short, Equity Market Neutral, Fixed Income Arbitrage, and Global Macro. The results demonstrate that nonlinear DR methods, particularly autoencoders, and KPCA combined with NN, significantly outperform other techniques. Our findings highlight the value of nonlinear transformations in enhancing predictive accuracy for HF returns time series.**

***Index Terms*—dimensionality reduction, hedge funds, PCA, KPCA, t-SNE, UMAP, autoencoders.**

I. Introduction

THE hedge fund industry has experienced substantial growth [1], due to investors' demand for alternative sources of return and portfolio diversification. Compared to traditional financial market forecasting, hedge funds introduce further layers of complexity, including the lack of transparency, and very diverse strategies [2]. A wide array of factors in investment models often lead to noise and overfitting causing poor decision-making. However, by effectively using DR techniques, hedge fund analysts can enhance predictive model accuracy, gain deeper insights into portfolio risk, and make smarter investment decisions.

DR techniques in machine learning (ML) are typically implemented for several key reasons: to mitigate the curse of dimensionality, prepare data for ML algorithms, boost model performance, and enhance interpretability and explainability [3]. DR algorithms are commonly distinguished based on linearity, interpretability, computational cost, data type, hyper-parameters, etc. [4], while the success of DR algorithms is highly dependent on data characteristics, quality and size [5]. A recent survey [3] highlights approaches developed for time-series analysis e.g. autoencoders, Principal Component Analysis (PCA) and KPCA, while PCA and UMAP are commonly used for financial times series [6]. Popular nonlinear techniques, t-SNE and UMAP, are shown to be successful for both visualization and DR in financial time series analysis [7]. Finally, more transparent approaches, e.g. based on fuzzy sets and fuzzy rough sets, are proposed in literature [8]. Still, the authors conclude that while nonlinear DR approaches often yield superior results on artificial datasets, this performance does not necessarily translate to real-world datasets [5].

In forecasting, DR algorithms serve as inputs to predictive models to enhance forecasting performance. The widespread adoption of ML models in financial forecasting is driven by their ability to capture nonlinear relationships and adapt to rapidly changing market conditions. According to a recent review [9], the predominant ML algorithms for stock market prediction are SVM and NN, followed by deep learning models. In [10], the authors demonstrate benefits of using ML approach to forecasting stock returns; the best performing models are boosting techniques and neural networks. Finally, many approaches tend to incorporate DR in their ML framework, e.g. [11]

In this research, we examine the effect of nonlinear DR techniques, selected from various category groups on performance of hedge fund returns forecasting. In our study, PCA will serve as a linear benchmark, and KPCA, t-SNE, UMAP

Topical area: Advanced Artificial Intelligence in Applications

and autoencoders will be exploited to uncover the underlying structure of the data, modeling nonlinearity from different perspectives. As part of ML forecasting framework, predominant ML algorithms are selected based on the surveyed research papers, i.e. linear and kernel SVM are chosen to match diverse combinations of linear/nonlinear DR algorithms, NNs as most effective in hedge fund return prediction, and XGB as advanced representative of boosting algorithms. The datasets used in this research contain time series of monthly returns for the period of 1999-2023 for 5 HF investment, diverse in types of risks, financial instruments and investment philosophy [12]. The research addresses the following questions aimed at exploring the effectiveness of nonlinear DR in case of hedge funds' returns forecasting:

RQ1: How does the selection of nonlinear DR techniques affect the predictive power of ML forecasting models for hedge funds' return time series?

RQ2: Which ML forecasting models demonstrate best performance when utilizing specific DR techniques?

RQ3: How do forecasting outcomes contribute to investment decision making for hedge fund managers in terms of explainability and meaningful insights?

By evaluating the performance of nonlinear DR in forecasting HF returns, this paper aims to address the broader investment community interested in the intersection of ML and financial modeling.

The rest of this paper is organized as follows. In Section II we outline the theoretical background of DR techniques, highlighting their benefits and drawbacks. In Section III we cover the problem setup, dataset analysis, and experimental design. Main results and discussion appear in Section IV.

II. Theoretical Background

Due to the increasing challenges set by enormous amounts of financial data, dimensionality reduction is one of the necessary steps during data preprocessing. When it comes to various technical and fundamental analysis indicators, information from social networks, as well as numerous historical data, this is a significant technique for improving the performance of algorithms, reducing model training, eliminating noise and irrelevant data for specific problems.

A. Principal Component Analysis

PCA is the widely used linear orthogonal DR technique with main goal to identify the principal components (PCs) that capture the maximum variance of data [13]. During transformation, the first PC represents a linear combination of the original features and explains the biggest amount of variance. Each subsequent component is performed with the next highest score of variances.

The main advantage of PCA is its speed and ability to efficiently process large datasets, as well as a clear mathematical interpretation of the variance of the data [13]. However, the principle of linearity limits its ability to capture nonlinear relationships that are present in many real-life datasets. Hence, this inability to model nonlinearity drives the need for the development of more advanced, nonlinear dimensionality reduction techniques, such as KPCA [14].

KPCA is a nonlinear approach of generalizing PCA. Using the kernel method, KPCA first maps data to a higher dimensional space [14]:

$$\mathbb{R}^d \rightarrow \mathbb{R}^D, \ where \ D \gg d \tag{1}$$

Once the kernel method has been applied to transform data into a linearly separable form, PCA is used to dimensionality reduction [5]. KPCA uses several kernel functions like RBF kernel, polynomial and sigmoid. The choice of the kernel function plays an important role, since the performance of the methods depends on the choice.

B. t-Distributed Stochastic Neighbor Embedding

t-SNE is a nonlinear technique of DR with primary purpose to visualize high dimension datasets into low dimension spaces, usually two or three-dimension spaces [15]. Furthermore, the main aim of t-SNE is to preserve complex relationships between original data.

Firstly, t-SNE applies SNE to the dataset which assigns a higher probability to similar pairs of high-dimensional objects, and a lower probability to different data points [5]:

$$p_{a|b} = \frac{exp\frac{-\|x_b - x_a\|^2}{2\sigma^2}}{\sum_{a \neq k}\frac{-\|x_k - x_a\|^2}{2\sigma^2}} \tag{2}$$

where $p_{a|b}$ is conditional probability, x_a high-dimensional data, x_b low-dimensional data and σ^2 given variance. This is achieved with a Gaussian kernel, which represents the similarity between data points as a conditional probability. Thereafter, t-SNE defined SNE in a low-dimension map using the Student's t-distribution to avoid the crowding problem.

Even though this method is widely recognized for its excellence in visualizing and preserving local structure, t-SNE produces high computational costs and time complexity, which is limited for large datasets.

C. Uniform Manifold Approximation and Projection

UMAP is one more nonlinear method for DR based on advanced mathematical concepts, manifold theory and topological data analysis. UMAP stands out in preserving local and global data structure, which is a significant advantage over t-SNE, which focuses on local structure [16]. At the same time, UMAP overcomes the limitation of t-SNE such as execution speed and feature limitation [17].

The algorithm consists of two key steps, graph construction for the high dimensional space and optimization of low dimensional graph layout. In the first step, UMAP constructs a probability distributions graph of nearest neighbors, while the optimization step implies stochastic gradient descent on individual observations [17]. Precisely, initial construction of the graph is crucial in ensuring preservation of both local and global structure. Moreover, two hyperparameters are key for embedding results and control of structure preservation, the number of nearest neighbors that the algorithm takes and the minimum distance between data points in a low-dimensional space [17].

D. Autoencoders

An autoencoder is an unsupervised NN architecture designed to compress (encode) input data into its essential attributes and then reconstruct (decode) the original input from this compressed representation [18]. Therefore, this NN consists of three components: encoder, code and decoder [19]. The encoder is defined by the function that reduces the input data into projection within latent space, while the decoder is defined by the function that decodes the projection into its reconstruction [19]. The code part incorporates the projection as the output from the encoder to the decoder. The output is the reconstruction of the input data [18].

Training of autoencoders aims to minimize the reconstruction loss which measures the difference between the decoder's reconstruction and the original input [19]. This loss function is used to optimize model weights via gradient descent during backpropagation. Furthermore, the performance of autoencoder depends on the tuning of the hyperparameters.

One of the main advantages of autoencoders is the denoising in the data as well as the detection of anomalies and adaptation to changing distributions. However, their implementation requires careful consideration of resources, given the computational complexity.

E. Comparative analysis of dimensionality reduction methods

For easier comparison of techniques, a comparative analysis is provided in Table I. The table summarizes all relevant aspects of applied DR methods, including goal, explainability, computational complexity, linearity, topology, strengths and weaknesses. This review allows a clear insight into the limitations of each method, as well as advantages depending on the specific requirements.

III. Problem Statement and Methodology

A. The problem setup

Our specific case study addresses a problem of forecasting HFs' returns using complex and broad feature space. For investment managers, both predictive accuracy and understanding of key drivers are important for informed investment decision making as they enable effective risk management and portfolio optimization. Given the inherent complexity of financial data, dimensionality reduction becomes a necessary step to simplify data, remove noise and redundancy, and adapt data for building robust and scalable forecasting models for HFs investment returns.

Thus, we address the research questions Q1-Q3 (section 1) and analyze the effectiveness of advanced nonlinear DR techniques in predictive ML models' performance i.e. accuracy and interpretability. The case study includes the analysis of HFs returns for five various investment strategies that suit different investment styles and risk tolerance, as to provide credible and general insights.

B. Datasets

As a follow-up, we further analyze the five datasets from our previous research in [12].

Similarly, this case study is intended to analyze and predict monthly returns of five hedge funds strategies from Morning CISDM Database: Commodity Trading Advisors (CTA), Equity Long Short (ELS), Equity Market Neutral (EMS), Fixed Income Arbitrage (FIA), and Global Macro (GM) strategies. Each strategy suits a different investment style and risk tolerance, allowing traders to choose approaches that align with their financial goals, time horizons, and market outlooks. CTA mostly uses managed futures contracts and trend-following approaches to forecast commodity price. ELS and EMN strategies primarily focus on analyzing and trading companies' stocks listed on stock exchanges. ELS typically involves taking long positions in undervalued stocks and short positions in overvalued ones, while also utilizing options to hedge risks and leverage to enhance potential returns. EMN aims to maintain minimal correlation with the broader equity market by hedging against factors such as currency fluctuations, sector exposure, and market volatility. FIA targets undervalued debt securities of traded companies, such as bonds and other fixed income products. To mitigate risks associated with high-yield positions, FIA may also involve equity positions in the issuing firms as a hedging mechanism. Finally, the GM strategy is based on analyzing macroeconomic trends in countries or regions, using instruments like stocks, bonds, and commodities, and managing risk through derivatives.

TABLE I.
DEFINING CHARACTERISTICS OF FIVE DIMENSIONALITY REDUCTION METHODS

Method	Goal	Explainability	Computational Complexity	Linearity	Topology	Strength	Weakness
PCA	Maximize variance	Medium	$O(d^2n+n^3)$	Linear	Random projection	Computational efficiency	Linear projection
KPCA	Linearly separate data	Low	$O(n^3)$	Nonlinear	Manifold	Capturing nonlinear patterns	High training time
t-SNE	Preserve local structure	Low	$O(n^2)$	Nonlinear	Manifold	Exceptional visualization	Provide only 2 to 3 features
UMAP	Preserve local and global structure	Low	$O(n^{1.14})$	Nonlinear	Manifold	Strong mathematical foundations	Finding the spurious structure of the multiplicity
Autoencoder	Effective compression	Low	Depends on the structure	Nonlinear	Manifold	Denoising	Training data required

Following the recommendations of financial experts and literature review, for each HF strategy we include data on four groups of factors describing trends, debt, capital market conditions, and macroeconomic aspects. Several factors were selected from each group:

- **Interest rates** are represented by a 3-month treasury bill, 5-year constant maturity, and 10-year constant maturity rates, minus federal funds rates.
- **Credit rating categories** include investment grade, high yield below investment grade, and high yield below investment grade.
- **Elements of Fama-French 5-factor model** [20]: NYSE, AMEX, and NASDAQ stock exchanges minus federal funds rate, average return on the small capitalization stock portfolios minus the average return on the large stock portfolios, average return on the value portfolios minus the average return on the growth portfolios, average return on the robust operating profitability portfolios minus the aver-age return on the weak operating profitability portfolios, and average return on the conservative investment portfolios minus the average return on the aggressive investment portfolios.
- **Primitive trend following strategies** on bonds, commodities, interest rates and stocks.

The data was collected monthly over a period of 25 years, from January 1999 to December 2023. For each HF, the final dataset consists of 15 features and 300 instances in total.

C. Data preprocessing and experimental setup

As part of the data preprocessing, the five HFs returns series were divided into training set which include the first 22.5 years, and test set, i.e. the last 2.5 years. Since the data are on different scales, we applied min-max normalization as the preprocessing step. The data also shows peaks and drawdowns, such as “dotcom” stock market bubble in the 2000s, he global economic crisis from 2007 to 2009, and Covid-19 in 2019 [12]. There is also evidence of strong correlations between factors within interest rate group and credit rating categories, as well as performance of conservative against aggressive portfolios and performance of value against growth portfolios.

To address the research questions, we aim to identify a ML prediction algorithm that performs best on transformed data. The forecasting algorithms are implemented in Python, using default hyperparameter values to isolate the influence of DR techniques. The selected algorithms range from simple linear models, such as linear SVM (epsilon value 0.1), to more complex ones, NN with two hidden layers with 64 and 32 neurons and a RELU activation function, as default hyperparameters.

TABLE II.
PERFORMANCE OF PREDICTION MODELS EVALUATED USING RRMSE

		CTA		**ELS**		**EMN**		**FIA**		**GM**	
		STD	**RRMSE**	**STD**	**RRMSE**	**STD**	**RRMSE**	**STD**	**RRMSE**	**STD**	**RRMSE**
Non-reduced	SVM - linear kernel	0.0510	0.0828	0.2661	0.4504	0.0415	0.0675	0.2346	0.4513	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1950	0.3397	0.0415	0.0675	0.0229	0.3778	0.0698	0.1513
	NN	0.0847	0.1086	0.2051	0.3261	0.0615	0.1078	0.1825	0.2813	0.1073	0.1812
	XGB	0.0744	0.1035	0.1349	0.2742	0.0351	0.0684	0.4164	0.6141	0.0654	0.1290
PCA	SVM - linear kernel	0.0510	0.0828	0.1887	0.3789	0.0415	0.0675	0.2720	0.4455	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1176	**0.1849**	0.0415	0.0675	0.2006	0.4124	0.0698	0.1513
	NN	0.0554	0.0840	0.1933	0.3248	0.0574	0.0891	0.1637	0.3025	0.0815	0.1394
	XGB	0.0546	0.0892	0.1725	0.3361	0.0491	0.0794	0.1390	0.2489	0.0904	0.1524
KPCA	SVM - linear kernel	0.0510	0.0828	0.1856	0.3637	0.0415	0.0675	0.3028	0.4773	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1202	0.2004	0.0415	0.0675	0.1775	0.3687	0.0698	0.1513
	NN	0.0514	0.0825	0.1730	0.2749	0.0493	0.0740	0.1504	0.2390	0.0675	0.1351
	XGB	0.0618	0.0969	0.1651	0.3031	0.0438	0.0731	0.1504	0.2609	0.0779	0.1482
t-SNE	SVM - linear kernel	0.0510	0.0828	0.1793	0.2877	0.0415	0.0675	0.1676	0.3209	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1917	0.3543	0.0415	0.0675	0.2020	0.3986	0.0698	0.1513
	NN	0.1326	0.2006	0.6024	0.8879	0.2847	0.4307	0.7113	1.0910	0.3966	0.5741
	XGB	0.0752	0.1086	0.1844	0.2958	0.0372	0.0683	0.1496	0.2353	0.0723	0.1297
UMAP	SVM - linear kernel	0.0510	0.0828	0.1839	0.3027	0.0415	0.0675	0.3211	0.5134	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1723	0.2920	0.0415	0.0675	0.2396	0.4991	0.0698	0.1513
	NN	0.0958	0.1981	0.3228	0.7567	0.1823	0.3648	0.3520	0.8028	0.2826	0.5249
	XGB	0.0745	0.1072	0.2449	0.3906	0.0335	0.0699	0.1471	0.2409	0.1059	0.1680
Autoencoder	SVM - linear kernel	0.0510	0.0828	0.1770	0.2933	0.0415	0.0675	0.2245	0.3694	0.0698	0.1513
	SVM - RBF kernel	0.0510	0.0828	0.1494	0.2732	0.0415	0.0675	0.2395	0.4932	0.0698	0.1513
	NN	0.0573	**0.0822**	0.1522	0.2300	0.0299	**0.0571**	0.1246	**0.2071**	0.0645	**0.1078**
	XGB	0.0635	0.0932	0.2678	0.4082	0.0472	0.0860	0.1905	0.2904	0.0998	0.1717

We also employ nonlinear SVM (RBF kernel and epsilon value 0.1) to account for the nonlinear nature of financial data and to explore how a nonlinear feature extraction technique aligns with a relatively simple nonlinear predictor. Finally, we utilize XGB (100 decision trees with a maximum depth of 6), as a representative of ensemble learning methods particularly effective in prediction tasks.

The success of the proposed methods will be measured using standard evaluation metrics for regression: mean absolute error (MAE) and relative root mean squared error (RRMSE). RRMSE is normalized error using the average of actual values and is often expressed as a percentage.

IV. Experimental results

The effectiveness of DR methods is validated by comparing predicted values with actual hedge fund returns across all five HF strategies. The performance of ML predictive models, using RRMSE and MAE, is shown in Tables II and III.

As expected, more sophisticated techniques, such as autoencoders, deliver better performance in all cases except for the ELS strategy. This exception may stem from the Fama-French asset pricing factors already being carefully selected by domain experts. From an investment perspective, the analysis of individual principal component loadings indicates that factors such as investment-grade (BAA), high-yield below investment grade (BBB), and 5-year and 10-year constant maturity rates have a significant impact on most strategies. This is also expected, as these are key drivers of performance across various industries and sectors.

Furthermore, the empirical results are used to address main research questions. The prediction results of HFs return series confirm that dimensionality reduction, both linear and nonlinear, in most cases, positively affects the predictive capacity of ML forecasting models. However, nonlinear DR and nonlinear ML predictors together potentially generate more noise in the data e.g. combination of t-SNE/UMAP with NN increases error.

In general, linear PCA provides equal or better prediction results compared to non-reduced input datasets. It was expected that adding nonlinear NN and XGB predictors to non-reduced and/or PCA-reduced data will significantly reduce error. However, that was not the case for most strategies, even though linear DR did contribute to reducing error in general. This indicates the potential presence of nonlinear dependence in the original data which required analysis of more complex DR algorithms.

Among the tested nonlinear DR methods, t-SNE and UMAP resulted in higher prediction errors across strategies

TABLE III.
PERFORMANCE OF PREDICTION MODELS EVALUATED USING MAE

		CTA		**ELS**		**EMN**		**FIA**		**GM**	
		STD	**MAE**	**STD**	**MAE**	**STD**	**MAE**	**STD**	**MAE**	**STD**	**MAE**
Non-reduced	SVM - linear kernel	0.0294	0.0375	0.0434	0.0593	0.0120	0.0154	0.0296	0.0486	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0318	0.0454	0.0120	0.0154	0.0289	0.0378	0.0147	0.0283
	NN	0.0488	0.0391	0.0335	0.0414	0.0177	0.0256	0.0230	0.0270	0.0227	0.0308
	XGB	0.0428	0.0414	0.0220	0.0390	0.0101	0.0170	0.0525	0.0569	0.0138	0.0235
PCA	SVM - linear kernel	0.0294	0.0375	0.0308	0.0537	0.0120	0.0154	0.0343	0.0445	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0192	**0.0233**	0.0120	0.0154	0.0253	0.0454	0.0147	0.0283
	NN	0.0319	0.0363	0.0316	0.0426	0.0166	0.0197	0.0206	0.0321	0.0172	0.0239
	XGB	0.0314	0.0406	0.0282	0.0471	0.0142	0.0180	0.0175	0.0260	0.0191	0.0259
KPCA	SVM - linear kernel	0.0294	0.0375	0.0303	0.0511	0.0120	0.0154	0.0382	0.0465	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0196	0.0262	0.0120	0.0154	0.0224	0.0407	0.0147	0.0283
	NN	0.0296	0.0371	0.0282	0.0349	0.0142	0.0159	0.0190	0.0234	0.0143	0.0247
	XGB	0.0355	0.0429	0.0270	0.0415	0.0126	0.0169	0.0190	0.0269	0.0164	0.0266
t-SNE	SVM - linear kernel	0.0294	0.0375	0.0293	0.0367	0.0120	0.0154	0.0211	0.0345	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0313	0.0487	0.0120	0.0154	0.0255	0.0433	0.0147	0.0283
	NN	0.0763	0.0866	0.0984	0.1065	0.0822	0.0933	0.0897	0.1043	0.0837	0.0876
	XGB	0.0432	0.0451	0.0301	0.0378	0.0107	0.0165	0.0189	0.0229	0.0153	0.0227
UMAP	SVM - linear kernel	0.0294	0.0375	0.0300	0.0393	0.0120	0.0154	0.0405	0.0505	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0281	0.0385	0.0120	0.0154	0.0302	0.0552	0.0147	0.0283
	NN	0.0551	0.0997	0.0527	0.1117	0.0526	0.0912	0.0444	0.0909	0.0597	0.0934
	XGB	0.0429	0.0443	0.0400	0.0497	0.0097	0.0177	0.0185	0.0240	0.0224	0.0275
Autoencoder	SVM - linear kernel	0.0294	0.0375	0.0289	0.0382	0.0120	0.0154	0.0283	0.0370	0.0147	0.0283
	SVM - RBF kernel	0.0294	0.0375	0.0244	0.0374	0.0120	0.0154	0.0302	0.0544	0.0147	0.0283
	NN	0.0330	**0.0339**	0.0248	0.0282	0.0086	**0.0140**	0.0157	**0.0209**	0.0136	**0.0182**
	XGB	0.0366	0.0392	0.0437	0.0503	0.0136	0.0207	0.0240	0.0276	0.0211	0.0295

(e.g., GM and FIA), indicating its limitations for this forecasting task. On the other hand, autoencoders are proven to contribute to predictive power to most ML models, and for all analyzed HF strategies. For all HF strategies, the best performance is achieved with two nonlinear DR techniques, KPCA and autoencoders, combined with NN forecasting model.

When considering algorithm characteristics, KPCA is computationally intensive and sensitive to high dimensionality, while autoencoders are better suited for processing larger datasets. From the perspective of HF managers, improved forecasting helps reduce performance uncertainty to some extent, supporting more informed decision-making when selecting among diverse hedge fund strategies. However, both approaches lack explainability, making it challenging for managers to justify their decisions.

V. Conclusion

This paper extends the research initiated in [12] on identifying and extracting key features in financial time series forecasting. In this study, the influence of various nonlinear DR methods was analyzed for HF performance forecasting. An evaluation of effectiveness of KPCA, t-SNE, UMAP and autoencoders was examined on real-world data. Reduced data serves as inputs to SVM with linear and nonlinear kernel, NN and XGB as predominant ML predictors for financial time series. To provide credible conclusions, we utilized HF returns data that represent different investment styles like CTA, ELS, EMN, FIA, and GM [12]. The four group of factors (15 features in total) covering interest rate, credit rating, trend and Fama-French model were included into analysis, over 25 years period.

The empirical findings address the primary research questions. The analysis shows that both linear and nonlinear DR contribute to ML forecasting results for diverse HF strategies. Linear PCA was not effective enough, even when combined with nonlinear ML predictors. The best performance was achieved with nonlinear DR, KPCA and autoencoders when paired with neural networks. Other critical DR characteristics were discussed for the most effective combinations. From the point of view of HF managers, the results provide better forecasting accuracy and contribute to more informed investment decision making. Besides the necessity for domain ML knowledge (e.g. hyper-parameters), the lack of explainability was identified as the main drawback, as expected.

Our future work will be oriented towards the development and application of investment-based DR methods that offer explainability as a baseline for effective decision-making (e.g. [12]). Also, investment metrics will be included to validate investment performance.

Acknowledgment

This study was supported by the University of Belgrade Faculty of Organizational Sciences.

References

[1] HFR Industry Reports. [Online]. Available: www.hfr.com/hfr-industry-reports/

[2] W. Wu, J. Chen, Z. Yang, and M. L. Tindall, "A cross-sectional machine learning approach for hedge fund return prediction and selection," *Manag. Sci.*, vol. 67, no. 7, pp. 4577–4601, 2021. https://doi.org/10.1287/mnsc.2020.3696

[3] M. Ashraf *et al*., "A survey on dimensionality reduction techniques for time-series data," *IEEE Access*, vol. 11, pp. 42909–42923, 2023, doi: 10.1109/ACCESS.2023.3269693.

[4] R. Zaheer, M. K. Hanif, M. U. Sarwar, and R. Talib, "Evaluating the effectiveness of dimensionality reduction on machine learning algorithms in time series forecasting," *IEEE Access*, 2025. https://doi.org/10.1109/ACCESS.2025.3551741

[5] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, 100378, 2021. https://doi.org/10.1016/j.cosrev.2021.100378

[6] IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10929010

[7] A. M. Lopes and J. A. T. Machado, "Dynamical analysis of the Dow Jones index using dimensionality reduction and visualization," *Entropy*, vol. 23, no. 5, p. 600, 2021. https://doi.org/10.3390/e23050600

[8] Z. Wang, H. Chen, X. Yang, J. Wan, T. Li, and C. Luo, "Fuzzy rough dimensionality reduction: a feature set partition-based approach," *Information Sciences*, vol. 644, Art. no. 119266, 2023.

[9] D. Kumar, P. K. Sarangi, and R. Verma, "A systematic review of stock market prediction using machine learning and statistical techniques," *Mater. Today Proc.*, vol. 49, pp. 3187–3191, 2022. https://doi.org/10.1016/j.matpr.2020.11.399

[10] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Rev. Financ. Stud.*, vol. 33, no. 5, pp. 2223–2273, 2020. https://doi.org/10.1093/rfs/hhaa009

[11] C. Lin, "Key financial indicators analysis and stock trend forecasting based on a wrapper feature selection method," in *2024 19th Conf. on Computer Science and Intelligence Systems (FedCSIS)*, Belgrade, Serbia, Sep. 2024, pp. 755–759.

[12] A. Radosavcevic, A. Poledica, and I. Antovic, "Discovery of key factors in hedge funds investment strategies using optimal IBA-based logical polynomials," in *Proc. Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst.*, Cham, Switzerland: Springer, 2024, pp. 335–346. https://doi.org/10.1007/978-3-031-74000-8_28

[13] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *J. Soft Comput. Data Mining*, vol. 2, no. 1, pp. 20–30, 2021. https://doi.org/10.30880/jscdm.2021.02.01.003

[14] L. J. Cao *et al*., "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1–2, pp. 321–336, 2003. https://doi.org/10.1016/S0925-2312(03)00433-8

[15] R. Silva and P. Melo-Pinto, "t-SNE: A study on reducing the dimensionality of hyperspectral data for the regression problem of estimating oenological parameters," *Artif. Intell. Agric.*, vol. 7, pp. 58–68, 2023. https://doi.org/10.1016/j.aiia.2023.02.003

[16] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by UMAP to visualize physical and genetic interactions," *Nat. Commun.*, vol. 11, no. 1, p. 1537, 2020. https://doi.org/10.1038/s41467-020-15351-4

[17] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization," *J. Mach. Learn. Res.*, vol. 22, no. 201, pp. 1–73, 2021.

[18] B. Ghojogh *et al*., "Feature selection and dimensionality reduction in pattern analysis: A literature review," *arXiv preprint arXiv:1905.02845*, 2019. https://doi.org/10.48550/arXiv.1905.02845

[19] Q. Fournier and D. Aloise, "Empirical comparison between autoencoders and traditional dimensionality reduction methods," in *Proc. 2nd Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, pp. 211–214, IEEE, June 2019. https://doi.org/10.1109/AIKE.2019.00044

[20] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *J. Financ. Econ.*, vol. 33, no. 1, pp. 3–56, 1993. https://doi.org/10.1016/0304-405X(93)90023-5

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 417–422
DOI: 10.15439/2025F4362
ISSN 2300-5963 ACSIS, Vol. 43

Encouraging Sustainable Fashion Choices Through a Mobile Application

Bianka Šimková
Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovakia
Email: bianka.simkova.2@student.tuke.sk

***Abstract*—This work presents the design and implementation of a mobile application for home wardrobe management, aimed at promoting sustainability in fashion by encouraging more conscious clothing choices. The theoretical part of the article deals with the concept of "fast fashion" and its impact on different areas of the fashion industry and defines the problem with existing apps in this field. The practical section includes the design of an information system based on a user preferences survey, and includes a visualization of the system architecture. This section also describes the implementation process of two parts of the system, namely the creation of a clothing classification application using the CRISP-DM methodology and the development of a mobile app for the Android operating system, as well as the testing process. The conclusion evaluates the outcomes of the project and outlines potential directions for future enhancements.**

I. Introduction

THE motivation for writing this article is the ever-growing fashion industry, which offers us an overabundance of clothes at very low prices. This results in a excessive number of garments in the wardrobe of an average individual and clutter in the wardrobe. Therefore, the aim of this article is to design and implement a home wardrobe management information system that will contribute to sustainability in fashion.

Consumerism still dominates the world today, and much of this is influenced by the fashion industry. The term "fast fashion" emerged in the 1990s, when retailers began offering clothes to the masses at low prices, but at the expense of quality, in order to keep up with the pace of changing trends and emulate more luxurious garments [1]. However, the large volumes of clothing produced lead to increased resources for disposal and consequently high levels of waste, which has a negative impact on the environment [2]. Extending the lifetime of products is one of the most effective environmental strategies as it has a great potential to prevent waste and reduce production, consumption (both in sourcing and disposal) and transportation. This has been conceptually known, documented, and recognised for a long time - for example, in the waste management hierarchy, where prevention is the highest priority [3]. This is also why terms such as sustainability and "slow fashion" have been frequently mentioned in recent years.

A recent study based on a sample of 2,000 respondents found that the average individual owns approximately 136 items of clothing, indicating a relatively high volume of wardrobe content [4]. Often, we may not have an overview of individual items of clothing, nor how many times a piece of clothing has been used and how long it has been in our wardrobe. This factor, combined with the ongoing marketing strategies employed by retailers, can result in impulsive purchases of clothing items that are redundant, seldom worn, or do not fit properly. This inspired the idea of developing a mobile application — a wardrobe database — that can categorize garments based on photos and assess whether similar items are already present in the wardrobe or were owned in the past, along with data on their frequency of use. This feature can assist during shopping by discouraging unnecessary purchases, helping to keep the closet more organized and less cluttered.

II. Related Work

According to Klepp et al. [3], the lifetime of garments can be described in a way that makes them comparable, using the following categories:

- Number of wears;
- Years (Duration of use);
- Number of users (owners);
- Number of cleaning cycles.

Another important aspect in the care of garments is the way they are cleaned, which in addition to having an impact on the lifetime of the garment, also has an impact on the environment. It has been shown that the correct washing machine setting for garments made of polyester will reduce the release of microplastics by 12 to 16% [6]. Accordingly, effective garment care also requires attention to the material's origin, particularly the fibre composition of the fabric, such as:

- Synthetic fibres - such as polyester, nylon or acrylic.
- Natural fibres - such as cotton, linen or wool.

Also, washing synthetic materials with hard surfaces, such as jeans, increases the risk of damaging the material. When washing synthetic fabrics, it is also advisable to use less detergent and avoid bleach. Finally, it is recommended not to use a tumble dryer to dry synthetic materials, as heat and mechanical forces can cause the release of microplastic fibres. Last but not least, a person's preferences in choosing and wearing clothes, including on the basis of hair colour, complexion, eyes and body type, are also a major factor.

As Acerbi et al. [5] argue, fashion is not shaped by random decisions; rather, individuals exhibit strong preferences for

specific cultural traits and tend to emulate certain styles and patterns. Fast fashion capitalizes on this behavior by employing innovative production and distribution strategies to drastically reduce the time required to bring garments from design to retail — often within weeks rather than months. As a result, the number of fashion seasons has expanded from the traditional two per year (spring/summer and autumn/winter) to as many as 50 to 100 micro-seasons annually [9]. This acceleration contributes to the continual shortening of the fashion cycle — the lifespan of a fashion trend — which is typically divided into five phases: introduction, rise, peak, decline, and obsolescence [10].

Many authors consider the rise of social media as the main cause of shorter and less predictable fashion cycles, as "trend drivers" [7]. A 2015 study confirmed that there is a direct and significant relationship between social media and the fashion industry [8]. There are some existing apps that focus on creating personal virtual wardrobe, such as Open-Wardrobe Outfit Planner ++[1], Acloset - AI Fashion Assistant[2], Whering:Digital Closet Stylist[3], or Style DNA: Fashion AI Stylist [4]. However, the applications are rather commercially tuned, i.e. built on the basis of recommendations of similar products that the user already owns. This consequently leads to unconsidered purchases and an excessive volume of clothing items, as mentioned in the introduction. Therefore, it would be advisable for the application to be able to assess, based on a photography of the garment, whether or not the user already owns a similar item and its frequency of use. Furthermore, existing apps either lack garment usage statistics or provide them in a rather superficial manner.

III. Proposed App Design

Initially, the requirements for the app were derived from survey responses published across various platforms. The survey was created with the intent to give us an insight into people's attitudes and behaviours related to fashion and clothes shopping. It consisted of 7 questions on demographic data, 8 statements related to personal preferences that respondents rated based on a Likert scale, and 1 question to select from the options of the preferred requirements of the wardrobe database application.[5] In total, 18 functional requirements were implemented in the app, namely: adding garments to your wardrobe; filtering and sorting garments by type, colour, pattern, material, recommended care, colour, number of wears, purchase price and date of purchase; adding care information to garments; recording garment washes; creating and saving an outfit; viewing saved outfits; planning an outfit; filtering saved outfits based on occasion and season; viewing wardrobe usage statistics; viewing usage statistics for a specific garment; automatic classification of garments by photo into type and colour; offering and accepting an outfit for exchange; adding personal preferences for style of clothing; estimating whether an outfit fits in the wardrobe. From the functional requirements, a list of ten use cases was created and made along with a UML diagram of the relationships between them to visualize more complex functionalities of the system that users expect and are of value to them.

Furthermore, four non-functional requirements were defined, focusing on security, performance, and usability. These included: mobile app response time of less than two seconds; support for at least 100 users in the database; encryption of stored data; and a simple, user-friendly graphical user interface (GUI).

The application consists of a 3-tier architecture as shown in Fig.1. The presentation layer consists of a user interface in the form of a xml layout and fragments. The business layer consists of an application core written in the Java programming language and a clothing classification API written in the Python programming language and deployed on Microsoft Azure. The data tier is managed by the Firebase platform. Attributes of garments and outfits, as well as the user's personal preferences, are stored in the Cloud Firestore database, and photos of garments and outfits are stored in Cloud Storage.

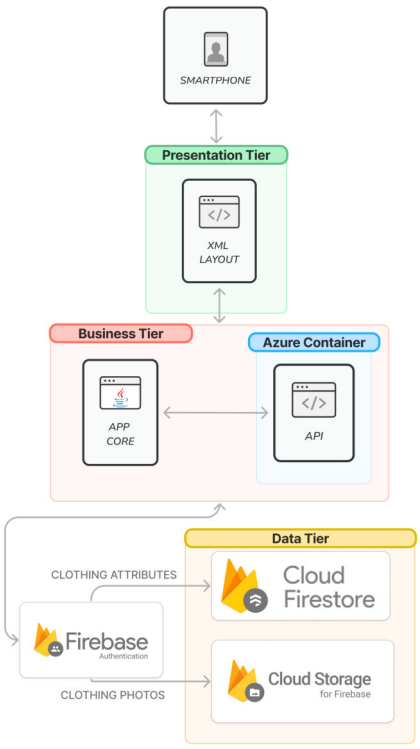


Fig. 1. App Architecture.

[1]https://play.google.com/store/apps/details?id=com.open_wardrobe_mobile
[2]https://play.google.com/store/apps/details?id=com.looko.acloset
[3]https://play.google.com/store/apps/details?id=com.whering.app
[4]https://play.google.com/store/apps/details?id=style.dna.app&hl=sk
[5]Link to the survey: https://docs.google.com/forms/d/1sv9lIR2Uj6OpbMFe94TsRsnGgMJTwuWdhJSrItx7oZ0/

IV. Implementation

A. Methodology

The development took approximately 6 months. In the first part of the implementation, we worked on creating an API for clothing classification using the Python programming language. In the second part, we programmed an Android application in Java programming language and integrated the developed API into it. In modelling the SVM classifier, we followed the steps of the CRISP-DM methodology, which consists of understanding the objective, understanding the data, data preparation, modelling, evaluation and deployment.

B. Clothing Classifier

For training the SVM classifier we utilized the dataset *Clothing dataset (full, high resolution),*[6] which contains over 5,000 photos of clothing items in both original and reduced quality. An analysis of the image orientations — portrait versus landscape — revealed a significantly higher frequency of portrait-format photos, suggesting that the average user tends to capture clothing images in portrait mode. The standard for the aspect ratio of photos taken with a smartphone is 4:3, so we resized all these photos to 160x120 pixels, and converted them to an array of RGB pixel codes. There are 20 categories of clothing, but the number of categories varied considerably, which would have caused us problems in modelling, so we reduced these categories to 8 main ones: short-sleeved top, long-sleeved top, pants, dress, outerwear, shorts, skirt and shoes. After merging and reducing the categories, the frequencies are more balanced.

We used 80% of our preprocessed category and pixel dataset to train the model and 20% to test the model, maintaining the ratio of category frequencies. For training the SVM classifier, we chose the Gaussian Radial Basis Function (RBF) kernel function and the decision function shape One vs. One. We evaluated the model with the metrics of accuracy, precision recall and F1 score. The classification accuracy is 0.74, which provides a good basis for deploying the model in practice. Precision, recall, and F1 score are above 0.7 for most classes. For categorizing and calculating similarity between colours, we used the webcolours library.[7]

C. API Containerization

The SVM classifier was incorporated in an application to process clothing photography using FastAPI, a modern web framework for creating APIs using Python. The application provides two main endpoints that process prediction and classification requests and return JSON responses:

- classify - takes as the image of the garment as an input parameter and outputs the type and colour of the garment;
- predict - takes as the image of the garment, and the name of the user's colour palette, as an input parameters and outputs the type and colour of the garment, and a boolean value that tells whether the colour is in the user's predefined colour palette.

We containerized the application using Docker, with Gunicorn serving the FastAPI app through Uvicorn workers. The resulting Docker image was uploaded to Docker Hub and deployed to Microsoft Azure as a continuously running container.

D. Mobile App

The mobile app was developed in Android Studio in the Java programming language. The frontend of the final version of the mobile application consists of 23 fragments. The bottom of the screen features a navigation menu that is used to switch between the five main modules of the application.

1) Closet: This module can be accessed by the middle button in the bottom menu and will show us the wardrobe contents with options to filter, add new garments, view statistics, and view garments marked for washing. In the process of adding a garment by the "+" button, the aforementioned classifier is also used, which classifies the garment into type and colour. After adding a garment, the detail of the stored garment is displayed, which also contains the statistics of the particular garment (Fig. 2), which are automatically recalculated based on the planning of outfits - sets of garments. Additionally, the user can also access the aggregated statistics for the whole closet (Fig. 3). The "Filter" button will show the options to filter and sort the garments, and after confirming the selection, it will show us the closet with only the garments that meet the required characteristics.

2) Outfits: The first button in the bottom menu will launch the module related to storing, viewing, filtering, and planning outfits. The ability to filter and sort outfits works the same way as for clothing. When adding, all the outfits stored in the wardrobe are displayed, and the user selects the garments that make up the outfit. Scheduled outfits can be viewed by the user in the "OOTD" (Outfit Of The Day) section.

3) Settings: This module is accessed by the far-right button of the bottom menu and displays the login screen or, if the user is logged in, the option to log out and set personal preferences.

4) DoesItFit: The fourth button in the bottom menu gives us the ability to view the saved preferences and also the ability to estimate the fit to the wardrobe, where the user uploads a photo of the garment and the application here uses the classifier to classify the garment into type and colour, verifying that the colour of the garment is within the user's chosen colour palette and displaying similar garments in the wardrobe, i.e. of the same type and colour (Fig. 4).

5) Swap: By clicking on the second button in the bottom menu, we can see the garments offered for exchange. Clicking on a particular garment will show the details and the option to request the garment. Then the user who provided the garment for exchange can see in the details of their garment the email addresses of the users who requested the garment, and after selecting and confirming the selection of one of the email addresses, the garment will be reassigned to that user.

The backend of the application is written in the Java programming language and consists of 53 classes, which

[6]https://www.kaggle.com/datasets/agrigorev/clothing-dataset-full/code
[7]https://webcolours.readthedocs.io/en/latest/



Fig. 2. Clothing item detail.

are divided into five packages: `activities` (as mentioned previously in this section); `api` (classes associated with linking the application to the classifier); `entities` (constructors and methods of `ClothingItem`, `Outfit`, `SwapItem` and `Webcolours` objects); `recycler-view` (classes that modify the view of entities); `tools` (methods for working with the database, data visualization, image editing, calculating statistics, etc.). The remaining two classes are subclasses of the `Activity` class: `WelcomeActivity` and `MainActivity` that runs throughout the application and is responsible for controlling the bottom menu to change fragments.

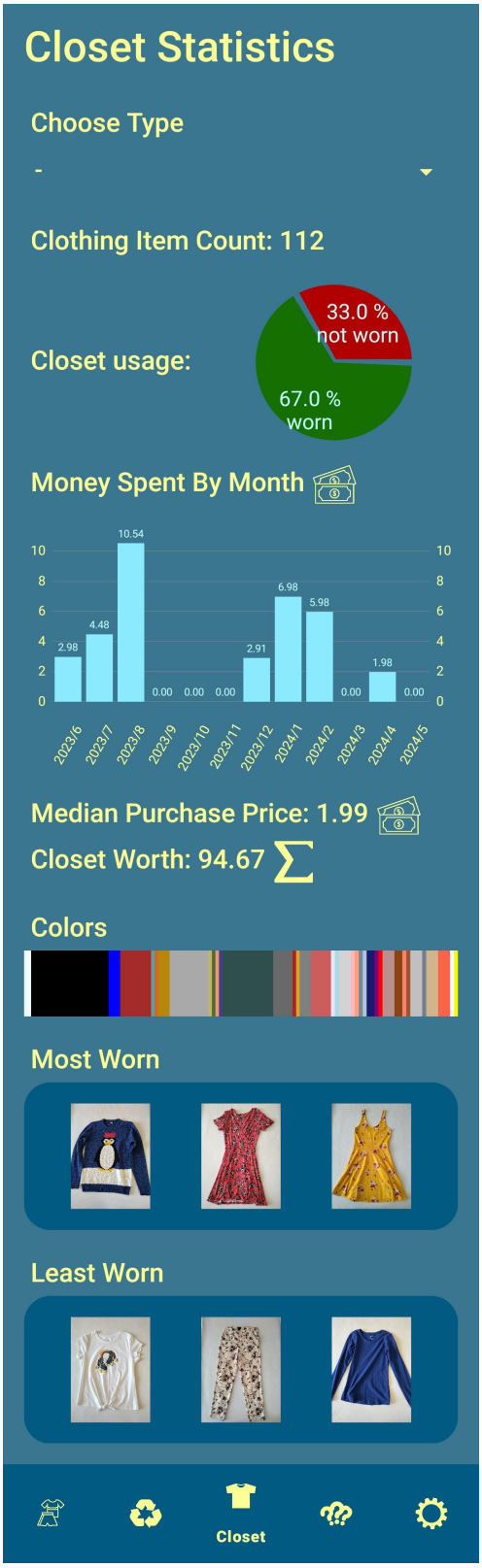


Fig. 3. Closet statistics.



Fig. 4. Application output on finding similar garments based on photo.

E. Database

We used the Firebase Firestore database to store attributes related to clothing items and outfits. The "user_clothes" collection contains documents named according to the unique user ID (UID) generated during registration via Firebase Authentication. Each of these documents includes two subcollections: "clothes" and "outfits", which contain documents named according to the IDs of the outfits. Each document captures 16 features of the corresponding item, including statistical attributes such as `timesWorn`, `timesWashed` and `wearPercentile`. The methods used for calculating statistics are contained in the `StatisticsCalculator` class, which includes, for example, a method to calculate garment usage frequency based on percentiles as shown in Algorithm 1 and 2. The underlying logic for these statistical measures is inspired by the concepts discussed in the article by Klepp et al. [3], as referenced in the introduction.

Algorithm 1 Calculate wear percentiles of clothing items

Require: List of clothing items with `timesWorn`
Ensure: List of wear percentiles (as integers from 0 to 100)

Initialize empty list $percentiles$
for all $calculatedItem \in clothingItems$ **do**
  $countLessOrEqual \leftarrow 0$
  for all $item \in clothingItems$ **do**
    if $item \neq calculatedItem$ **and** $item.timesWorn \leq calculatedItem.timesWorn$ **then**
      $countLessOrEqual \leftarrow countLessOrEqual + 1$
    end if
  end for
  Append $countLessOrEqual$ to $percentiles$
end for
$countAll \leftarrow$ size of $clothingItems - 1$
for $i = 0$ to $percentiles.size() - 1$ **do**
  $percentileFloat \leftarrow \frac{percentiles[i]}{countAll} \times 100$
  $percentiles[i] \leftarrow \lfloor percentileFloat \rfloor$
end for
return $percentiles$

Algorithm 2 Get wear frequency label

Require: A clothing item with computed percentile
Ensure: A frequency label (string)

$percentile \leftarrow clothingItem.getWearPercentile()$
if $percentile \leq 25$ **then**
  return "very low"
else if $percentile \leq 50$ **then**
  return "low"
else if $percentile \leq 75$ **then**
  return "medium"
else if $percentile \leq 100$ **then**
  return "high"
else
  return null
end if

V. Evaluation

Functional and integration testing of the application was also carried out in the implementation phase. At the end of the implementation, we evaluated the app using Firebase Test Lab. After evaluation by the Test Lab, we took the results of this testing into account and patched the version. We conducted acceptance testing in person with five target users who were asked to complete 13 tasks such as: use the app to add a garment to the wardrobe; use the app to create an outfit; use the app to view statistics about your wardrobe usage; use the app to check if the garment fits into your wardrobe. Immediately after testing, we had each participant complete a System Usability Scale (SUS) survey. All users rated the application rather positively, which resulted in the SUS score of 93.5. In addition, the survey included a question about which functionalities users found most or least useful. The most useful features were rated by users as the garment care record, viewing statistics, filtering garments, saving outfits, and estimating whether a garment fits them.

Among all features, testers found the clothing swap option to be the least beneficial.

In order to validate the non-functional requirement "mobile app response time of less than two seconds", we created a `StartupAndFrameBenchmark` class in the Android Studio environment, in which we defined tests to detect the duration of the app launch and the duration of the screen rendering after some action, i.e. the app response time. The results of this testing show that 99% of the measured responses were less than or equal to 205.0 milliseconds, which is approximately 0.2 seconds. The remaining non-functional requirements related to the database. Since we store data in Firestore database, we have fulfilled these requirements, since the capacity of Firestore database is 50 000 active users per month, Firebase automatically encrypts data in database. We also enhanced security by setting the Firebase rules so that each logged in user can only read and write data under the document that is identical to their UID or under the "swap" collection as mentioned before.

VI. Conclusion and future work

The aim of this work was to discuss the influence of different factors on fashion and propose a design and implementation of a mobile application with the intent of promoting sustainability in fashion. The implemented app lays the groundwork for encouraging more conscious consumption habits by helping users manage their wardrobes more efficiently and make sustainable fashion choices. The app also allows for possible future extensions, primarily focused on improving the user experience, such as: adding wardrobe sharing; adding automatic outfit suggestion; automatic background removal on photos of clothes; improving the accuracy of the classifier, or using neural networks in classification. The App and API are available on github.[8]

References

[1] A. Joy, J. Sherry Jr., A. Venkatesh, J. Wang, and R. Chan, "Fast fashion, sustainability, and the ethical appeal of luxury brands," *Fashion Theory, vol. 16, no. 3, pp. 273–295, 2012,* https://doi.org/10.2752/175174112X13340749707123.

[2] L. Cavusoglu and M. Dakhli, "The impact of ethical concerns on fashion consumerism: a review," *Markets, Globalization & Development review, vol. 1, no. 2, art. no. 5, 2016,* https://doi.org/10.23860/MGDR-2016-01-02-05.

[3] I. G. Klepp, K. Laitala, and S. Wiedemann, "Clothing lifespans: what should be measured and how," *Sustainability, vol. 12, no. 15, art. no. 6219, 2020,* https://doi.org/10.3390/su12156219.

[4] Search Department Statista, "Average number of apparel items owned by consumers worldwide from 2017 to 2019 [Infographic]," Statista, 2019, https://www.statista.com/statistics/1008465/average-number-of-apparel-items-owned-by-consumers-worldwide/.

[5] A. Acerbi, S. Ghirlanda, and M. Enquist, "The logic of fashion cycles," *PLOS ONE, vol. 7, no. 3, 2012,* https://doi.org/10.1371/journal.pone.0032541.

[6] G. Dalla Fontana, R. Mossotti, and A. Montarsolo, "Assessment of microplastics release from polyester fabrics: The impact of different washing conditions," *Environmental Pollution, vol. 264, 2020,* https://doi.org/10.1016/j.envpol.2020.113960.

[7] G. G. Berk and M. R. DeLong, "Communicating Fashion: Trend Research and Forecasting," University of Minnesota Libraries Publishing, 2023, https://open.lib.umn.edu/communicatingfashion/chapter/chapter-6-trend-lifecycles-and-adopter-categories/.

[8] N. Ahmad, A. Salman, and R. Ashiq, "The impact of social media on fashion industry: Empirical investigation from Karachiites," *Journal of Resources Development and Management, vol. 7, 2015,* https://ssrn.com/abstract=2603275.

[9] D. Drew and G. Yehounme, "The apparel industry's environmental impact in 6 graphics," World Resources Institute, 2017, https://www.wri.org/insights/apparel-industrys-environmental-impact-6-graphics.

[10] M. Pillai, "Understanding cycles of fashion," *International Journal of Advances in Engineering and Management (IJAEM), vol. 2, no. 12, pp. 177–191, 2020,* https://doi.org/10.35629/5252-0212177191.

[8] https://github.com/biankasim/

Thematic Sessions

PARTS 4 and 5 of FedCSIS 2025 Proceedings contain contributions originating from the Thematic Sections. Let us briefly introduce each one of them (in alphabetical order).

I. AI in Agriculture

Artificial Intelligence (AI) is increasingly used in agriculture for a variety of uses, from plant disease detection to weeding automation, soil status monitoring, crop prediction, irrigation management, and decreased use of resources for improving quality and productivity. This Thematic Session welcomes contributions related to a wide variety of interdisciplinary research and applications related to the application of intelligence systems in agriculture. AI can, in fact, provide highly positive effects on precision agriculture by optimizing, automating, and forecasting several aspects of farming and revolutionizing the sector, providing helpful information and driving decisions using multiple sources of data and different sensors.

Moreover, in the climate change era, AI can improve sustainability by optimizing resources such as water and soil management as well as fostering novel methods and approaches. We welcome innovative contributions, early results, and position papers addressing one or more of the topics listed below. We intend to foster informal discussions and bring together researchers, practitioners, and industry experts to explore the challenges and opportunities of intelligence systems and agriculture.

This Thematic Session was organized by:

- Charvat, Karel, Czech Center for Science and Society, Czech Republic
- Martinelli, Massimo, Institute of Information Science and Technologies, National Research Council of Italy, Italy
- Moroni, Davide, Institute of Information Science and Technologies, National Research Council of Italy, Italy
- Procházka, Ales, University of Chemistry and Technology and Czech Technical University CIIRC, Czech Republic

II. AI in Digital Humanities, Computational Social Sciences and Economics Research

This thematic session is dedicated to the computational study of Social Sciences, Economics and Humanities, including all subjects like, for example, education, labour market, history, religious studies, theology, cultural heritage, and informative predictions for decision–making and behavioural-science perspectives. While digital methods, intelligence systems, and AI have been emerging topics in these fields for several decades, this thematic session is not only limited to discoveries in these domains, but also dedicated to the reflections of these methods and results within the field of computer science. Thus, we are in particular interested in interdisciplinary exchange and dissemination with a clear focus on computational and AI methods for intelligence systems.

Since there is a clear methodological overlap between these three domains and often similar algorithms and AI approaches are considered, we see this thematic session as place for interdisciplinary learning, discussing a joint toolbox as a support for scholars from these field with human and context-aware agents. The aim of this thematic session is thus to bridge the gap between scientific domains, foster interdisciplinary exchange and discuss how research questions from other domains challenge current computer science. In particular, we are interested in communications between researchers from different fields of computer science, social sciences, economics, humanities, and practitioners from different fields

This Thematic Session was organized by:

- Cooper, Anthony-Paul, Durham University, United Kingdom and University of Turku, Finland
- Dörpinghaus, Jens, BIBB and University of Koblenz, Germany
- Helmrich, Robert, BIBB and University of Bonn, Germany
- Speckesser, Stefan, Brighton University, United Kingdom
- Tiemann, Michael, BIBB and University of Koblenz, Germany

III. Advances in Programming Languages

Programming languages are the most fundamental tools for programmers. With the right programming languages, one can significantly reduce the cost of building new applications and maintaining existing ones. Over the last few decades, there have been substantial advances in programming language technology within traditional paradigms such as functional, logic, object-oriented, and aspect-oriented programming. Today the growing influence of artificial intelligence on programming paradigms and the development of software systems is rapidly reshaping the field. Above all, programming tools based on generative artificial intelligence have become a standard, transforming programming practices in both academia and industry. The methods and approaches for selecting, adapting, and evaluating AI techniques for programming language design and implementation have become a critical area of research. Within this context, new advances in programming language research are also contributing to the design and implementation of autonomous and intelligent systems, such as those developed under the umbrella of agent-oriented programming. The primary motivation has always been, and will continue to be, to enable a more effective formulation of programmers' ideas. As a result, research in programming languages remains a perpetual activity and lies at the core of computer science. Hence, new language features and programming paradigms—particularly those enhanced by AI techniques—as well as improved compile-time and run-time mechanisms, can be anticipated in the near future.

This Thematic Session was organized by:

- Janousek, Jan, Czech Technical University, Czech Republic
- Rangel Henriques, Pedro, Universidade do Minho, Portugal
- Slivnik, Boštjan, University of Ljubljana, Slovenia
- Varanda Pereira, Maria Joao, Instituto Politecnico de Braganca, Portugal

IV. Computer Aspects of Numerical Algorithms

Numerical algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy–to–use scalable tools for solving large scale problems. This Thematic Session is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation and utilization of libraries of efficient numerical algorithms. The goal is the meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

This Thematic Session was organized by:

- Bylina, Beata, Maria Curie-Skłodowska University, Poland
- Bylina, Jarosław, Maria Curie-Skłodowska University, Poland
- Cyganek, Bogusław, AGH University of Kraków, Poland
- Lirkov, Ivan, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- Stpiczyński, Przemysław, Maria Curie-Skłodowska University, Poland

V. Challenges for Natural Language Processing

Natural Language Processing technologies enable Intelligence Systems to analyze, interpret, and generate human language, enhancing their capabilities to communicate and make informed decisions. This Thematic Session is dedicated to natural language processing methods that address challenging, non–obvious problems with the goal of advancing human language technologies. Special attention is given to tools that approach multilingual tasks, utilize cross–lingual learning and process natural languages that are not widely represented in other events.

This Thematic Session was organized by:

- Kobyliński, Łukasz, Institute of Computer Science, Polish Academy of Sciences, Poland
- Kubis, Marek, Adam Mickiewicz University, Poland

VI. Computational Optimization

In the real world, Intelligence Systems address numerous problems in the fields of engineering, economics, medicine, and other domains. Many of these problems can be formulated as optimization tasks, in particular, we may consider challenges that are frequently characterized by non–convex, non–differentiable, discontinuous, noisy, or dynamic objective functions and constraints that ask for adequate computational methods.

The aim of this Thematic Session is to stimulate communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

This Thematic Session was organized by:

- Fidanova, Stefka, Institute of Information and Communication Technologies at Bulgarian Academy of Sciences, Bulgaria
- Mucherino, Antonio, IRISA, University of Rennes, France
- Zaharie, Daniela, West University of Timisoara, Romania

VII. Data Science in Health, Ecology and Commerce

Data Science in Health, Ecology and Commerce is a thematic session on the intersections of data analysis, data economics, information systems and data-based research. The workshop is explicitly focusing on the interaction of those four fields and interdisciplinary research. Here, data–driven solutions can be generated by understanding complex real–world (e.g. smart health services, interaction of ecological factors on wellbeing or the economic effects of sustainable infrastructures) related problems, critical thinking and analytics to derive knowledge from (big) data. The past years have shown a forthcoming interest in innovative data technology and analytics solutions that link and utilize large amounts of data across individual digital ecosystems. Innovative applications scenarios in the field of health, smart cities or sustainable value systems merge data from various IoT devices, social media or application systems and demonstrate the great potential for gaining new insights, supporting decisions or providing smarter services. Together with inexpensive sensors and computing power we are ahead of a world that bases its decisions on data. However, we are only at the beginning of this journey, and we need to further explore the required methods and technologies as well as the potential application fields and the impact on society and economy. This endeavor needs the knowledge of researchers from different fields applying diverse perspectives and using different methodological directions to find a way to grasp and fully understand the power and opportunities of data science.

This Thematic Session was organized by:

- Bumberger, Jan, Helmholtz Centre for Environmental Research, Germany
- Fernandes De Muylder, Cristiana, Universidade FUMEC, Brazil
- Franczyk, Bogdan, Wrocław University of Economics, Poland and Leipzig University, Germany
- Häckl, Dennis, Leipzig University, Germany
- Militzer-Horstmann, Carsta, Leipzig University, Germany
- Reinhold, Olaf, Cooperative State University Saxony, Germany

VIII. Education & AI systems

Etymologically, the word “school” originally meant “leisure,” or from the ancient Greek σχολή (scholē). Today, this foundational sense of leisure has often been lost as modern education focuses heavily on efficiency and structure. However, recent advances in AI technology, particularly generative AI (G–AI), have opened doors to newly creative ways of transmitting and sharing knowledge. This thematic session seeks to explore the transformative potential of AI-driven tools, especially chatbots and conversational agents, in education.

This session welcomes perspectives ranging from critical assessments to enthusiastic endorsements of AI's role in education. Questions include:

- What is now possible in education with the massive advent of new AI technologies?
- How can educational approaches be transformed to meet contemporary needs?
- What standards and best practices should be developed?
- What are the positive aspects, threats, and challenges of AI implementation?
- How can AI tools make education more creative, high–quality, and accessible, especially for underprivileged groups?
- How can G–AI enhance informed decision-making while addressing potential threats and security issues?

This Thematic Session was organized by:

- Dan, Daniel, MODUL University, Austria
- Nakayama, Minoru, Institute of Science Tokyo, Japan
- Wróblewska, Warsaw University of Technology, Poland

IX. Information Systems Management

This thematic session is a forum for the exchange of ideas for practitioners and theoreticians working in the broad area of information systems management in organizations. We invite you to submit papers from three complementary directions: information systems management in organizations, the use of information systems to strengthen the position of managers, and management information systems in the context of sustainable development. We are particularly interested in aspects of the adaptation of modern technologies in management information systems, i.e. the possibilities of using artificial intelligence, robotic process automation, and business intelligence in supporting the work of managers, decision–making, and setting directions for the strategic development of organizations. We also we expect papers on the impact of information systems on sustainable development and the assessment of this process.

This Thematic Session was organized by:

- Bicevska, Zane, University of Latvia, Latvia
- Chmielarz, Witold, University of Warsaw, Poland
- Leyh, Christian, Technische Hochschule Mittelhessen, Germany
- Sołtysik-Piorunkiewicz, Anna, University of Economics in Katowice, Poland

X. Model Driven Approaches in System Development

For many years, various approaches in system design and implementation differentiate between the specification of the system and its implementation on a particular platform. People in software industry have been using models for a precise description of systems at the appropriate abstraction level without unnecessary details. Model–Driven (MD) approaches to the system development increase the importance and power of models by shifting the focus from programming to modeling activities. Models may be used as primary artifacts in constructing software, which means that software components are generated from models.

In the AI era, MD approaches provide a structured framework for the design, development and optimization of complex systems by creating models that represent different system components, learning algorithms, neural networks and/or other components of a system. At the same time, adaptability is increased as the entire system can be updated quickly without having to rewrite large parts of the code, making it easier to integrate new functions or intelligent components.

A goal of the proposed workshop is to bring together people working on MD approaches, techniques and tools as well as domain–specific modeling (DSM) and domain–specific languages (DSL) and applying them in all aspects of intelligent or other information systems, in order to share their experiences, develop new ideas, evaluate and improve MD approaches, and disseminate their application. The intent is to strive for an interdisciplinary nature of MD approaches in software development, as well as research topics expressed by, but not limited to, acronyms such as Model Driven Software Engineering (MDSE), Model Driven Development (MDD), Domain Specific Modeling (DSM), and OMG's Model Driven Architecture (MDA). Apart from analyzing application perspectives of MD approaches in intelligent systems, another important aspect nowadays that MDASD intends to cover is how to deploy modern AI/ML based approaches in the software process based on a wide spectrum of MD approaches.

This Thematic Session was organized by:

- Milašinović, Boris, University of Zagreb, Croatia
- Ristić, Sonja, University of Novi Sad, Serbia
- Tolvanen, Juha-Pekka, MetaCase, Finland

XI. International Forum on Cyber Security, Privacy, and Trust

In today's digital age, information security serves as a backbone for protecting both user data and electronic transactions. Safeguarding communications and data infrastructures in an increasingly interconnected world has become more critical than ever. Security evolved into a vital scientific discipline, encompassing multifaceted complexities that require the collaboration of experts from computer science, engineering, information systems, and intelligence systems communities. Information security continues to advance through well–established technical research directions, including access control (user authentication and authorization), protocol security, software security, and data cryptography. Furthermore, emerging topics in organizational and intelligence–driven security have gained prominence alongside these traditional research areas.

The International Forum on Cyber Security, Privacy, and Trust thematic session emphasizes the diverse developments and deployments in cyber information security. It aims to highlight current challenges and present the latest research contributions. The forum serves as a platform to explore technical aspects of cyber security, innovative user privacy techniques, and trust mechanisms. Additionally, it broadens

its scope to address emerging topics, such as social, organizational, and intelligence system–driven security research directions. NEMESIS'25 provides an inclusive forum for presenting theoretical and applied research papers, case studies, implementation experiences, and work–in–progress results in cyber security.

This Thematic Session was organized by:

- Debar, Hervé, Télécom SudParis, Institut Mines-Télécom, France
- Felkner, Anna, NASK—National Research Institute, Poland
- Ioannidis, Sotirios, Technical University of Crete, Greece
- Kadobayashi, Youki, Nara Institute of Science and Technology, Japan
- Świątkowska, Joanna, European Cyber Security Organisation, Belgium

XII. Self Learning and Self Adaptive Systems

The proposed thematic session is focused on advanced Artificial Intelligence (AI) techniques and their applications in various fields. One of the advanced forms of AI methods is the Reinforcement Learning (RL) algorithms. RL or Self-learning systems are AI agents able to acquire and renew knowledge over time without hard coding. These are adaptive systems whose functions improve by a learning process based—typically—on the method of trial and error. A self-learning system initially interacts with its users or surrounding environment by attempting and observing the changes produced by its actions. This session will focus on the design, implementation, and exploitation of self-learning features within an Intelligent Environment or some of its components. The thematic session will represent an opportunity for academia and industry to debate the state–of-the-art challenges and open issues.

This Thematic Session was organized by:

- Naeem, Muddasar, Università Giustino Fortunato, Italy
- Coronato, Antonio, Università Giustino Fortunato, Italy
- Syed Tahir Hussain Rizvi, Wenn AsA, Norway

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 429–440 ISSN 2300-5963 ACSIS, Vol. 43
DOI: 10.15439/2025F5652

Multi-level Optimization-based Ensemble Machine Learning for Efficient Crop Yield Prediction in Saudi Arabia

Raya Aldawoud*, Zohra Sbaï*†
*Department of Computer Science, Prince Sattam bin Abdulaziz University, Al Kharj, Saudi Arabia
†National Engineering School of Tunis, Tunis El Manar University, Tunis, Tunisia

z.sbai@psau.edu.sa

***Abstract*—Accurate crop yield prediction is essential for enhancing food security, optimizing resource use, and supporting smart farming initiatives. Traditional statistical models often fail to capture the nonlinear interactions between environmental, climatic, and agronomic variables. To address these challenges, this research evaluated seven ensemble machine learning models: Random Forest, Bagged Decision Tree, AdaBoost, Gradient Boosting, Stochastic Gradient Boosting, eXtreme Gradient Boosting, and CatBoost. Being interested in the region of Saudi Arabia, we collected and integrated multi-source data that resulted in meteorological factors (temperature, humidity, precipitation), vegetation indices, pesticide usage, and historical yield records of thirty eight crop categories. While selecting only eight crops to study in the optimization phase, each model was assessed under four experimental configurations: baseline models, hyperparameter tuning, outliers' removal, and Bayesian optimization. Results showed that the optimized models performed up to 47% better than the default models. More precisely, hyperparameter tuning showed marginal gains, and Bayesian optimization and outliers' removal led to noticeable performance improvements. However, the effectiveness of each strategy was crop-specific.**

***Index Terms*—Crop Piield Prediction, Ensemble Machine Learning, Bagging, Boosting, Hyperparameter Tuning, Bayesian Optimization**

I. Introduction

ENSURING sustainable food production is a global challenge, especially in arid regions like Saudi Arabia, where agriculture is constrained by extreme climate, water scarcity, and limited arable land. Crop yield prediction has emerged as a critical tool to enhance agricultural planning, reduce resource waste, and support national food security strategies. Accurate forecasting allows for informed decisions in land management, irrigation scheduling, and fertilizer application, while also aiding in mitigating the risks of climate variability and supply chain disruptions.

Despite the advances in machine learning (ML) and deep learning (DL) methods for yield prediction, many existing studies are either based on synthetic or global datasets and rarely reflect the local agro-climatic and environmental realities of the Gulf region in general. Moreover, predictive models are often developed for major cereal crops like wheat and maize, with limited attention to a broader range of fruits and vegetables that are vital to the regional diet and agricultural economy. The present study addresses these gaps by developing an ensemble learning-based framework to forecast the yield of several crops grown in Saudi Arabia.

We propose a framework that leverages multiple ensemble learning techniques including Random Forest, Bagging Decision Trees, Gradient Boosting, Stochastic Gradient Boosting, XGBoost, AdaBoost, and CatBoost to exploit the strengths of individual models while enhancing robustness and accuracy. For each crop, the models are trained and evaluated using historical yield data alongside environmental variables such as temperature, precipitation, humidity, NDVI, VCI, WDRVI, and pesticide application rates.

In our research, we seek to develop predictive models specifically tailored to key crops grown in Saudi Arabia by utilizing real, localized agricultural data. The study aims to systematically compare the performance of various ensemble learning algorithms across multiple crop types to determine their relative strengths and limitations. Also, it seeks to identify the most effective model for each crop by evaluating their performance using standard regression metrics such as the coefficient of determination (R^2), Mean Absolute Error (MAE),Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Ultimately, the research aims to provide a scalable and reliable modeling framework that supports yield forecasting, enhances strategic food security planning, and promotes the adoption of smart agriculture practices in arid and data-constrained environments.

By combining local data with advanced ensemble learning strategies, this research contributes to the growing body of precision agriculture literature and offers practical insights for decision-makers aiming to modernize agriculture in Saudi Arabia. More precisely, the contributions of the present research are the following:

- Unified ensemble-based prediction framework tailored to the Saudi agro-environmental context.
- Comprehensive comparative analysis of seven ensemble machine learning models across multiple crop categories, revealing insights into their relative performance under different optimization conditions.

- Multi-stage optimization pipeline to enhance models' accuracy and stability; including direct optimization techniques such as hyperparameter tuning and Bayesian search and methods of enhancing performance like outliers' removal.

The remainder of the paper is organized as follows. Section 2 recalls the works existing in the literature for crop yield forecasting. Section 3 details the methodology adopted to realize the aforementioned contributions. The main results are discussed in section 4. Finally, section 5 concludes the paper.

II. Related work

Crop yield prediction has emerged as a critical area of research in agriculture [32], especially in the face of climate change, population growth, and food security challenges. Accurate prediction of agricultural production enables farmers, and agribusiness stakeholders to make decisions about resource allocation, land management, and market planning. It also plays a role in supporting food sustainability and reducing risks in agricultural investment and planning [11], [18].

In recent years, artificial intelligence (AI) and machine learning (ML) have been increasingly adopted for crop yield forecasting due to their capacity to model complex, nonlinear relationships among environmental, soil, and agronomic variables [12], [18]. These technologies have shown results compared to traditional statistical methods, especially when dealing with large datasets.

Palanivel and Surianarayanan [13] reviewed ML methods for crop yield prediction, highlighting ANN and SVM for handling nonlinear agricultural data. Studies from India, China, and South Africa covered wheat, maize, rice, cotton, and potatoes. Key features: climate (rainfall, temperature, humidity) and soil. ANN showed high accuracy—e.g., $\pm 9\%$ error in wheat yield. A big data framework was proposed to improve prediction via distributed processing.

Ashfaq et al. [1] proposed an ML framework for winter wheat yield prediction in Multan, Pakistan (2017–2022), integrating meteorological, NDVI, soil, and spatial data. Three models—SVM, RF, and LASSO—were evaluated using NDVI (Landsat 8), climate, and soil features. RF outperformed others with $R^2 = 0.88$, 97% accuracy, and RMSE of 0.056 t/ha. NDVI and climate data boosted accuracy, with water-related features more impactful. The study also produced spatial yield maps and highlighted RF's suitability for heterogeneous, data-scarce regions.

Ilyas et al. [4] proposed an AI-based framework for automated crop classification and yield prediction using remote sensing and ensemble learning. A fuzzy hybrid ensembled model was developed, combining spatial filtering and data augmentation with a bagging ensemble classifier. Yield prediction used GB, RF, DT, and linear regression trained on FAO and World Bank data (2017–2021) for flaxseed, lentils, rice, sugarcane, and wheat. The ensemble classifier improved accuracy by +13% over GB and +24% over DT. GB regressor achieved the best performance with lowest MSE and highest accuracy.

Yewlea et al. [2] proposed RicEns-Net, a deep ensemble model for rice yield prediction in Vietnam's An Giang province using multi-modal data: Sentinel-1 (SAR), Sentinel-2 (MSI), Sentinel-3 (meteorology), NASA rainfall, and field data. From 557 observations, 15 features were selected via statistical filtering. RicEns-Net combines CNN, MLP, DenseNet, and AE, weighted by validation error. It outperformed all tested models, achieving MAE = 341.13 kg/ha, RMSE = 436.26 kg/ha, and adjusted R^2 = 0.589, with minimal train-test variance (ΔR^2 = 0.063).

Wang et al. [6] reviewed DL models for yield prediction across crops: corn, soybean, rice, wheat, tomato, and lettuce. They evaluated LSTM, CNN, CNN-RNN, BO-LSTM, ConvLSTM, DNN, and GRU against ML models (RF, SVR, GBR, KNN), using meteorological, soil, remote sensing, and management data. Notable results: BO-LSTM ($R^2 = 0.82$ for wheat), GRU ($R^2 = 0.98$ for corn), CNN-RNN ($R^2 = 0.9995$ for tomato). DL outperformed ML on large datasets but at higher computational cost and lower interpretability.

Morales and Villalobos [15] evaluated ML models for wheat and sunflower yield prediction using synthetic data from DSSAT models across five Spanish regions. Features included weather, soil, cultivar, and management practices. Models tested: Lasso, Ridge, RF, and ANN (ANN-2 to ANN-12). RF achieved best test results for both crops (e.g., wheat: RMSE = 5%, R^2 = 0.99 train; sunflower: RMSE = 12%, R^2 = 0.97). ANN models showed overfitting risks. Chronological validation confirmed RF's robustness over random splits, supporting its use in real-world forecasting.

El-Kenawy et al. [21] evaluated ML and DL models for potato yield prediction using a Kaggle dataset with agro-climatic, temporal, and spatial variables. Models included KNN, GB, XGBoost, MLP (ML), and GNN, GRU, LSTM (DL). GNN outperformed all, achieving $R^2 = 0.5172$ and MSE=0.0236. LSTM and GRU followed with $R^2 = 0.4474$ and 0.3872. DL models better captured spatial-temporal dependencies. The study applied extensive preprocessing, hyperparameter tuning, and emphasized DL's value for sustainable agriculture and food security.

Joshi et al. [3] explored deep transfer learning for winter wheat yield prediction in U.S. climate zones using satellite time-series and climate data (2008–2020). A BiLSTM replaced MLP to better model spatiotemporal features. Four transfer techniques were tested: TrAdaBoost.R2, Two-stage TrAdaBoost.R2, DANN, and Fine-tuning. Two-stage TrAdaBoost.R2 + BiLSTM yielded top results (MAE = 0.41/0.42; R^2 = 0.51/0.53). DANN underperformed due to domain mismatch. The study highlights BiLSTM's and domain similarity's roles in enhancing transfer learning efficacy.

Brandt et al. [5] developed an ensemble learning framework for predicting yields of winter wheat, barley, and rapeseed in Germany using data from 140,000–155,000 parcels (2019–2022). Features included Sentinel-2 imagery, meteorological data, and soil properties. Two ensemble strategies—stacking (PLSR, RF, SVR, CTB, LGB, XGBoost with Elastic Net meta-learner) and majority voting—were tested.

The voting ensemble achieved the best parcel-level R^2 scores: wheat (0.74), barley (0.68), rapeseed (0.66). High-resolution EO and agro-climatic fusion enhanced accuracy and scalability.

Rao et al. [11] proposed a supervised ML model to guide crop selection based on soil nutrients (N, P, K), temperature, humidity, pH, and rainfall, using a dataset of 2200 samples across 22 crops. Evaluated models included KNN, DT, and RF classifiers. RF achieved top test accuracy (99.32%) with both Gini and Entropy. DT reached up to 98.86%, while KNN scored 97.04%. Regression analysis (ENet, Lasso, Kernel Ridge) and stacking were also used for yield approximation. A mobile app with GPS/rainfall tracking was suggested for practical deployment.

In Saudi Arabia, multiple research papers have studied crop yield prediction. For instance, Assous et al. [12] developed a neural network model for yield prediction across Gulf countries using some features such as temperature, rainfall, nitrogen, and pesticide use, achieved a determination coefficient (R^2) of 0.93. Likewise, Al-Adhaileh and Aldhyani [9] explained that pesticide, temperature, and rainfall were among the most effective predictors for major crops in Saudi Arabia like wheat and sorghum, achieved 0.96 of R^2 and 0.0449 of root mean square error (RMSE) using a multilayer perceptron (MLP) model.

Another recent research by Islam et al. [14] used XGBoost for yield prediction like wheat, maize, sorghum, potatoes and rice in the Saudi Arabia and obtained an R^2 score of 0.97 and RSME of 15803.15. Their research focused on the role of data in promoting climate-resilient agriculture in the region.

Al-Gaadi et al. [25] developed a precision agriculture framework to predict potato yield in Eastern Saudi Arabia using multispectral imagery (Landsat-8, Sentinel-2). Vegetation indices (NDVI, SAVI, CNDVI, CSAVI) were correlated with field yield data from 30-ha plots. Linear regression models showed Sentinel-2 performed best ($R^2 = 0.65$, RMSE = 4.96%), with optimal prediction 60–70 days post-planting. The study highlights the effectiveness of high-resolution satellite monitoring for yield estimation in arid regions.

Al-Gaadi et al. [26] assessed the water footprint (WF) and productivity of carrots and onions in arid Saudi conditions using Sentinel-2, Landsat-8, and the SSEB model. Spectral indices (NDVI, SAVI, RDVI, EVI) and LST were used to predict yield and evapotranspiration. Linear regression with NIR band achieved $R^2 = 0.77$ (carrots) and 0.68 (onions). Estimated WF was 312 m^3/t (carrots) and 230 m^3/t (onions), confirming the value of satellite-based monitoring for yield and water use efficiency.

Li et al. [27] performed a national-scale analysis of Saudi agriculture (1990–2021) using a hybrid ML framework combining DBSCAN, CNNs, and spectral clustering to map 28,000+ center-pivot fields via Landsat NDVI composites. Delineation accuracy was high ($R^2 > 0.97$; producer's: 83.7–94.8%, user's: 90.2–97.9%). Results showed expansion (2010–2015) and post-2016 contraction tied to water policy. The study offers high-resolution field data for future yield, WF, and crop type studies in arid regions.

Ahmed [28] proposed a maize yield prediction framework for Saudi Arabia using a modified MLP optimized by Spider Monkey Optimization (SMO). Trained on FAO and World Bank data (rainfall, temperature, yield), the MLP-SMO model outperformed LASSO, XGBoost, LightGBM, RF, SVM, GRNN, and LSTM, achieving $R^2 = 0.98$ and RMSE = 0.11 Mg/Ha. Statistical tests (MAE, MBE, Wilcoxon) confirmed its robustness. The study highlights the potential of hybrid neural-metaheuristic models for accurate yield forecasting in arid regions.

Jabbari et al. [29] studied IoT adoption for crop monitoring and yield prediction among 550 farmers in Jizan, Saudi Arabia. Statistical analysis showed a strong correlation between awareness and perceived benefits ($r = 0.835$, $p < 0.001$). Perceived benefits explained 21% of adoption variance ($R^2 = 0.210$), while access to information and government support were also significant predictors ($R^2 = 0.191$). The study highlights the need for targeted training and institutional backing to boost IoT adoption in precision agriculture.

A summary of the approaches for crop yield prediction in Saudi Arabia is given in table I.

TABLE I
EXISTING WORKS OF CROP YIELD PREDICTION IN SAUDI ARABIA

Ref, Year	Crop	Model	R^2	Data Source
[14], 2024	Maize, Potatoes, Sorghum, Soybean, Wheat	Bagging RF XGBoost	$\geq$0.91	FAO + World Bank Open Data
[12], 2023	Dates Maize Potatoe Wheat	Neural Network Neural Network	0.974 0.792 0.499 0.930	FAO + Thier global data
[27], 2023	Wheat, Barley, Alfalfa, Fruits, Vegetables	DBSCAN, CNN (AlexNet), Spectral Clustering	0.97	GEE, Manual digitization
[28], 2023	Maize	MLP + Spider Monkey Optimization (SMO)	0.98	FAO, World Data Bank
[29], 2023	Various	Regression (correlation + MLR)	0.698	Questionnaire
[9], 2022	Potatoes Rice Sorghum Wheat	MLP MLP MLP MLP	0.99 0.90 0.99 0.99	Kaggle
[26], 2022	Carrots	Linear Regression	0.77	USGS, WorldClim.org, Tawdeehiya Farms
[25], 2016	Potatoes	Linear Regression	0.65	USGS, INMA Co.

Despite advances in ML/DL for crop yield prediction, significant gaps exist in the literature, particularly in terms of regional adaptability to arid climates, full ensemble model evaluation, and holistic data fusion methodologies. Existing models frequently lack calibration for regions such as Saudi

Arabia, and therefore struggle with generalizable, scalable architecture. While specific ensemble strategies have been verified, few studies have systematically compared a wide range of ensemble methods across many crops and regions, a gap that this work fills by evaluating seven algorithms for Saudi crops. Furthermore, many implementations miss critical data fusion, particularly the combined influence of environmental, temporal, remote sensing, and management variables, whereas this study distinguishes the varied importance of components such as NDVI versus precipitation in the Saudi context. Also, we claim that the suggested approach has great generalizability, allowing for unlimited use across many situations, provided the necessary data is available.

III. Methodology

Sustainable food production in arid regions like Saudi Arabia faces significant challenges due to extreme climate, limited water resources, and constrained arable land. Crop yield prediction serves as a crucial tool to improve agricultural planning, optimize resources use, and support national food security. While ML and DL methods have shown promise in this domain, most prior studies rely on global datasets and focus primarily on few crops. These approaches often fail to generalize to local conditions and underrepresent crops vital to the Saudi agricultural economy, such as fruits and dates.

This research proposes a framework based on ensemble machine learning tailored to the agro-climatic context of Saudi Arabia. By leveraging multi-source datasets and a range of ensemble models including Random Forest, Bagging Decision Trees, Gradient Boosting, Stochastic Gradient Boosting, XGBoost, AdaBoost, and CatBoost, this study aims to deliver accurate and scalable yield predictions.

As illustrated in figure 1, the methodology is structured as follows: data collection and preprocessing, model training using historical and environmental variables (e.g., temperature, precipitation, NDVI, pesticide usage), performance evaluation using standard regression metrics (R^2, MAE, RMSE, MAPE), and different optimization experiments. This structured approach ensures robust, context-aware model development that supports smart agriculture in arid environments.

A. Data collection

To develop an effective crop yield prediction framework tailored to Saudi Arabia, a diverse set of datasets was compiled from reputable global platforms. The features were selected to reflect environmental, agronomic, and remote sensing variables that have been shown to impact crop productivity in prior studies.

- Crop type and yield: Crop-specific production and yield data were obtained from the the Food and Agriculture Organization (FAO) [30]. The data included total yield values (in kg per hectare) for several crops cultivated in Saudi Arabia like wheat, maize, tomatoes, potatoes, barley, sorghum, and dates.
- Pesticide use: Annual pesticide usage was collected from FAO databases. This variable represents the total quantity of pesticides applied in agricultural operations per year.
- Weather data: Meteorological data, including average annual temperature, total precipitation, and average humidity, were extracted from the World Bank Climate Data API [31]. The data were aggregated on a yearly basis for the studied country.
- Vegetation indices: The Normalized Difference Vegetation Index (NDVI) is a widely used spectral index that quantifies vegetation health and density by measuring the difference between near-infrared (NIR) and red (RED) reflectance captured by satellite sensors [1]. NDVI is calculated using the formula:

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$$

This formula exploits the fact that healthy vegetation reflects more NIR light and absorbs more red light, while sparse or unhealthy vegetation shows the opposite pattern. NDVI values range from -1 to $+1$, where higher positive values typically indicate dense, healthy vegetation, and values near zero or negative indicate barren areas, water bodies, or built-up land.
In this study, NDVI images were extracted and processed using the Google Earth Engine (GEE) platform, which provides cloud-based access to satellite imagery and enables large-scale geospatial analysis. Landsat 5, 7, and 8 Surface Reflectance (SR) collections were utilized to ensure temporal continuity from 1990 to 2022. The satellite images were first filtered for cloud-free conditions and spatially clipped to specific agricultural regions in Saudi Arabia. For each image, NDVI was computed using GEE's built-in `normalizedDifference()` function, which efficiently implements the NDVI formula.

A summary of the data sources is given in Table II.

TABLE II
SUMMARY OF DATA SOURCES USED IN THE STUDY

Source	Type of Data	Time Range
FAO	Crop yield, pesticide use	1990–2022
GEE	Vegetation indices (NDVI)	1990–2023
World Bank Climate Data	Temperature, humidity, precipitation	1990–2022

B. Data pre-processing

Following the collection of raw datasets from various sources, basic pre-processing steps were applied to prepare the data for ML modeling. These steps were essential to ensure the consistency, completeness, and relevance of the input features.

- Vegetation indices computation:
To obtain a comprehensive overview of the crop condition, NDVI is used along with VCI (Vegetation Condition Index) and WDRVI (Wide Dynamic Range Vegetation Index). The three indices give complementary vegetation

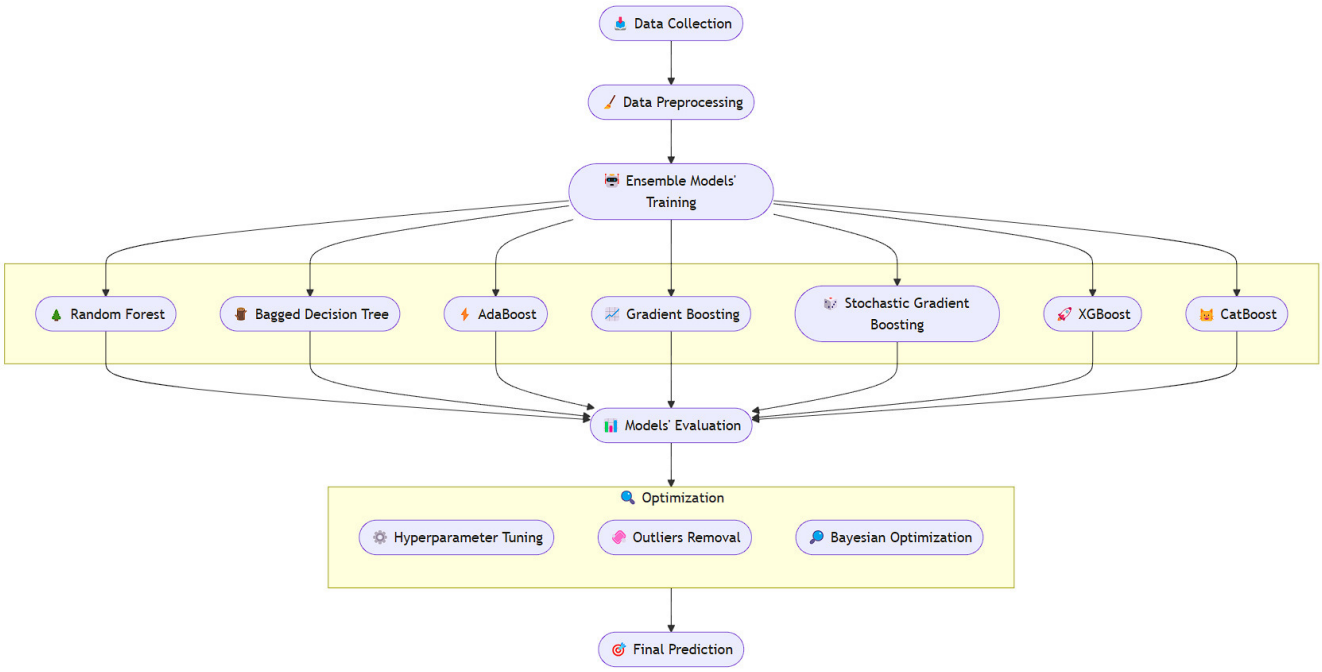


Fig. 1. Proposed framework

information, which is notably significant in agricultural surveillance and crop yield forecasting.

- VCI was calculated using the standard formula [4]:

$$VCI = \frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}} \times 100$$

- WDRVI was derived from red and near-infrared reflectance bands using:

$$WDRVI = \frac{(0.1 \cdot NIR) - RED}{(0.1 \cdot NIR) + RED}$$

where NIR and RED are the near-infrared and red reflectance values from MODIS [4].

- Data aggregation: All variables were aggregated at the annual level to align crop yield values with climate, pesticide, and remote sensing data.
- Missing data handling: Records with missing values were dropped or imputed using mean imputation where appropriate. This was particularly relevant in the early years with sparse pesticide data.
- Categorical encoding: The crop name (such as Wheat, Sorghum, Dates) was converted into numeric labels using label encoding to allow its use as a categorical feature in selected models. This is a critical step since machine learning models need numerical data. For presentation and evaluation, the original harvest names are kept in distinct variables or extracted using inverse_transform.

All datasets were harmonized to a common temporal resolution (annual), cleaned to remove missing values, and merged into a unified data frame containing 1117 records across all crop types and years. Descriptive data statistics are illustrated in figure 2.

C. Ensemble models

This research proposes a multi-stage ML pipeline to predict crop yields in various types of crop cultivated in Saudi Arabia. The methodology is centered on ensemble techniques, given that ensemble machine learning takes advantage of the complementary characteristics of many models to improve predictive accuracy, minimize variation and bias, and achieve improved generalization in complicated tasks. We implemented seven from the famous bagging and boosting models and proposed different optimizations based on regularization and hyperparameter tuning applied to the seven models. The analysis involved the evaluation of each model across all crops in the dataset, then the selection of the top-performing crops on which the optimization techniques were tested, and finally the in-depth training, tuning, and evaluation of all models per selected crop. The seven applied ensemble learning models fall in the bagging and boosting ensemble categories:

- Bagging models: Bagging models increase the accuracy of predictions by training several base learners simultaneously on random portions of the data.
  - Random Forest (RF) is an ensemble method that constructs multiple decision trees on bootstrapped samples of the data and averages their outputs. For regression, the prediction is given by:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} T_i(x)$$

	Year	Yield_Value	Temperature	Precipitation	Humidity	NDVI_Mean
count	1117.00	1117.00	1117.00	1117.00	1117.00	1117.00
mean	2006.93	12690.34	25.59	60.67	31.16	0.05
std	9.79	11713.60	0.41	5.93	0.45	0.01
min	1990.00	96.80	24.55	52.69	30.40	0.04
25%	1998.00	3974.50	25.29	56.09	30.81	0.04
50%	2007.00	10000.00	25.60	59.47	31.09	0.04
75%	2016.00	18946.80	25.88	64.34	31.41	0.05
max	2022.00	85800.50	26.21	75.59	32.04	0.06

	VCI	WDRVI_Mean	Pesticides_Value
count	1117.00	1117.00	1117.00
mean	46.48	0.03	4653.51
std	32.52	0.00	2717.36
min	0.00	0.02	994.00
25%	15.99	0.02	2683.37
50%	36.04	0.03	4216.71
75%	81.06	0.03	7255.65
max	100.00	0.03	10495.54

Fig. 2. Collected Data Statistics

where $T_i(x)$ is the prediction from the i-th DT and K is the total number of trees [14].

– Bagged Decision Tree (Bagged DT) is similar to RF but without feature randomness; trees are trained on bootstrapped datasets only.

- Boosting models: Boosting models construct ensembles in stages, with each model focusing on rectifying the faults of the previous ones.
  – Gradient Boosting (GB) iteratively minimizes a loss function via gradient descent:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

  – Stochastic Gradient Boosting (Stochastic GB) is a variant of GB that introduces randomness by sampling the training set at each iteration [20].
  – eXtreme Gradient Boosting (XGBoost) is an optimized version of GB with regularization:

$$\text{Obj} = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t)$$

  where $\Omega(f_t)$ is a complexity penalty for tree f_t [20].
  – AdaBoost sequentially trains weak learners and combines them with weights:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

  where h_m is the m-th weak learner and α_m is its weight.
  – CatBoost uses ordered boosting and handles categorical features natively to reduce prediction shift and overfitting [20].

This ensemble learning-based framework proposed for crop yield prediction was inspired by the work of Zhou [20], which details the theoretical underpinnings and practical successes of such models in various domains.

All seven models were trained and evaluated in their baseline definition (called baseline models in the sequel), then three optimizations were proposed for each model resulting in 21 other models named respectively OR, Hypertuned, and BayesSearchCV for models optimized using Outliers Removal, Hyperparameters tuning, and Bayesian optimization.

D. Model's optimization

To enhance the accuracy and generalizability of the ensemble models, a structured optimization pipeline was implemented. This process consisted of three sequential steps: hyperparameter tuning, outlier removal, and Bayesian optimization. Each step was selected to address specific challenges in predictive modeling and improve the models' robustness.

1) Hyperparameter tuning: The optimization phase began with hyperparameter tuning using GridSearchCV, a technique that systematically explores combinations of predefined hyperparameter values. This approach enables the selection of optimal configurations for each model, such as the number of estimators, tree depth, and learning rate, by evaluating model performance across cross-validated folds.

Hyperparameter tuning plays a crucial role in reducing underfitting or overfitting, and ensures that models are well-calibrated for the underlying data distribution. Its effectiveness is well-documented in the ML literature [22].

In this study, a distinct set of parameter grids was defined for each ensemble model to account for their structural and functional differences. The search space proposed for each model, the best parameters obtained, and the crop for which the best values are achieved are summarized in table III.

2) Outliers' removal: As an *indirect* optimization method, we relied on the outliers' removal method to clean the data by removing anomalies, which identifies and excludes data points that differ considerably from the dataset's anticipated trends or distribution. This may lead to better generalization and lower error metrics (e.g., RMSE and MAE).

TABLE III
SELECTED HYPERPARAMETERS FOR TUNING THE DEVELOPED ENSEMBLE MODELS

Model	Search Space	Best Parameters	Crop
Random Forset	n_estimators: [50, 100, 200, 300]	50	Pulses, Total
	max_depth: [None, 10, 20]	None	
	min_samples_split: [2, 5, 10]	2	
	min_samples_leaf: [1, 3]	1	
	max_features: ['sqrt', 'log2', 0.8]	0.8	
AdaBoost	n_estimators: [50, 100, 200, 300, 500]	300	Pulses, Total
	learning_rate: [0.001, 0.01, 0.1, 0.5]	0.001	
	loss: ['linear', 'square', 'exponential']	exponential	
XGBoost	n_estimators: [100, 200, 300]	200	Pulses, Total
	learning_rate: [0.02, 0.05, 0.1, 0.2]	0.2	
	subsample: [0.8, 0.9, 1.0]	0.8	
	colsample_bytree: [0.8, 0.9, 1.0]	0.8	
	max_depth: [3, 4, 5]	3	
	reg_alpha: [0, 0.001, 0.01, 0.1]	0.1	
	reg_lambda: [1, 0.1, 0.5, 1.0]	0.1	
CatBoost	iterations: [100, 200, 300]	300	Pulses, Total
	learning_rate: [0.01, 0.05, 0.1, 0.2]	0.05	
	depth: [3, 4, 5]	4	
	l2_leaf_reg: [1.0, 3.0, 5.0]	5.0	
Bagged Decision Tree	n_estimators: [10, 50, 100, 200]	10	Pulses, Total
	max_samples: [0.5, 0.6, 0.8, 1.0]	1.0	
	max_features: [0.5, 0.6, 0.8, 1.0]	1.0	
	estimator__max_depth: [None, 5, 10, 15]	None	
Gradient Boosting	n_estimators: [100, 200,300,500]	100	Fruit Primary
	learning_rate: [0.05, 0.1, 0.2]	0.1	
	subsample: [0.6, 0.8, 1.0]	0.8	
	max_depth: [3, 4, 5]	3	
Stochastic Gradient Boosting	n_estimators: [100, 200, 300, 400]	400	Pulses, Total
	learning_rate: [0.05, 0.1, 0.2]	0.05	
	subsample: [0.7, 0.8, 0.9]	0.8	
	max_depth: [3, 4, 5]	3	
	min_samples_split: [2, 5]	2	
	min_samples_leaf: [1,3]	3	

Once the models were tuned, Isolation Forest was applied to remove outliers from the data. This unsupervised anomaly detection technique isolates unusual observations by randomly partitioning the feature space. Its strength lies in its scalability and independence from the data distribution assumptions.

To maintain consistency, the cleaned dataset was passed into retrained versions of the same ensemble models, each with fixed hyperparameters tailored for robust generalization. For instance, RF and Bagged DT were retrained using 100 estimators, while AdaBoost used 100 weak learners. GB and SGB were configured with 200 and 100 estimators respectively, with subsampling enabled to enhance robustness. XGBoost was applied with additional regularization (`reg_alpha` = 0.1, `reg_lambda` = 1.0) and feature subsampling (`colsample_bytree` = 0.8), aiming to mitigate overfitting on reduced data. CatBoost was configured with 200 iterations, a moderate learning rate (0.05), and a regularization term (`l2_leaf_reg` = 3.0), while disabling verbose output for computational efficiency.

All models used a consistent random seed to ensure reproducibility across experimental runs. Outliers' removal helped to mitigate the impact of noise and extreme values, which can distort model learning and inflate error metrics. This is particularly beneficial in agricultural datasets where variability is high [23].

3) Bayesian optimization: In the final step, Bayesian Optimization was conducted using `BayesSearchCV`. Unlike grid search, which exhaustively tests all combinations, Bayesian optimization models the objective function and iteratively selects hyperparameters that are likely to yield the best results. This method improves the search efficiency, particularly in high-dimensional spaces, and has been shown to outperform traditional tuning methods in many ML applications [24].

For this study, specific search spaces for each model were designed to reflect their structural characteristics and known sensitivities. The parameters used in these spaces are defined in table IV. In the same table, we report the best values obtained and the crop for which the model performs better.

These customized search spaces allowed for targeted exploration of the hyperparameter landscape while maintaining computational efficiency. All models were optimized with a fixed random seed for reproducibility, and verbosity was disabled for resource-controlled environments.

IV. RESULTS AND DISCUSSION

Rigorous assessment of the developed models is critical for verifying the efficacy of models, ensuring repeatability, and quantifying predicted reliability, which helps to enable accurate choices and trustworthy real-world deployment.

Our objective was to develop and evaluate seven ensemble machine learning models in their base configuration in addition

TABLE IV
SELECTED HYPERPARAMETERS FOR BAYESIAN OPTIMIZATION OF THE DEVELOPED ENSEMBLE MODELS

Model	Search Space	Best Parameters	Crop
Random Forset	n_estimators: (100, 1000) max_depth: (3, 25) max_features: (0.1, 1.0)	100 3 0.9274781325161314	Pulses, Total
AdaBoost	n_estimators: (50, 500) learning_rate' (0.005, 2.0, loguniform) loss: ['linear', 'square', 'exponential']	53 0.018596594453385986) exponential	Pulses, Total
XGBoost	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, loguniform) max_depth: (3, 15) subsample: (0.5, 1.0) colsample_bytree: (0.5, 1.0)	384 0.1426991843302091 14 0.8350739741344673 0.705051979426657	Pulses, Total
CatBoost	iterations: (100, 1000) learning_rate: (0.005, 0.5, loguniform) depth: (3, 15) l2_leaf_reg: (0.1, 20.0, loguniform) border_count: (32, 255)	898 0.035612784787118226 3 1.2563877010366442 59	Other fruits, n.e.c.
Bagged Decision Tree	n_estimators: (10, 200) max_samples: (0.5, 1.0) max_features: (0.5, 1.0) estimator__max_depth: (3, 20)	10 1.0 1.0 3	Fruit Primary
Gradient Boosting	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, log-uniform) max_depth: (3, 15)	247 0.14690143571379652 14	Pulses, Total
Stochastic Gradient Boosting	n_estimators: (100, 1000) learning_rate: (0.005, 0.5, log-uniform) max_depth: (3, 15) subsample: (0.5, 1.0) min_samples_split: (2, 10) min_samples_leaf: (1, 10)	269 0.03878243104032394 14 0.7268326719031495 5 2	Pulses, Total

to the application of three techniques to study the models' performance enhancement. This resulted in the evaluation of 28 models resulting from the seven ensemble models discussed earlier multiplied by the four possible configurations. For this, we chose to assess these models on selected crops instead of studying all the crops obtained in the collected data. This choice is made on the basis of training the seven baseline models on each crop in the dataset and selecting the eight crops with best results according to the following metrics:

- R^2: Measures the proportion of variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating better model fit:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Root Mean Squared Error (RMSE): is a fundamental evaluation metric widely used in regression problems, including crop yield prediction. It quantifies the square root of the average of the squared differences between predicted values $\hat{y}_i$ and actual values y_i, and is formally defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n denotes the number of observations. The squaring of errors penalizes larger deviations more than smaller ones, making MSE particularly sensitive to outliers. This sensitivity allows it to highlight substantial prediction errors that may impact agricultural decision making.

- Mean Absolute Error (MAE): Represents the average magnitude of the prediction errors, without considering their direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Absolute Percentage Error (MAPE): is a regression metric that measures the average absolute percentage difference between predicted and actual values, making it useful for determining the accuracy of the model

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

The crops selected for further investigation and the metrics obtained for the best model for each crop are summarized in table V.

A. Baseline models' Performance

As mentioned in the previous section, to establish a benchmark for model performance, all ensemble models were initially evaluated using default (untuned) configurations. The evaluation was carried out across eight diverse crop categories using standard regression metrics: MAE, RMSE, R^2, and MAPE (table V). We aim here to provide more insights about the training of the seven ensemble models (with default configurations) on the selected crops.

TABLE V
BEST MODEL AND ITS EVALUATION METRICS FOR EACH CROP

Crop	Best Model	MAE	RMSE	R^2	MAPE
Cereals, primary	AdaBoost (Baseline)	0.015	0.019	0.829	0.042
Dates	Gradient Boosting (Baseline)	0.022	0.034	0.861	0.299
Fruit Primary	Stochastic Gradient Boosting (Baseline)	0.026	0.035	0.918	0.890
Other fruits, n.e.c.	Stochastic Gradient Boosting (Baseline)	0.075	0.103	0.923	0.252
Pulses, Total	AdaBoost (Baseline)	0.003	0.005	0.976	0.007
Sorghum	Stochastic Gradient Boosting (Baseline)	0.008	0.013	0.894	0.016
Wheat	Stochastic Gradient Boosting (Baseline)	0.014	0.019	0.864	0.051
Pumpkins, squash and gourds	Random Forest (Baseline)	0.088	0.119	0.880	0.534

To visualize the overall model performance across crops, figure 3 presents a heatmap of R^2 scores, capturing the comparative behavior of all base models.

The heatmap confirms that CatBoost and Stochastic GB seem to be strong performers overall, consistently showing high R^2 scores. AdaBoost and Gradient Boosting show very good performance across most crops. Random Forest generally performs well achieving 97% for Pulses crop. On the other hand, Bagged DT and XGBoost appear to have lower R^2 scores, particularly for Dates and Other fruits, which supposes that they might not be suitable for these crop types compared to the other models.

From the lens of crops, Pulses, Fruit, and Sorghum seem to be relatively well-predicted across most models, with many R^2 scores greater than 0.80 and a score ranging between 0.91 and 0.98 for Pulses crop. Whereas, Dates and Other fruits tend to have lower R^2 scores for several models, indicating that predicting the target variable for these crops might be more challenging.

Learning curves obtained by testing the models' effectiveness on progressively bigger subsets of the input data through cross-validation were studied to identify probable variance and bias concerns. To understand how the seven models learn and generalize from data, we visualize the learning curves for "Pulses, Total" and "Dates". For Pulses crop, figure 4 shows that the models are capable of extremely quick learning. Despite limited training instances (e.g., 5-10), most models get strong R^2 values (typically greater than 0.80) on both training and cross-validated data. The curves increase rapidly and converge soon. However, figure 5 shows that for Dates, the models demonstrate slower and less efficient learning. Low training instances (e.g., up to 15-20) show a significant disparity between training R^2 (typically 1.0) and CV R^2 (near 0 or even negative), indicating early overfitting.

B. Performance of the optimized Models

For direct and indirect optimization of the developed models across the selected crops, we performed the following tasks as explained in the methodology section.

- Hyperparameter tuning
- Outliers' removal
- Bayesian optimization

The R^2 scores for all models are summarized in table VI. Accoding to the table, the proposed optimizations have significantly improved the models' performance across the studied crops, particularly for formerly underperforming predictions. For Dates crop for example, the XGBoost model showed an increase in R^2 score by $\tilde{4}7\%$ after hyperparameter tuning. Also, Bayesian optimization enhances dramatically the prediction of Dates by Bagged Decision Tree (R^2 scores passes from 0.67 to 0.92). It is indeed clear from the table that almost all the optimizated models perform better than the default models.

We also highlighted in gray the cell showing the best R^2 score for each crop for model selection. For example, the best predictor for Fruits is the hypertuned Stochastic GB model with a score R^2 of 0.98.

The capacity to accomplish such consistent generalization over a broader range of crop varieties demonstrates the success of our optimization method and establishes these models as dependable tools for the forecasting of crop yields.

V. CONCLUSION

In this research, we developed and evaluated seven ensemble-based machine learning models to predict eight crop yields in Saudi Arabia. In order to enhance the baseline models' performance, three optimization methods were developed: direct optimization via hyperparameter tuning and Bayesian optimization and indirect optimization through outliers' removal.

The results demonstrated that Stochastic Gradient Boosting and XGBoost are the best performers, each model achieving the highest R^2 values for three different crops. Then, CatBoost and AdaBoost were champions in predicting Other fruits and Cereals, resp. However, bagging models (RF and Bagged DT), and the GB model did not outperform the developed models in any of the selected crops.

It should be noted that the optimization strategies had varied effects. Hyperparameter tuning showed marginal gains, while Bayesian optimization and outlier removal (via Isolation Forest) led to noticeable performance improvements. However, the effectiveness of each strategy was dependent on the crop.

In summary, the proposed framework provided a scalable and interpretable approach to smart agriculture in Saudi Arabia. Its crop-level insights support data-driven decision-making for farmers and policymakers.

For future investigations, it is promising to expand the dataset to include soil characteristics, crop management practices, and economic indicators to improve model generalization. Also, it is propitious to explore deep learning architec-

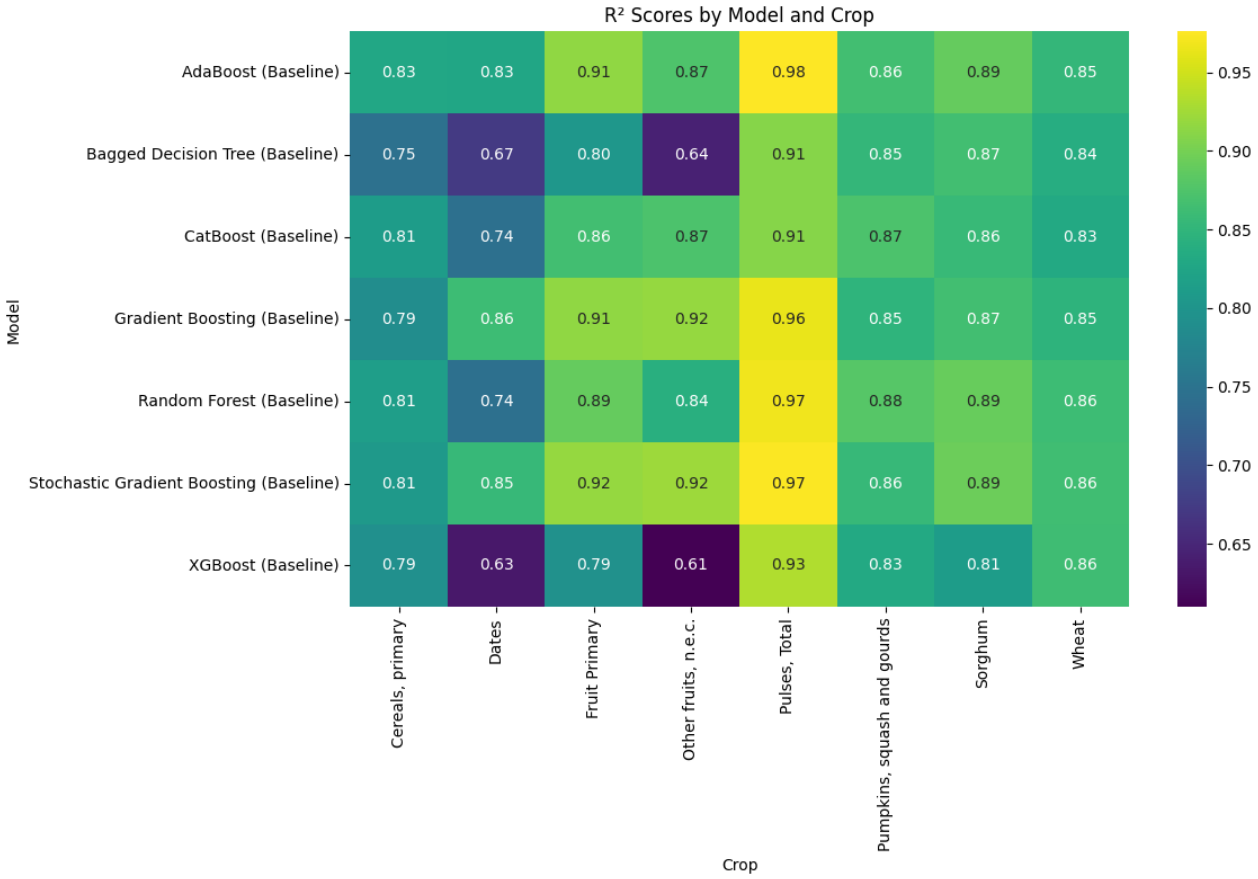


Fig. 3. Heatmap of R^2 values for baseline models across crops

tures such as LSTM or CNN to capture temporal dynamics more effectively where larger amount of data is available. Finally, exploring the performance of deep ensemble learning on the studied crops appears promising, given the encouraging results reported in [33].

REFERENCES

[1] M. Ashfaq, I. Khan, A. Alzahrani, M. U. Tariq, H. Khan, and A. Ghani, *Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion*, IEEE Access, vol. 12, pp. 40947–40961, 2024. doi:10.1109/ACCESS.2024.3376735.

[2] A. D. Yewlea, L. Mirzayeva, and O. Karakuş, *Multi-modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction*, Preprint submitted to Elsevier, Feb. 2025.

[3] A. Joshi, B. Pradhan, S. Chakraborty, R. Varatharajoo, S. Gite, and A. Alamri, *Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data*, Remote Sensing, vol. 16, no. 24, 4804, 2024. https://doi.org/10.3390/rs16244804.

[4] Q. M. Ilyas, M. Ahmad, and A. Mehmood, *Automated Estimation of Crop Yield Using Artificial Intelligence and Remote Sensing Technologies*, Bioengineering, vol. 10, no. 2, article 125, pp. 1–24, 2023. doi:10.3390/bioengineering10020125.

[5] P. Brandt, F. Beyer, P. Borrmann, M. Möller, and H. Gerighausen, *Ensemble learning-based crop yield estimation: a scalable approach for supporting agricultural statistics*, GIScience & Remote Sensing, vol. 61, no. 1, pp. 2367808, 2024. DOI: 10.1080/15481603.2024.2367808.

[6] Y. Wang, Q. Zhang, F. Yu, N. Zhang, X. Zhang, Y. Li, M. Wang, and J. Zhang, *Progress in Research on Deep Learning-Based Crop Yield Prediction*, Agronomy, vol. 14, no. 10, article 2264, pp. 1–26, 2024. doi:10.3390/agronomy14102264.

[7] J. Ansarifar, L. Wang, and S. V. Archontoulis, *An Interaction Regression Model for Crop Yield Prediction*, Scientific Reports, vol. 11, article 17754, 2021. doi:10.1038/s41598-021-97221-7.

[8] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, *A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction*, IEEE Access, vol. 9, pp. 63406–63439, 2021.

[9] M. H. Al-Adhaileh and T. H. H. Aldhyani, *Artificial Intelligence Framework for Modeling and Predicting Crop Yield to Enhance Food Security in Saudi Arabia*, PeerJ Computer Science, vol. 8, article e1104, 2022. doi:10.7717/peerj-cs.1104.

[10] F. M. Talaat, *Crop Yield Prediction Algorithm (CYPA) in Precision Agriculture Based on Internet of Things (IoT) Techniques and Climate Changes*, Neural Computing and Applications, vol. 35, pp. 17281–17292, 2023. doi:10.1007/s00521-023-08619-5.

[11] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, *Crop Prediction Using Machine Learning*, Journal of Physics: Conference Series, vol. 2161, 012033, 2022. doi:10.1088/1742-6596/2161/1/012033.

[12] H. F. Assous, H. AL-Najjar, N. Al-Rousan, and D. AL-Najjar, *Developing a Sustainable Machine Learning Model to Predict Crop Yield in the Gulf Countries*, Sustainability, vol. 15, no. 12, article 9392, 2023. doi:10.3390/su15129392.

[13] K. Palanivel and C. Surianarayanan, *An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques*, International Journal of Computer Engineering and Technology (IJCET), vol. 10, no. 3, pp. 110–118, 2019.

[14] M. M. Islam, M. Alharthi, R. S. Alkadi, R. Islam, and A. K. M. Masum, *Crop Yield Prediction through Machine Learning: A Path Towards Sustainable Agriculture and Climate Resilience in Saudi Ara-*

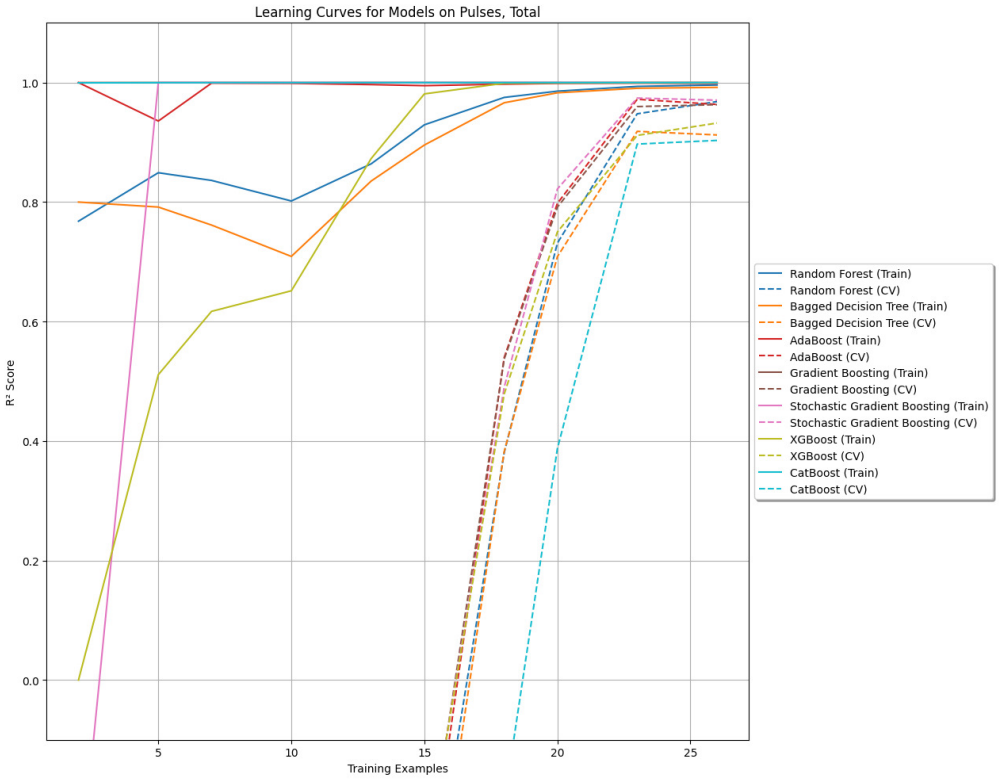


Fig. 4. Learning Curves for models on the crop: ”Pulses, Total”

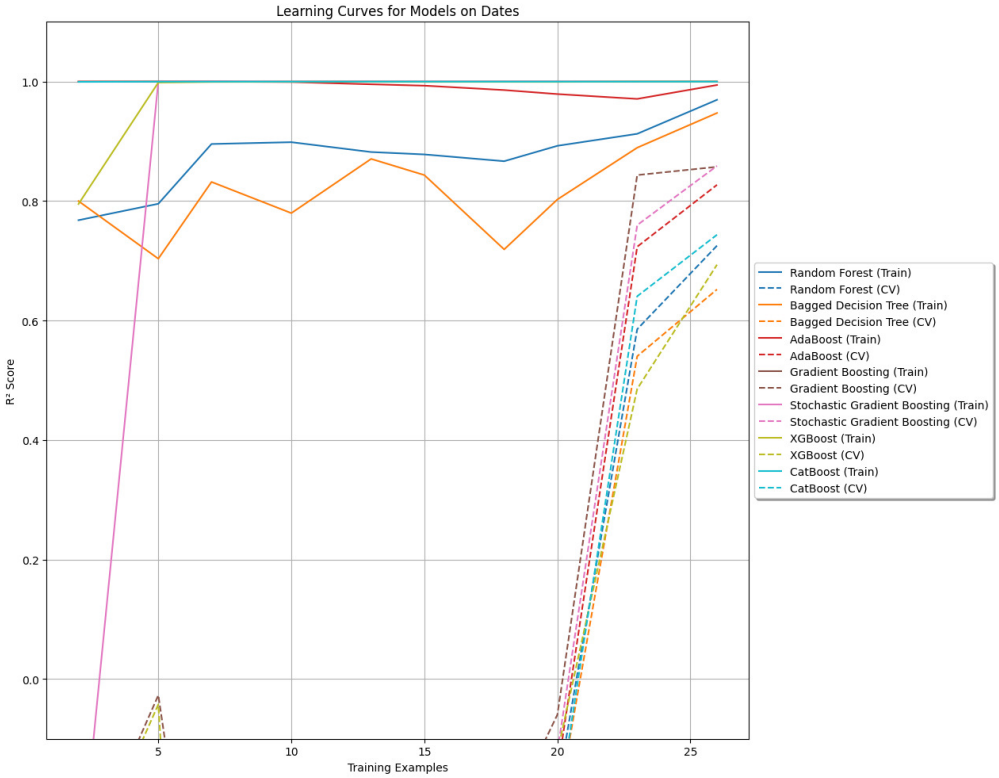


Fig. 5. Learning Curves for models on the crop: ”Dates”

TABLE VI
R^2 SCORE FOR EACH MODEL PER CROP

Model	Cereals, primary	Dates	Fruit Primary	Other fruits, n.e.c.	Pulses, Total	Sorghum	Wheat	Pumpkins, squash and gourds
AdaBoost (Baseline)	0.829	0.826	0.915	0.873	0.976	0.889	0.852	0.865
AdaBoost (Hypertuned)	0.775	0.951	0.961	0.943	0.983	0.807	0.816	0.893
AdaBoost (OR)	0.908	0.919	0.899	0.825	0.975	0.767	0.889	0.915
AdaBoost (BayesSearchCV)	0.827	0.947	0.962	0.947	0.982	0.828	0.827	0.893
Bagged Decision Tree (Baseline)	0.745	0.672	0.803	0.644	0.909	0.865	0.838	0.852
Bagged Decision Tree (Hypertuned)	0.804	0.915	0.940	0.890	0.951	0.883	0.828	0.932
Bagged Decision Tree (OR)	0.897	0.885	0.900	0.891	0.971	0.906	0.885	0.946
Bagged Decision Tree (BayesSearchCV)	0.823	0.926	0.931	0.866	0.928	0.861	0.853	0.930
CatBoost (Baseline)	0.811	0.744	0.865	0.873	0.910	0.856	0.831	0.872
CatBoost (Hypertuned)	0.705	0.896	0.923	0.963	0.971	0.863	0.906	0.911
CatBoost (OR)	0.846	0.735	0.825	0.885	0.949	0.905	0.841	0.959
CatBoost (BayesSearchCV)	0.692	0.866	0.921	0.963	0.963	0.849	0.914	0.931
Gradient Boosting (Baseline)	0.789	0.861	0.915	0.918	0.963	0.865	0.847	0.848
Gradient Boosting (Hypertuned)	0.767	0.956	0.981	0.930	0.980	0.830	0.889	0.881
Gradient Boosting (OR)	0.852	0.895	0.956	0.953	0.981	0.829	0.865	0.919
Gradient Boosting (BayesSearchCV)	0.747	0.952	0.967	0.834	0.983	0.847	0.621	0.892
Random Forest (Baseline)	0.813	0.743	0.889	0.841	0.968	0.891	0.862	0.880
Random Forest (Hypertuned)	0.773	0.906	0.941	0.891	0.953	0.846	0.869	0.948
Random Forest (OR)	0.904	0.866	0.879	0.919	0.974	0.913	0.880	0.953
Random Forest (BayesSearchCV)	0.829	0.927	0.946	0.892	0.974	0.873	0.868	0.930
Stochastic GB (Hypertuned)	0.694	0.958	0.983	0.936	0.993	0.856	0.907	0.887
Stochastic GB (Baseline)	0.807	0.855	0.918	0.923	0.974	0.894	0.864	0.855
Stochastic GB (OR)	0.848	0.912	0.934	0.941	0.967	0.850	0.861	0.911
Stochastic GB (BayesSearchCV)	0.799	0.960	0.977	0.949	0.985	0.821	0.898	0.804
XGBoost (Baseline)	0.793	0.634	0.793	0.610	0.933	0.811	0.862	0.831
XGBoost (Hypertuned)	0.784	0.933	0.860	0.833	0.961	0.845	0.802	0.911
XGBoost (OR)	0.875	0.886	0.872	0.638	0.959	0.920	0.828	0.980
XGBoost (BayesSearchCV)	0.696	0.814	0.821	0.767	0.959	0.839	0.927	0.896

bia, AIMS Agriculture and Food, vol. 9, no. 4, pp. 980–1003, 2024. doi:10.3934/agrfood.2024053.

[15] A. Morales and F. J. Villalobos, *Using Machine Learning for Crop Yield Prediction in the Past or the Future*, Frontiers in Plant Science, vol. 14, article 1128388, 2023. doi:10.3389/fpls.2023.1128388.

[16] S. V. Joshua, A. S. M. Priyadharson, R. Kannadasan, A. A. Khan, W. Lawanont, F. A. Khan, A. U. Rehman, and M. J. Ali, *Crop Yield Prediction Using Machine Learning Approaches on a Wide Spectrum*, Computers, Materials and Continua, vol. 72, no. 3, pp. 5663–5679, 2022. doi:10.32604/cmc.2022.027178.

[17] C. Trentin, Y. Ampatzidis, C. Lacerda, and L. Shiratsuchi, *Tree Crop Yield Estimation and Prediction Using Remote Sensing and Machine Learning: A Systematic Review*, Smart Agricultural Technology, vol. 9, article 100556, 2024. doi:10.1016/j.atech.2024.100556.

[18] T. van Klompenburg, A. Kassahun, and C. Catal, *Crop yield prediction using machine learning: A systematic literature review*, Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020.

[19] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, *Crop Yield Prediction using Machine Learning and Deep Learning Techniques*, Procedia Computer Science, vol. 218, pp. 406–417, 2023.

[20] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series, CRC Press, 2012.

[21] E.-S. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, *Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture*, Potato Research, vol. 68, pp. 759–792, 2025. doi:10.1007/s11540-024-09753-w.

[22] B. Bischl, M. Binder, P. Lang, T. Pfisterer, and A. Richter, *Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 11, pp. 12785–12810, 2023. doi:10.1109/TPAMI.2023.3243326.

[23] G. Pang, C. Shen, L. Cao, and A. van den Hengel, *Deep Learning for Anomaly Detection: A Review*, ACM Computing Surveys, vol. 54, no. 10, Article 200, pp. 1–38, Jan. 2020. doi:10.1145/3439950.

[24] J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 2951–2959, 2012.

[25] K. A. Al-Gaadi, R. Madugundu, E. Tola, and S. El-Hendawy, *Prediction of Potato Crop Yield Using Precision Agriculture Techniques in the Eastern Region of Saudi Arabia*, Journal of the Saudi Society of Agricultural Sciences, vol. 15, no. 2, pp. 82–89, 2016.

[26] K. A. Al-Gaadi, R. Madugundu, E. Tola, S. El-Hendawy, and S. Marey, *Satellite-Based Determination of the Water Footprint of Carrots and Onions Grown in the Arid Climate of Saudi Arabia*, Remote Sensing, vol. 14, no. 23, pp. 1–20, 2022.

[27] T. Li, O. M. L. Valencia, K. Johansen, and M. F. McCabe, *A Retrospective Analysis of National-Scale Agricultural Development in Saudi Arabia from 1990 to 2021*, Remote Sensing, vol. 15, no. 3, pp. 1–21, 2023.

[28] S. Ahmed, *A Software Framework for Predicting the Maize Yield Using Modified Multi-Layer Perceptron*, Sustainability, vol. 15, no. 4, pp. 3017, 2023.

[29] A. Jabbari, A. Humayed, F. A. Reegu, M. Uddin, Y. Gulzar, and M. Majid, *Smart Farming Revolution: Farmer's Perception and Adoption of Smart IoT Technologies for Crop Health Monitoring and Yield Prediction in Jizan, Saudi Arabia*, Sustainability, vol. 15, no. 19, pp. 14541, 2023, doi: 10.3390/su151914541.

[30] https://www.fao.org/faostat/en/#home, last accessed on May 2025.

[31] https://climateknowledgeportal.worldbank.org/, last accessed on May 2025.

[32] G. Ignesti, D. Moroni, M. Martinelli, *Towards the actual deployment of robust, adaptable, and maintainable AI models for sustainable agriculture*, Position Papers of the 19thConference on Computer Science and Intelligence Systems (FedCSIS), 2024, doi: 10.15439/2024F2991.

[33] Z. Sbai, *Deep Learning Models and Their Ensembles for Robust Agricultural Yield Prediction in Saudi Arabia*, Sustainability, vol. 17, no. 13, pp. 1-26, 2025, doi: 10.3390/su17135807.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 441–449
DOI: 10.15439/2025F2314
ISSN 2300-5963 ACSIS, Vol. 43

Leveraging Large Language Models for Automated Export Control Screening: Evaluating LLMs Framework

Salem Alotaibi*†, Alexei Lisitsa*, Antony McCabe*‡, Joanna MacSween§
*Department of Computer Science, University of Liverpool, Liverpool, UK
Emails: s.alotaibi8@liverpool.ac.uk, a.lisitsa@liverpool.ac.uk, tonymcc@liverpool.ac.uk
†Department of Computer Science and Artificial Intelligence, University of Bisha, Bisha, KSA
Email: salotaibi@ub.edu.sa
‡Computational Biology Facility, University of Liverpool, Liverpool, UK
§Export Control Officer, Legal and Governance, University of Liverpool, Liverpool, UK
Email: macsween@liverpool.ac.uk

***Abstract*—Export control (EC) compliance is a critical yet labour-intensive process within research institutions, where the classification of sensitive technologies and cross-border disclosures often depends on expert interpretation of complex legal frameworks. This paper investigates the potential of large language models (LLMs), specifically in this study ChatGPT-4o and LLaMA-3.3, to support EC screening through a multi-stage, expert-in-the-loop framework. The methodology includes prompt variation, regulatory conditioning, reflective reasoning, and expert-informed evaluation to simulate real-world compliance workflows. Using a curated dataset of UK research project descriptions and the UK Strategic Export Control List, we assess model performance across over 1,400 outputs. Results show that while both models benefit from domain-specific grounding, ChatGPT-4o consistently produces more stable and interpretable classifications. Prompt sensitivity, bias behaviour, and ambiguity handling are also examined to highlight model limitations. The findings suggest that LLMs can support early stage EC assessment but require structured prompting and human oversight to ensure regulatory alignment.**

***Index Terms*—Export control, large language models, regulatory compliance, prompt engineering, dual-use research, natural language processing, legal AI.**

I. Introduction

EXPORT CONTROL (EC) regulations play a critical role in governing the dissemination of sensitive technologies, intellectual property, and research outputs. While this study focuses specifically on the UK's export control regime—including mechanisms such as the Open General Export Licence (OGEL)[1], Standard Individual Export Licence (SIEL)[2], and the SPIRE application system [3], we reference international frameworks such as the Export Administration Regulations (EAR)[4] and the International Traffic in Arms Regulations (ITAR)[5] to provide broader regulatory context. These frameworks aim to mitigate national security and foreign policy risks by restricting unauthorised access to dual-use or controlled items. Dual-use items include those with both civilian and military applications, such as advanced materials, sensors, and encryption technologies, which are particularly sensitive in academic research contexts. Relevant legal instruments include the UK Export Control Order 2008 [6], the EU Dual-Use Regulation (EU) 2021/821 [7], the US EAR [4], and the UK Strategic Export Control List [8].

Within the academic sector, ensuring compliance with these frameworks is increasingly complex due to international collaboration, evolving research domains (e.g., quantum computing, artificial intelligence), and growing scrutiny from government bodies.

Despite their significance, export control processes remain predominantly manual and highly dependent on expert interpretation. Compliance officers are often required to assess project descriptions against regulatory texts that are both voluminous and legally dense. This results in considerable administrative burden and interpretive inconsistency, especially in institutions managing large volumes of research proposals or operating across multiple jurisdictions.

Recent developments in natural language processing (NLP), particularly through the advent of LLMs, offer new avenues for supporting regulatory compliance. Models such as GPT-4 and LLaMA have demonstrated potential in legal classification, policy reasoning, and context-aware document analysis [9], [10], [11]. However, several limitations remain. Prior work highlights that existing LLM pipelines often fail to generalise across domains, lack grounding in domain-specific legal contexts, and exhibit limited capacity for justification or error correction [12], [13], [14]. As noted by Hussain et al. [9], automated compliance in sensitive areas such as export control demands a level of interpretability and consistency that many current approaches have yet to achieve.

This paper introduces a multi-stage, expert-in-the-loop framework for evaluating the capability of LLMs to assess research disclosures under export control criteria. The framework is designed to emulate key components of institutional compliance workflows, allowing for iterative refinement through prompt variation, domain-specific grounding, reflective reasoning, and expert-informed logic. Rather than

relying on retraining or symbolic transformation, the proposed approach evaluates whether LLMs can reason effectively over full regulatory documents and project metadata.

The contributions of this paper are threefold. First, we propose a structured methodology for assessing LLM performance across different stages of reasoning, incorporating domain knowledge and expert feedback. Second, we conduct a comprehensive evaluation of ChatGPT-4o and LLaMA-3.3 across 1400 outputs using varied prompts, conditions, and error correction procedures. Third, we analyse model behaviour under ambiguity and potential bias, offering practical insights into reliability and risk in compliance decision-making.

The remainder of the paper is structured as follows: Section II reviews related work on automated legal compliance and the application of LLMs in regulatory domains. Section III outlines our proposed five-stage framework and design rationale. Section IV details the experimental setup, data sources, and evaluation procedures. Section V presents empirical findings on model performance across prompts, stages, and error correction conditions. Section VI contextualises these findings with respect to model generalisability, domain adaptation, and deployment feasibility. Finally, Section VII concludes with key insights and directions for future work.

II. Background and Related Work

Automating regulatory compliance has long posed a significant challenge for organisations operating within heavily regulated sectors such as finance, healthcare, and export control. The automation of regulatory compliance has traditionally relied on symbolic and rule-based approaches, including formal logic, ontologies, and information retrieval methods. These methods often apply sentence-level analysis and handcrafted rules to extract obligations or classify provisions. Traditional ML-based compliance checking in this area has relied on manually curated datasets and conceptual models, [15], [16]. While rule-based and symbolic methods offer transparency, many ML techniques particularly deep learning models require additional strategies to ensure interpretability and alignment with legal reasoning [9]. These methods typically focus on completeness and semantic coverage but struggle with ambiguity and cross-paragraph consistency [17]. This limitation becomes particularly pronounced in domains where legal language is dense and referential.

The emergence of large language models has prompted a shift in how regulatory artefacts are analysed. Unlike their symbolic predecessors, models such as ChatGPT-4o and LLaMA-3.3 are capable of interpreting language at the paragraph or document level, capturing contextual relationships and legal nuance in a way that more traditional techniques cannot [18]. Recent work by Hassani et al. [9] illustrates that LLMs outperform both keyword-based systems and fine-tuned BERT variants in legal classification tasks, particularly in domains like General Data Protection Regulation (GDPR) and food safety. An extension of this work [12] further examines paragraph-level grounding and contextual relevance. Similarly, [19] evaluated ChatGPT and LLaMA in the interpretation of clinical guidelines, highlighting their ability to follow structured protocols. Recent benchmarks by Yin et al. [20] also demonstrate LLM performance in semantic mapping of Sustainable Development Goals (SDGs), reinforcing their applicability in policy-oriented contexts.

LLMs have also shown potential for detecting inconsistencies in legal and technical requirements. In [21], they demonstrated that ChatGPT can be used to flag contradictions in natural language software specifications, highlighting broader utility for regulatory oversight. Similarly, Zhang et al. [11] showed that hallucination risks persist in extractive clinical NLP tasks, emphasising the need for cautious deployment in compliance settings.

Recent implementations have integrated LLMs into structured compliance systems. For example, the Gracenote platform [10] uses GPT-4 and prompt engineering to generate obligations registers, support consultation tools, and perform regulatory horizon scanning. LangChain pipelines have further enhanced LLM performance through document embeddings and retrieval-augmented generation (RAG). Al-Turki et al. [22] developed a human-in-the-loop LLM architecture for building regulation compliance, translating legal texts into structured YAML formats using GPT-4 with iterative expert feedback. Complementary work by Chen et al. [13] integrates deep learning and ontologies to check BIM (Building Information Modelling) compliance, combining LLMs with symbolic representation. These advances illustrate the growing role of LLMs as context-aware, domain-adaptable engines for compliance support.

This emphasis on context-aware interpretability and structured reasoning is echoed in recent work on legal NLP and compliance-oriented modelling. For example, a Bi-LSTM-based architecture has been used to detect explicit cause-and-effect relationships in court judgments, using labelled sentence structures to support traceable and explainable inference over legal arguments [23].

A risk-based quality control framework for legally regulated software further demonstrates how domain-specific structural representations can enhance traceability, regulatory robustness, and the auditability of automated compliance workflows [24].

The translation of natural language requirements into access control policies using LLMs has also been explored in the context of policy automation, offering insights relevant to export screening and classification workflows [25].

However, despite these developments, the use of LLMs in export control compliance remains largely underexplored. This domain involves a particularly challenging regulatory landscape governed by overlapping national and international regimes such as the UK Export Control Order, the EU Dual-Use Regulation, and the US EAR. In [14], the authors demonstrated a similar application in financial auditing, highlighting LLMs' potential to support layered and context-sensitive regulatory checks. The complexity of export classifications ranging from technical specification alignment to licensing decisions which requires a level of interpretive reasoning that

few systems currently offer.

Based on the literature, this paper highlights three key limitations:(1) the lack of annotated export control datasets for training or benchmarking; (2) minimal evaluation of LLMs across different legal jurisdictions; and (3) limited deployment of LLMs for core export tasks such as EC classification, red-flag screening, or licensing eligibility.

In this context, this paper explores a framework to evaluate whether ChatGPT-4o and LLaMA-3.3 can directly assess the applicability of export control regulations based on full-text legal documents and project descriptions. Unlike approaches that rely on intermediate representations or symbolic transformation, our framework ingests unstructured regulatory texts and infers, from context, whether export restrictions apply to a given scenario. This mirrors how compliance officers perform early-stage screening and allows us to evaluate the practical value of LLMs in complex classification tasks, particularly in domains like export control where project-specific variables and legal precision are critical.

III. Methodology

This work proposes a five-stage, expert-in-the-loop framework to evaluate and iteratively improve the ability of LLMs to classify research outputs under export control regulations. The methodology simulates regulatory decision-making by combining baseline reasoning, regulatory conditioning, error analysis, reflective prompting, and expert-guided optimisation. Each stage builds upon the previous to promote contextual learning and alignment with expert judgement.

The framework is summarised in Fig. 1 and includes the following stages:

1) **Baseline Inference (Stage 1)**: In this stage, the language model is prompted to classify domain-specific research descriptions without prior instruction. This establishes a baseline for the model's default reasoning capabilities. The model also generates a natural language justification for each prediction
2) **Domain Conditioning (Stage 2)**: The UK Strategic Export Control List is introduced to the model as context. Models are then re-evaluated on the same dataset to assess domain knowledge integration.
3) **Disagreement Analysis (Stage 3)**: Model predictions are compared with expert annotations. Disagreements are categorised to identify systematic reasoning issues.
4) **Model Reflection (Stage 4)**: In this stage, disagreement examples are reprocessed using self-prompting techniques that guide the model to reflect on its earlier decision. Prompts are crafted to encourage regulatory reasoning, contextualization, and consistency with prior outputs. The goal is to assess whether reflective prompting improves alignment without additional training
5) **Expert-in-the-Loop Optimisation (Stage 5)**: Model outputs from the reflection stage are reviewed again by the expert, who provides feedback on both classification decisions and explanations. This feedback is used to refine prompts and adjust reasoning templates, forming a closed loop for improving the model's interpretability and regulatory conformity.

This iterative process promotes improved regulatory alignment and explanation quality without the need for model retraining.

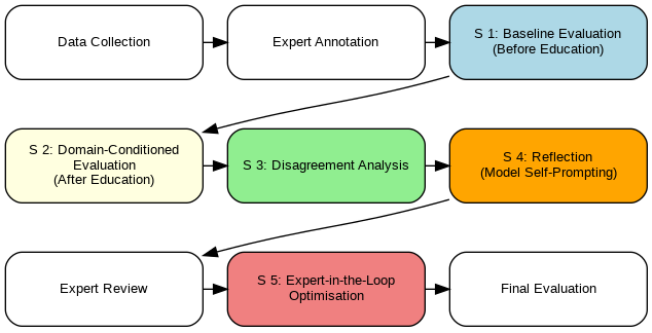


Fig. 1. Five-Stage Expert-Enhanced Framework for Export Control Classification.

IV. Experimental Setup

To evaluate the proposed methodology, a series of experiments were conducted using ChatGPT-4o and LLaMA-3.3. The setup includes data sources, model prompting, evaluation metrics, and bias analysis.

A. Data Collection

Two primary sources were used:

- **Regulatory Document:** The UK Strategic Export Control List was retrieved from official UK government publications.[1], [2], [8].
- **Project Dataset:** 50 research project descriptions were collected from the UKRI EPSRC funding portal[1]. Each entry included the fields: Title, Abstract, Project Partners, Organisation, and Department.

Each project description was independently annotated by a compliance officer from the University of Liverpool. Labels were binary, indicating whether the project should be subject to export control. These expert annotations were treated as ground truth across all experiments. Seven additional cases were marked as ambiguous for later bias and uncertainty evaluation.

B. Prompt Design and Evolution

Prompt engineering was central to model evaluation. Each stage involved a different prompting strategy:

1) Initial Prompt:

> I will give you a description of many grant projects from EPSRC in the UK. The description includes Title, Abstract, Project Partners, Organisation, and Department. I need you to tell me whether the owner of this project should apply for export control items or not and give me an explanation for your answer.

[1]https://www.ukri.org/councils/epsrc/

2) Prompt Diversity: To evaluate model robustness and sensitivity to prompt phrasing, five semantically similar variants of the initial prompt were generated per model. These new prompts were then used to rerun both the Stage 1 and Stage 2 evaluations, producing six runs per model (original + 5 variants), resulting in 24 experimental conditions and 1200 model outputs in total.

3) Reflective Prompting: Incorrect predictions were re-evaluated using prompts designed to promote self-correction:

> You previously answered that export control applies. Re-assess this decision considering the regulation provided. Identify any mistakes in your previous reasoning and provide a revised classification and justification.

4) Final Structured Prompt: A compliance-driven decision flow was embedded into a structured Yes/No prompt:

> I will describe a research or teaching project. Follow this step-by-step Yes/No flow to assess whether export controls apply:
>
> 1) Is the project exporting materials, technology, software, or information outside the UK? (Consider collaborators, overseas travel, or postgraduate students overseas.)
> → If NO, export controls do not apply. If YES, continue.
> 2) Do you know or suspect that any participant (staff, student, collaborator, or end user) intends to use the materials for military or WMD purposes?
> → If YES, export controls apply — recommend contacting `exportcontrol@liverpool.ac.uk`. If NO, continue.
> 3) Are any items subject to US export controls?
> → If YES, flag as potentially restricted and recommend further review. If NO, continue.
> 4) Are any items on the UK Strategic Export Control List, or is the project working with sanctioned entities?
> → If NO, export controls do not apply. If YES, continue.
> 5) Does the project fall under any exemptions (public domain, basic scientific research, or patent filing)?
> → If YES, export controls may not apply. If NO, export controls apply — recommend contacting `exportcontrol@liverpool.ac.uk`.
>
> Always reference UK Government and University of Liverpool guidance in your reasoning [2], [26].

C. Model Configurations

Two LLMs were tested: ChatGPT-4o, accessed via the OpenAI API, and LLaMA-3.3, accessed via the Hugging Face Inference API.

D. Evaluation Metrics

The following metrics were used: Accuracy, Precision, Recall, F1-score, and confidence scores on incorrect predictions. Confidence scores were obtained by instructing the model to label each decision with one of three categories: Low (0–40%), Medium (41–70%), or High (71–100%), based on how confident it was in its prediction. Prompt variance was measured as the standard deviation in accuracy across prompt variants.

E. Ambiguity and Bias Evaluation

Ambiguous cases were used to test confidence stability across prompts. Additionally, a bias test was conducted by modifying the institutional affiliation of partners (e.g., replacing EU with non-EU universities) to observe if geopolitical factors affected model judgement.

While this multi-stage design may appear complex, it is intended to emulate the layered decision-making practices of institutional export control review. Compliance officers often rely on contextual interpretation, iterative clarification, and expert oversight—requirements that cannot be captured by one-shot classification alone. The staged approach allows the model to integrate domain knowledge, reflect on prior outputs, and adapt to expert feedback, thereby aligning more closely with real-world compliance workflows.

V. Results

A. Initial Prompt Performance

Initial classification performance was evaluated using the original prompt, which produced three output labels: `YES`, `NO`, and `MAYBE`. In Stage 1, ChatGPT-4o tended to produce confident binary classifications (40% YES, 60% NO, 0% MAYBE), whereas LLaMA-3.3 showed greater uncertainty, assigning 30% YES, 15% NO, and 55% MAYBE. Following domain-specific conditioning in Stage 2, both models showed improved calibration. ChatGPT-4o's MAYBE rate increased slightly to 2%, while LLaMA-3.3 reduced its MAYBE usage to 8%, achieving better balance across output categories. These shifts indicate a positive effect of domain exposure in aligning outputs with regulatory expectations.

B. Prompt Sensitivity and Variance

To complement the overall classification performance analysis, we examined the extent to which prompt phrasing influences model reliability. Each model was evaluated using five semantically equivalent prompt variants differing in structure and wording. Although this introduced some fluctuation, the overall impact on classification metrics was relatively minor, particularly for ChatGPT-4o.

As illustrated in Fig. 3 and Fig. 2, ChatGPT-4o maintained stable F1-scores across all five prompts in Stage 1, ranging from 60.61% to 70.59%, with a standard deviation of 4.33 percentage points. In Stage 2, the F1-score range expanded to 64.52%–75.29%, with a standard deviation of 4.24 percentage points, reflecting slightly greater variability under conditioning. Similarly, its accuracy fluctuated by only 6.97 percentage points between the best- and worst-performing prompt.

In contrast, LLaMA-3.3 exhibited greater prompt sensitivity in Stage 1, with F1-scores ranging from 63.41% to 72.22% and an accuracy spread of 11.62 percentage points. In Stage 2, the accuracy range narrowed to 4.65 percentage points, though F1-score variability remained comparable (62.86%–70.27%, standard deviation of 3.36 percentage points). This suggests modest gains in output consistency, but not a clear improvement in classification performance.

Prompt paraphrasing thus introduced measurable variation in classification performance across both models. Structured prompts that incorporated regulatory keywords or decision templates tended to elicit more consistent and context-aware outputs, whereas open-ended variants were more sensitive to ambiguity. ChatGPT-4o also exhibited lower overall variance in prediction confidence, suggesting more stable calibration under prompt variation.

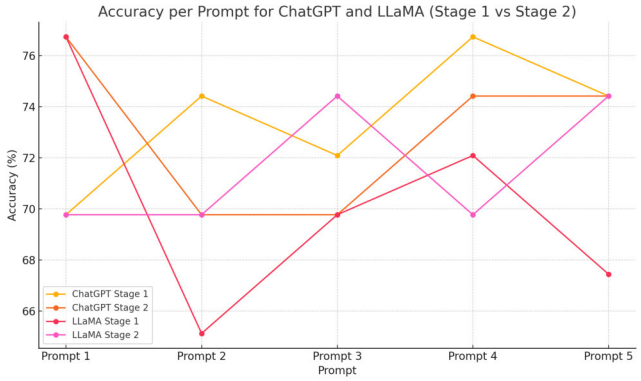


Fig. 2. Accuracy per prompt across ChatGPT-4o and LLaMA-3.3 in Stage 1 and Stage 2.

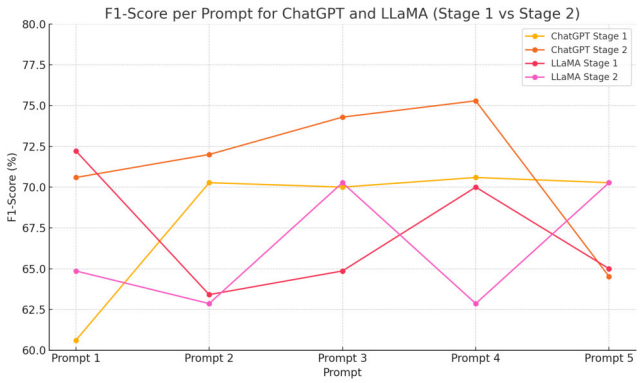


Fig. 3. F1-score per prompt for ChatGPT-4o and LLaMA-3.3 in Stage 1 and Stage 2.

C. Effect of Domain Conditioning

To evaluate the effect of domain-specific education, we compared model performance across all five prompt variants in Stage 1 (pre-conditioning) and Stage 2 (post-conditioning). This setup reflects how each model generalises its export control reasoning before and after exposure to the UK Strategic Export Control List.

Across all five prompt variants, ChatGPT-4o maintained a strong baseline with an average accuracy of 73.49% in Stage 1, and a comparable 73.02% in Stage 2. While the difference in accuracy was negligible, the model exhibited slightly improved calibration through more consistent outputs and reduced variance in confidence scores. LLaMA, by contrast, demonstrated more visible changes in response to domain conditioning. Its average accuracy improved from 70.23% to 71.63%. More notably, LLaMA-3.3's confidence in incorrect predictions decreased by approximately 8.8 percentage points which indicating better uncertainty management after incorporating regulatory context.

D. Reflective Inference Impact

Stage 4 introduced reflective prompting to evaluate whether large language models could revise incorrect classifications when explicitly asked to reassess their prior outputs. ChatGPT-4o corrected 4 out of 11 previous errors, resulting in a 36.36% resolution rate and an F1-score increase from 64.52% to 78.79%. In contrast, LLaMA-3.3 corrected none of its 11 prior misclassifications, and its F1-score declined due to additional false negatives.

A representative example illustrates this dynamic: ChatGPT-4o initially classified a cybersecurity research project with UK defence partners (e.g., RAF, Rolls-Royce) as controlled due to perceived dual-use concerns. After reflection, it revised the label to "not controlled," noting that the project lacked direct transfers or controlled components, and instead recommended internal review. This revision aligned with the expert's nuanced judgment and indicates improved reasoning alignment through structured metacognitive intervention.

These findings suggest that reflective prompting can enhance decision quality in models with sufficient contextual comprehension, such as ChatGPT-4o, but may be ineffective for models like LLaMA-3.3, which lack adequate task adaptation capacity or grounding.

E. Final Prompt Effectiveness

Stage 5 evaluated an expert-informed prompt based on institutional decision logic. As shown in Table I, this prompt led to the highest F1-scores for both models: 82.40% for ChatGPT-4o and 81.08% for LLaMA-3.3. ChatGPT-4o achieved balanced improvements in both precision and recall, while LLaMA-3.3's performance was driven by a substantial increase in recall 93.75%

TABLE I
PERFORMANCE USING EXPERT-INFORMED FINAL PROMPT (STAGE 5)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ChatGPT-4o (Stage 1)	73.49	58.58	82.67	68.35
ChatGPT-4o (Stage 5)	86.00	77.80	87.50	82.40
LLaMA-3.3 (Stage 1)	70.23	54.89	86.67	67.10
LLaMA-3.3 (Stage 5)	83.72	71.43	93.75	81.08

The table and figure demonstrate that the expert-informed prompt (Stage 5) led to the best overall performance. ChatGPT-4o reached 82.40% with balanced precision and recall, while LLaMA-3.3 attained its highest F1-score 81.08% primarily through a notable increase in recall 93.75%.

F. Ambiguity Handling and Bias Evaluation

Model behaviour on ambiguous export control cases was assessed using five prompts with uncertain control status. ChatGPT-4o maintained high and stable average confidence

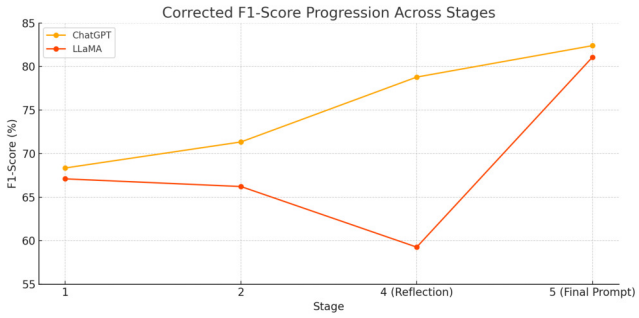


Fig. 4. F1-score progression across stages for both models. Stage 5 shows peak performance, with ChatGPT-4o reaching 82.40% and LLaMA-3.3 achieving 81.08%, driven by improved precision and recall.

levels across all prompts in Stage 1 85–92% and Stage 2 82–91%, as shown in figures 5 and 6. In contrast, LLaMA-3.3 exhibited broader fluctuations, with confidence levels ranging from 65% to 80%.

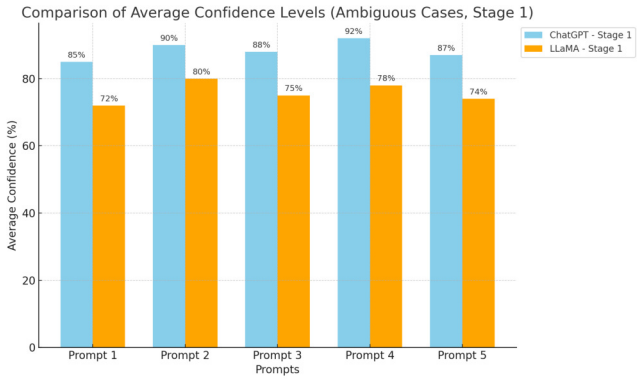


Fig. 5. Average confidence levels for ambiguous cases Stage 1. ChatGPT-4o shows higher and more consistent confidence than LLaMA-3.3 across all prompts.

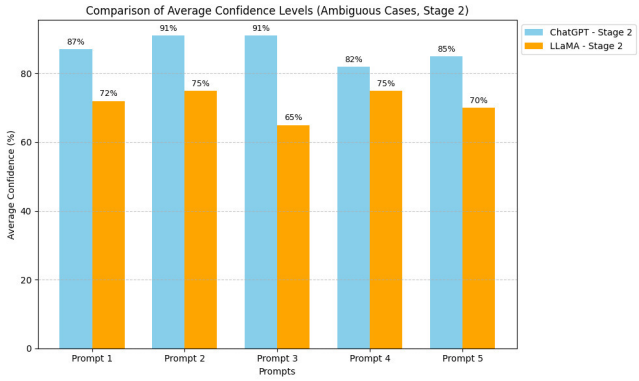


Fig. 6. Average confidence levels for ambiguous cases Stage 2. ChatGPT-4o maintains stable high confidence; LLaMA-3.3 remains more variable and generally lower.

Figure 7 illustrates the variance in confidence across stages. ChatGPT-4o's variance remained low and stable, at 43.7 in Stage 1 and 44.6 in Stage 2, whereas LLaMA-3.3's was substantially higher, at 180.3 in Stage 1 and 173.4 in Stage 2, indicating persistent volatility.

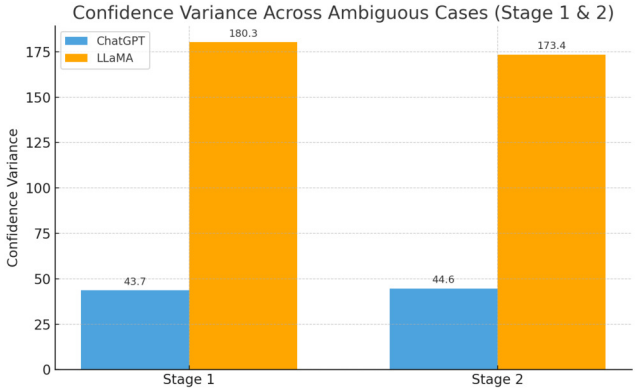


Fig. 7. Confidence variance on ambiguous cases. ChatGPT-4o shows stable variance across stages, while LLaMA-3.3's variance remains significantly higher, despite domain conditioning.

These findings suggest that ChatGPT-4o is better calibrated under uncertainty, whereas LLaMA-3.3 continues to exhibit erratic confidence distributions, reducing reliability in ambiguous classifications.

G. Bias in Partner Attribution

To evaluate the influence of institutional affiliation on compliance classification, we tested models on paired inputs with identical technical descriptions but different international partners. As shown in Table II, the inclusion of a non-EU collaborator (King Saud University) resulted in a classification change from "not controlled" to "controlled" for both ChatGPT-4o and LLaMA-3.3, despite no changes in project content, scope, or end-use.

The shift in decision was reflected in the model's generated justifications, which included references to U.S. export control frameworks (e.g., ITAR/EAR), reclassification of the technology as dual-use, and heightened concern over sensitive or military end-use. By contrast, substitution with a European partner (University of Amsterdam) produced no change in classification or in the model's accompanying rationale.

These findings indicate a model-level inclination to infer elevated risk based on geopolitical affiliation. These results suggest that models may be over-weighting partner nationality or geopolitical affiliation, even when the technical content of the project remains unchanged. This raises concerns about fairness and legal validity in automated screening, highlighting the need for explicit bias mitigation strategies in export control applications.

VI. Discussion

This study aimed to critically evaluate the capacity of large language models to reason about export control classifications using a staged framework. The experimental objectives focused on four dimensions of model behaviour: (1) prompt variation sensitivity, (2) the effect of domain-specific conditioning

TABLE II
IMPACT OF ADDING FOREIGN PARTNERS ON EXPORT CONTROL COMPLIANCE DECISIONS

Title	Project Partners	Organisation	ChatGPT-4o	LLaMA-3.3
XXXXXX	BT, Intel, Open Networking Foundation, Ori Industries 1 Ltd, Yale University, King Saud University (Saudi Arabia)	Univ. of Oxford	YES	YES
XXXXXX	BT, Intel, Open Networking Foundation, Ori Industries 1 Ltd, Yale University, University of Amsterdam	Univ. of Oxford	NO	NO

Note: For both examples shown above, the LLMs' original classification decisions (before the addition of new partners) were "NO."

("education"), (3) the impact of reflective prompting for error correction, and (4) the value of expert-informed structured prompts. In addition, model responses to ambiguous cases and bias-inducing input modifications were assessed to explore practical risks in regulatory deployment.

Prompt variation tests revealed that both models were somewhat sensitive to surface-level differences in input phrasing, though this effect was less pronounced than expected. ChatGPT-4o demonstrated stronger robustness to linguistic variation across all stages, aligning with findings by Bogireddy and Dasari [18], who observed consistent performance from ChatGPT-4o across diverse tasks. LLaMA-3.3, on the other hand, showed substantial variance in Stage 1, with up to 11.6 percentage points difference in accuracy across prompts. This instability improved notably post-conditioning. These results suggest that while prompt phrasing can influence model responses, its effect is comparatively modest in classification contexts, particularly for better calibrated models like ChatGPT-4o. Nevertheless, this sensitivity may still pose risks in high-stakes regulatory settings if users craft prompts inconsistently.

Domain-specific conditioning through the provision of the UK Strategic Export Control List was introduced in Stage 2 to simulate regulatory education. ChatGPT-4o's accuracy rose by just 0.1%, while LLaMA-3.3 improved by 0.6%. The more significant impact was observed in confidence calibration: LLaMA's confidence on incorrect predictions dropped by nearly 9 percentage points, indicating improved uncertainty management. More sophisticated grounding techniques, such as retrieval-augmented generation (RAG), may be required to meaningfully shift classification outcomes, as RAG models have been shown to improve factual accuracy, specificity, and task performance by combining pre-trained parametric memory with non-parametric knowledge sources [27].

Stage 4 explored whether models could improve their decisions through structured reflection. ChatGPT-4o successfully revised 36.4% of its prior errors, demonstrating an ability to align with regulatory expectations after self-assessment. This supports research by Lawal et al. [25], who found that prompting LLMs to reflect on prior outputs can lead to more policy aligned decisions. In contrast, LLaMA-3.3 failed to correct any outputs, reinforcing concerns around the model's adaptability and interpretive reasoning under pressure. The divergent performance highlights the need for models to not only generate initial outputs reliably but also to re-engage critically when prompted, particularly in domains that require explainability and self-consistency.

Stage 5 incorporated an expert-designed prompt that mirrored institutional export control procedures. This structured, decision-flow-based format led to the highest performance for both models, especially in recall. LLaMA-3.3's recall increased to 93.75%, a substantial leap from earlier stages. These results mirror findings from Al-Turki et al. [22], who showed that expert-guided input schemas can compensate for model instability. Unlike earlier stages, which tested the models' inherent flexibility, this final configuration emphasised alignment with operational logic. The results underscore the importance of designing prompts that do not merely inform the model but actively shape its reasoning observable in the generated outputs in a way that mirrors human expert processes. This strategy loosely aligns with principles of chain-of-thought prompting, where structured input sequences encourage more consistent and interpretable decision-making.

The final experimental tasks evaluated model behaviour in ambiguous or politically sensitive cases. ChatGPT-4o exhibited stable confidence across all uncertainty cases, while LLaMA-3.3 continued to show high variance. Bias testing further revealed problematic behaviours: both models flagged higher regulatory risk when international collaborators were from non-EU regions, despite identical technical content. This highlights the urgent need for robust bias mitigation strategies if LLMs are to be deployed in real-world compliance workflows.

In summary, the results suggest that while domain-specific input and expert-informed prompts can improve regulatory alignment, gains from prompt variation and naive domain conditioning remain modest. Models like ChatGPT-4o show stronger baseline reliability, while LLaMA-3.3 benefits more from external structure. Ultimately, consistent improvements were achieved only when expert logic was embedded directly into the input format, pointing toward the value of hybrid frameworks that combine human expertise with adaptable LLM reasoning.

Although Stage 5 achieved the highest performance metrics, it also required the most expert intervention and prompt engineering effort. In contrast, Stage 2 produced modest

but reliable gains with minimal human oversight, offering a more scalable compromise between accuracy and automation. Future deployment strategies may adopt a hybrid model, where expert input is reserved for edge cases flagged by lower-confidence outputs.

VII. Conclusion and Future Work

This study investigated the use of LLMs for export control classification of research disclosures, proposing a multi-stage, expert-in-the-loop framework. Through prompt variation, domain-specific conditioning, reflective prompting, and expert-informed refinement, we evaluated the classification performance and behavioural consistency of ChatGPT-4o and LLaMA-3.3. Our findings suggest that structured prompting and regulatory context can meaningfully enhance model reliability, with ChatGPT-4o demonstrating more consistent and interpretable outputs. However, challenges remain, particularly in handling ambiguous phrasing, edge cases, and politically sensitive affiliations.

Future work will address several areas. First, we plan to significantly expand the dataset, incorporating a broader range of research descriptions and licensing scenarios. This will enable more rigorous statistical analysis and improve generalisability. Second, we aim to evaluate additional LLM architectures, including newer open-source models and multilingual variants to assess comparative strengths across regulatory regimes. Third, we intend to explore the integration of automated outputs into institutional review processes, to assess whether human-machine collaboration improves both decision accuracy and accountability. Finally, we will extend this approach to adjacent compliance domains, such as sanctions screening and dual-use research oversight, where similar regulatory reasoning is required.

We also intend to experiment with retrieval-augmented generation (RAG) architectures to assess whether real-time document retrieval can enhance legal grounding and factual consistency, while monitoring for risks such as retrieval drift or hallucinated justifications.

Acknowledgments

Salem is a PhD student funded by the University of Bisha, Saudi Arabia.

References

[1] Export Control Joint Unit, Department for International Trade, and Department for Business and Trade, "Open general export licences (ogels)," 2025, last updated: 6 March 2025. [Online]. Available: https://www.gov.uk/government/collections/open-general-export-licences-ogels

[2] UK Department for International Trade and Export Control Joint Unit, "Do i need an export licence?" 2019, last updated: 13 August 2019. [Online]. Available: https://www.gov.uk/guidance/beginners-guide-to-export-controls

[3] UK Department for Business and Trade, "Spire export control system," 2024. [Online]. Available: https://www.spire.trade.gov.uk

[4] U.S. Department of Commerce, Bureau of Industry and Security, "Export administration regulations (ear)," 2024, accessed: May 2024. [Online]. Available: https://www.bis.gov/regulations/ear

[5] U.S. Department of State, Directorate of Defense Trade Controls, "International traffic in arms regulations (itar)," 2024. [Online]. Available: https://www.pmddtc.state.gov/ddtc_public?id=ddtc_public_portal_itar_landing

[6] UK Government, "The export control order 2008," 2008, statutory Instrument No. 3231. [Online]. Available: https://www.legislation.gov.uk/uksi/2008/3231/contents/made

[7] European Parliament and Council, "Regulation (eu) 2021/821 setting up a union regime for the control of exports, transfer, brokering and transit of dual-use items," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0821

[8] Export Control Joint Unit, Department for International Trade, and Department for Business and Trade, "Export controls: Military goods, software and technology," 2024, last updated: 10 April 2024. [Online]. Available: https://www.gov.uk/guidance/export-controls-military-goods-software-and-technology

[9] S. Hassani, M. Sabetzadeh, D. Amyot, and J. Liao, "Rethinking legal compliance automation: Opportunities with large language models," in *2024 IEEE International Requirements Engineering Conference (RE)*, 2024, pp. 432–440. [Online]. Available: https://doi.org/10.1109/RE59067.2024.00051

[10] J. Ioannidis, J. Harper, M. S. Quah, and D. Hunter, "Gracenote.ai: Legal generative ai for regulatory compliance," in *Proceedings of the 3rd International Workshop on AI for Legal Professionals (LegalAIIA)*, 2023. [Online]. Available: https://doi.org/10.2139/ssrn.4494272

[11] H. Zhang, N. Jethani, S. Jones, N. Genes, V. J. Major, I. S. Jaffe, and et al., "Evaluating large language models in extracting cognitive exam dates and scores," *PLOS Digit Health*, vol. 3, no. 12, p. e0000685, 2024. doi: 10.1371/journal.pdig.0000685. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000685

[12] S. Hassani, "Enhancing legal compliance and regulation analysis with large language models," in *2024 IEEE International Requirements Engineering Conference (RE)*, 2024. doi: 10.1109/RE59067.2024.00065 pp. 507–511. [Online]. Available: https://doi.org/10.1109/RE59067.2024.00065

[13] N. Chen, X. Lin, H. Jiang, and Y. An, "Automated building information modeling compliance check through a large language model combined with deep learning and ontology," *Buildings*, vol. 14, no. 7, p. 1983, 2024. [Online]. Available: https://doi.org/10.3390/buildings14071983

[14] A. Berger, L. Hillebrand, D. Leonhard, T. Deußer, T. B. F. D. Oliveira, T. Dilmaghani, and R. Sifa, "Towards automated regulatory compliance verification in financial auditing with large language models," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 4626–4635. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386518

[15] S. Wilson, F. Schaub, Y. Agarwal, A. Acquisti, L. Cranor, and N. Sadeh, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2016, pp. 1330–1340. [Online]. Available: https://doi.org/10.18653/v1/P16-1126

[16] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. C. Briand, "AI-enabled automation for completeness checking of privacy policies," *IEEE Transactions on Software Engineering*, vol. 48, no. 11, pp. 4647–4674, 2022. doi: 10.1109/TSE.2021.3124332. [Online]. Available: https://doi.org/10.1109/TSE.2021.3124332

[17] K. M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh, "Identifying the provision of choices in privacy policy text," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1294 pp. 2774–2779. [Online]. Available: https://doi.org/10.18653/v1/D17-1294

[18] S. R. Bogireddy and N. Dasari, "Comparative analysis of chatgpt-4 and llama: Performance evaluation on text summarization, data analysis, and question answering," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024. doi: 10.1109/ICCCNT61001.2024.10725662 pp. 1–7. [Online]. Available: https://doi.org/10.1109/ICCCNT61001.2024.10725662

[19] S. Pandya, T. E. Bresler, T. Wilson, Z. Htway, and M. Fujita, "Decoding the nccn guidelines with ai: A comparative evaluation of chatgpt-4.0 and llama 2 in the management of thyroid carcinoma," *The American Surgeon*, vol. 91, no. 1, pp. 94–98, 2025. [Online]. Available: https://doi.org/10.1177/00031348241269430

[20] H. Yin, A. Aryani, and N. Nambiar, "Evaluating the performance of large

language models for sdg mapping," arXiv, techreport arXiv:2408.02201, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.02201

[21] S. Fantechi, L. Gnesi, L. Passaro, and L. Semini, "Inconsistency detection in natural language requirements using chatgpt: A preliminary evaluation," in *Proceedings of the 2023 IEEE 31st International Requirements Engineering Conference (RE)*, 2023. [Online]. Available: https://doi.org/10.1109/RE57278.2023.00045

[22] D. Al-Turki, H. Hettiarachchi, M. M. Gaber, M. M. Abdelsamea, S. Basurra, S. Iranmanesh, H. Saadany, and E. Vakaj, "Human-in-the-loop learning with llms for efficient rase tagging in building compliance regulations," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3512434 Early Access. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3512434

[23] Łukasz Kurant, "Mechanism for detecting cause-and-effect relationships in court judgments," in *Annals of Computer Science and Information Systems*, vol. 35, 2023. doi: 10.15439/2023F4827 pp. 1041–1046. [Online]. Available: https://doi.org/10.15439/2023F4827

[24] M. Esche, L. Ho, M. Nischwitz, and R. Meyer, "Risk-based continuous quality control for software in legal metrology," in *Annals of Computer Science and Information Systems*, vol. 35, 2023. doi: 10.15439/2023F6171 pp. 451–461. [Online]. Available: https://doi.org/10.15439/2023F6171

[25] S. Lawal, X. Zhao, A. Rios, R. Krishnan, and D. Ferraiolo, "Translating natural language specifications into access control policies by leveraging large language models," in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 2024. doi: 10.1109/TPS-ISA62245.2024.00048 pp. 361–370. [Online]. Available: https://doi.org/10.1109/TPS-ISA62245.2024.00048

[26] University of Liverpool Legal and Compliance, "Export controls: How export control legislation applies to collaborating internationally," 2024, accessed: May 2024. [Online]. Available: https://www.liverpool.ac.uk/legal/exportcontrols/

[27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Y. Chang, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://doi.org/10.48550/arXiv.2005.11401

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 451–460
DOI: 10.15439/2025F4862
ISSN 2300-5963 ACSIS, Vol. 43

AraXLM: Evaluating Arabic Diacritization Tools for Cross-Language Plagiarism Detection

Mona Alshehri
0000-0002-4193-6230
Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: ma2250@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: n.beloff@sussex.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: m.white@sussex.ac.uk

***Abstract*—In recent years, plagiarism detection systems have evolved from basic lexical matching and n-gram overlap methods to Deep Learning (DL) models capable of capturing semantic relationships between texts. While these DL-based approaches have achieved notable success across various languages, their effectiveness in Arabic remains limited due to inherent linguistic ambiguities, particularly the omission of diacritical marks. This absence hinders accurate semantic interpretation and limits the ability of models to detect paraphrased or semantically obfuscated content in Arabic texts. This paper presents an evaluation of Arabic Text Diacritization (ATD) tools as the initial phase of a plagiarism detection framework designed for Arabic–English cross-lingual model text analysis (AraXLM). It describes the first stage of the framework, which focuses on assessing the performance of state-of-the-art ATD tools. An empirical analysis was conducted on six ATD models using Word Error Rate (WER), Diacritic Error Rate (DER), both with and without case endings (CE), and Bilingual Evaluation Understudy (BLEU) metrics. The results show that tools such as Shakkelha produced lower DER and high BLEU values, indicating high accuracy in diacritic restoration, while Fine-Tashkeel demonstrates the lowest WER and highest BLEU, reflecting best word-level performance. In contrast, CAMeL Tools and Mishkal display comparatively higher error rates across both metrics. These findings suggest that incorporating accurate diacritization models into Arabic NLP tasks, such as Machine Translation (MT) and Plagiarism Detection (PD), improves text normalisation and the quality of semantic embeddings. Thus, the AraXLM framework, supported by effective diacritization pre-processing, enhances linguistically aware detection of plagiarism involving Arabic text, where precise semantic alignment between languages is essential.**

***Index Terms*—Deep Learning (DL), Arabic Text Diacritization (ATD), Word Error Rate (WER), Diacritic Error Rate (DER), Case Ending (CE), Machine Translation (MT), Plagiarism Detection (PD), Bilingual Evaluation Understudy (BLEU).**

I. Introduction

DIACRITIC marks are a fundamental characteristic of the Arabic language. These symbols, placed above or below letters, indicate pronunciation and disambiguate meaning [1], [2]. Each mark has a specific position, Unicode representation, and functional role, influenced by syntactic and phonetic context. Diacritics are essential for accurate word interpretation, enhancing reading fluency, improving speech recognition systems, and facilitating natural language understanding. However, most Arabic texts omit these marks, which introduces semantic ambiguity and creates homographic conflicts [3].

A. Problem Statement

Recently, NLP techniques have started to include diacritical marks in their approaches to enhance language processing performance [4]. Most techniques used in plagiarism detection for the Arabic language preprocess the source document by removing the diacritic marks [5],[6]. The performance of these tests has still not improved due to pre-processing the document and not using semantic-based algorithms. The absence of diacritical marks in Arabic text introduces ambiguity that negatively impacts NLP tasks such as machine translation and plagiarism detection. Despite advances in deep learning and NLP, many systems either ignore diacritics or preprocess them away, limiting semantic accuracy and cross-lingual performance [7], [8]. Therefore, linguistic knowledge, such as vowel marks in Arabic language, needs to be addressed and evaluated in NLP tasks for enhancing semantic alignment across languages.

B. Research Questions

In order to address the issues that have been identified of applying and evaluating the ATD models, we will commence with these research sub-questions:

1) what is the most effective ATD tool for integrating into our proposed framework (AraXLM).
2) What issues are identified when conducting our experimental study, particularly for Arabic language?

The main research question is: given 2 sentences in different language, S1(English) and S2(Arabic), is S2 plagiarized from S1?

In order to answer this broad question, we defined the sub-questions that help to provide a complete answers to the overall research question. Our previous proposed framework was designed and presented for cross-language Plagiarism Detection (PD), including ATD as a linguistic feature [9]. The aim of our research contributes to detect the similarity between Arabic and English sentence pairs, including the diacritization process for Arabic sentences (see Figure 1). Diacritization is the process of adding diacritical marks (such as fatḥah, ḍammah, kasrah, shaddah, tanwīn, and sukūn) to unvowelled Arabic letters in order to indicate correct pronunciation and clarify meaning. This process is essential for various Natural Language Processing (NLP) tasks, such as MT, as it reduces the ambiguity caused by the absence of these marks. This paper presents the results of evaluating several tools concerning the accuracy of ATD.

In summary, our experimental design evaluates six Arabic diacritization tools using publicly available corpora. This study aims to identify the most effective Arabic ATD models to integrate into our proposed framework [9]. The evaluation employs three standard metrics: Word Error Rate (WER), Diacritic Error Rate (DER), and Bilingual Evaluation Understudy (BLEU). This multi-metric analysis provides a comprehensive comparison and deeper insight into the types of errors produced by each tool, with enhancing semantic accuracy in ML-based models. The results indicate that ML-based approaches achieve lower DER and higher BLEU scores, reflecting improved diacritic restoration and closer alignment with reference texts. However, these models produced higher WER, highlighting challenges associated with word-level tokenisation in Arabic, despite their character-level and semantic accuracy strengths.

II. Background

Arabic diacritization is a task in Arabic natural language processing (ANLP). These diacritics ensure correct pronunciation, disambiguation, and improved MT. However, the complex morphology of Arabic and the high level of ambiguity in unvowelled text make diacritization challenging task [10]. Over the years, researchers have developed various approaches to address this task, which can be categorized into three main types: rule-based, machine learning-based, and hybrid-based methods. Each approach employs different methodologies to overcome the challenges and develop accuracy in this field.

A. Rule-Based Diacritization Tools

Rule-based approaches apply diacritic marks based on predefined linguistic features (e.g., morphology, syntax, lexicon) [11]. In these tools, sequences of Arabic letters are mapped to specific diacritic categories through encoded rules derived from morphological analysers, syntactic parsers, and lexicon lookup tables [12]. However, pure rule based systems can fail with out of vocabulary (OOV) words and ambiguous contexts where multiple diacritic patterns may be valid. Moreover, validating the hundreds of rules and large lexicons required for comprehensive coverage can be time-consuming, error prone, and challenging to scale as the language evolves [12]. In contrast, rule-based diacritization tools achieve good result on words or structures explicitly defined within their linguistic frameworks, by applying deterministic rules derived from formal Arabic grammar.

B. Machine Learning Based Diacritization Tools

ML-based approaches for Arabic diacritization leverage neural networks (NNs) to learn and predict diacritic marks from text patterns. These approaches require datasets for training the model on diacritizing text pattern. ML algorithms can be categories to different types of Neural Networks (NNs): Artificial Neural Networks (ANNs), Convolution Neural Network (CNNs), Recurrent Neural Network (RNNs), and Transformer Neural Network (TNNs). Each type is used to process specific data and extract features that are relevant to the task.

FNNs are the simplest form of neural networks, consisting of input, interconnected hidden layers, and output layers. They process input data in a forward direction without cycles or loops [13]. FNNs have been used as baseline models for Arabic diacritization for capturing local patterns without requiring morphological pre-processing [14]. In ANNs types, back-propagation concept adjusts each weight in the layers. This step is done to eliminate the error (Loss) between the predicted output and the true target [15]. Thus, FNN is a subclass of ANN because the data flows on forward propagation only. However, there is no feedback loop in each node as outputs to inputs during inference. There-

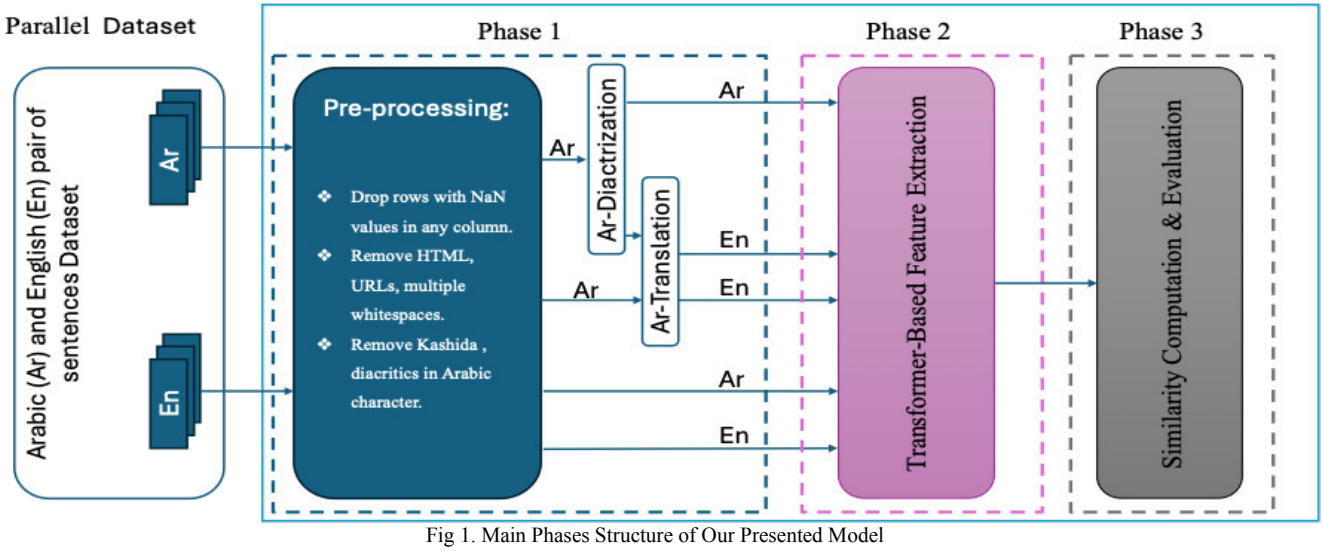


Fig 1. Main Phases Structure of Our Presented Model

fore, their ability to capture syntactic diacritization, such as case endings, is limited.

CNNs apply convolutional filters over embedding vectors to detect local patterns and have been used in image and text processing tasks [16]. In Arabic diacritization, CNNs have been employed to recognize diacritic marks in images of Arabic text, such as in Optical Character Recognition (OCR) systems, rather than for direct sequence labelling of plain text.

RNNs enable each hidden unit to receive inputs from both the current and previous time steps, thereby allowing the network to capture temporal dependencies within sequential data [17]. By combining the current input with the previous hidden state, the network updates the current hidden state, which is then used to generate the output. This sequential modelling allows RNNs to predict diacritics for each character in context [18]. However, RNNs encounter difficulties when processing long sequences due to the gradient problem. The gradient represents the scale of change used to update the network during training. Over extended sequences, gradients can either vanish (become very small) or explode (grow very large), depending on the activation functions employed [19]. This occurrence causes the network to forget earlier information in the sequence, thereby limiting its ability to learn long-range dependencies.

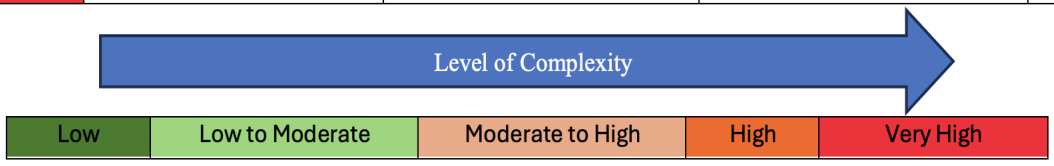
The Transformer architecture was introduced to process data in parallel by replacing recurrence and convolution with positional encoding and self-attention mechanisms [20]. In this architecture, the encoder generates contextualized representations of the input sequence, which are used by attention mechanisms in the decoder to focus on relevant parts of the input during output generation. The transformer architecture can consist of encoder and/or decoder layers. Recently, TNNs have been employed in ATD models to capture both local and global contextual information in parallel [21]. As a result, TNNs models have transformed NLP by introducing attention mechanisms that address key limitations of earlier approaches [22]. TABLE I shows the main differences between neural network approaches types. For example, TNN leverage attention mechanism for tasks requiring long sequence dependency, such as NLP task. The emergence of NN types with ATD have driven the advancement of structure models.

C. *Hybrid Based Diacritization Tools*

Hybrid-based diacritization approaches combine rule-based and machine learning (ML) methods to leverage both linguistic knowledge and contextual information in text. These tools apply ML components to handle lexical disambiguation and rule-based frameworks for syntactic case endings, capitalizing on the strengths of both paradigms. [23]. Recent research has adopted hybrid approaches for Arabic diacritization, combining rule-based linguistic knowledge with ML techniques to enhance accuracy and address the

TABLE I.
OVERVIEW OF NEURAL NETWORKS AND THEIR FEATURES

NN Type	Complexity	Data Type	Task	Feature	Authors
FNN		Structured, tabular data	Traditional ML (Classification, Regression)	One Directional (forward) Connected Layer	[13], [14]
ANN		Structured, tabular data	Traditional ML (Classification, Regression)	Fully Connected Layer	[15]
CNN		Spatial Data (Image, Videos)	Object Detection, Image recognition, Image Processing Computer Vision	Convolution Layer, Pool Layer, Flattening Layer and Fully Connected Layer.	[16]
RNN		Sequential Data (Text, Time Series)	Weather (time series) Forecasting, Speech Recognition	Recurrent Layers with hidden states that maintain a memory of previous inputs	[17], [18]
TNN		Parallelized Data (Text, image, Audio)	NLP, Text Summarization, Question Answering, Translation, PD	Attention Mechanism	[19], [20]



complexities inherent in the language [24], [25]. However, these systems, incorporating ML components, remain under development for requirement of large and good quality annotated corpora for effective training [26], [27]. Therefore, integrating rule-based and ML models require alignment and coordination to ensure consistency and resolve conflicts between these components when deployed together in hybrid tools.

III. Experimental Setup

In ATD, several studies have established systematic benchmarks frameworks, based on manually curated and expertly validated datasets, as the gold-standard for evaluating Arabic diacritization tools [28], [29]. To determine the most effective ATD tools for Phase 1 of the proposed framework [9], we conducted experiment on six diacritization tools. This experiment focused on evaluating their ability to restore diacritic marks on Arabic sentences for subsequent processing. We employed a repeated measures (within-subjects) design [30], in which each sentence (experimental unit) was processed by every diacritization tool (treatment) in the same order. This setup was designed to test the hypothesis that diacritization enhances the semantic clarity of Arabic text and aligns it more closely with its English counterparts for the next Phase.

A. Experiment Design

The dataset used in this experiment was designed by Alasmary in 2024 and consists of 742 sentences collected from internet sources covering various topics such as science, art, sport, and culture [28]. The author introduced this benchmark, known as the Character-based Arabic Tashkeel Transformer (CATT) dataset, to evaluate different ATD tools. All sentences were manually diacritized by native Arabic speakers and subsequently validated by experts to establish a gold-standard reference.

Preparation of the Arabic dataset for assessment involved several steps to ensure data quality. For example, text normalization was performed to eliminate noise caused by foreign symbols (non-Arabic characters). Second, numerical digits were converted from their digital (numeric) form into textual form, thereby ensuring the dataset is fully normalized and ready for use without additional pre-processing. This conversion is used during preparation, because Arabic numbers are represented in different numeral systems depending on regional preferences and historical influences [31]. TABLE II shows three Arabic numeral systems may appear in Arabic texts. These pre-processing steps ensure the dataset is standardised and ready for direct use without extra modification.

TABLE II.
NUMERL SYSTEMS IN ARABIC LANGUAGE [30]

System	Digit
1-Arabic numerals-European	0 1 2 3 4 5 6 7 8 9
2-Arabic-Indic	٩ ٨ ٧ ٦ ٥ ٤ ٣ ٢ ١ ٠
3-Eastern Arabic-Indic	۰ ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

B. Evaluation Metrics

We conducted our experiment using six Arabic diacritization tools and evaluated their performance by calculating the WER and the DER metrics. However, Arabic grammar requires different evaluation rules for diacritics on the final letter of words [32] as follow:

- Full Diacritization, including case endings (CE).
- Core-Word Diacritization, excluding CE.

Therefore, WER and DER can be reported with and without case endings. Each rule is used to provide distinct insights into performance measurements. For example, error rates with CE measures overall performance, including syntactic case endings that influence sentence meaning. Whereas error rates without CE measures the model's ability to restore word-internal morphology.

The `diacritization_evaluation`[1] Python package is used to calculate DER and WER metrics, relying on a predefined set of Arabic characters stored in its configuration files. The DER is computed by the file `der.py` as the percentage of mismatched diacritics-those that differ from the reference gold-standard text file over the total number of diacritic comparisons (matches and mismatches), rounded to two decimal places (see Equation 1). Similarly, the WER is calculated by the file `wer.py` as the percentage of mismatched words-those that differ from the reference gold-standard text file over the total number of word comparisons (matches and mismatches) (see Equation 2).

$$DER = \frac{Number\ of\ mismatched\ diacritics}{matched + mismatched} \quad (1)$$

$$WER = \frac{Number\ of\ mismatched\ words}{matched + mismatched} \quad (2)$$

Moreover, the BLEU (Bilingual Evaluation Understudy) metric is used to evaluate the quality of ATD models alongside WER and DER, providing a comprehensive assessment of diacritization performance. BLUE metric developed for machine translation (MT) evaluation, but it can be applied to measure the output of Arabic diacritization tools [33]. There are two types of BLEU metrics: word-level and character-level. In word-level BLEU, n-gram overlaps (default n=4) between the gold-standard dataset and the output file from the ATD models are computed, with mismatched diacritic marks counted as errors. Character-level BLEU, on the other hand, treats each Unicode code-point as a separate token for comparison.

C. Tools Performance

The following diacritization tools were compared in the experiment: CAMel Tools [34], Farasa [27], Mishkal [35], i2Text[2] online tool, Fine-Tashkeel [29] and Shakkelha [36]. These tools and their associated Python packages were implemented using the NVIDIA A100 GPU runtime on Google

[1] https://pypi.org/project/diacritization-evaluation/#files
[2] https://www.i2text.com/diacritize-arabic-text

Colab Pro. Since the experimental setup relies on multiple packages and modules written in Python, the scripts and files used in our tests were modified to suit our needs. The modifications are detailed as follows:

- Farasa Tool: The `diacritization_evaluation` package DER and WER using a predefined set of Arabic characters. However, the output generated by Farasa included a combined diacritic not originally present in the package's character file. To ensure accurate error rate calculation, we added this new diacritic mark to the set. Notably, this mark results from the combination of Shaddah and Sukun into a single diacritic, which is not standard in Arabic but appeared during the diacritization process by the Farasa tool.
- Mishkal tool, The output text file from the Mishkal tool represents the Tanween vowel as a combination double vowel (e.g., `Fatha+Fatha`). However, Tanween and Fatha have different Unicode representations when processed as diacritic marks. To handle this, we updated the diacritization_evaluation pack-age to recognize `Fatha + Fatha` as Tanween. During testing, Mishkal produced other compinged marks, such as `Fatha + Kasra`, `Kasra + Kasra`, and `Sukun + Sukun`, which we similarly added to the package's set of recognized diacritic characters.
- Fine-Tashkeel, We tokenised the input file and set the parameter `max_new_tokens=512` when invoking `model.generate` to process texts longer than the model's default context window of 128 tokens. This tool took long time (e.g., 5 hours) to diacritize the text file.

IV. Results of The Experiment

The results obtained from applying the selected ATD models using CATT benchmark are presented below:

A. WER and DER Evaluation

TABLE III and Figure 2 presents the evaluation results using WER and DER, both with and without case endings (CE). In the context of Arabic diacritization, lower WER and DER values indicate fewer errors and therefore better tool performance.

TABLE III.
WER AND DER IN % OF EVALUATED ATD TOOLS

Diacritization Tool	Case Ending %		No Case Ending %	
	WER	DER	WER	DER
CAMeL Tools	75.85	71.71	56.84	69.86
Farasa	65.79	18.04	59.99	18.47
Mishkal	60.92	74.59	39.33	73.16
Fine-Tashkeel	35.70	67.73	30.24	67.28
i2text	61.04	21.14	39.30	17.39
Shakkelha	47.45	13.22	37.76	12.0

The results show that Shakkelha and i2text yield the lowest DER values, particularly when case endings are excluded, reflecting their effectiveness in diacritic restoration at the character level. In contrast, the generation of merged and non-standard diacritic combinations during the diacritization process affected the performance and the output quality of tools such as Mishkal and Farasa. These irregular patterns introduced inconsistencies in the evaluation and interpretation of the diacritized output, which is reflected in their elevated DER scores despite achieving moderate WER. This indicates that while token-level segmentation can be acceptable, the character-level precision required for reliable diacritic restoration remains insufficient in these tools.

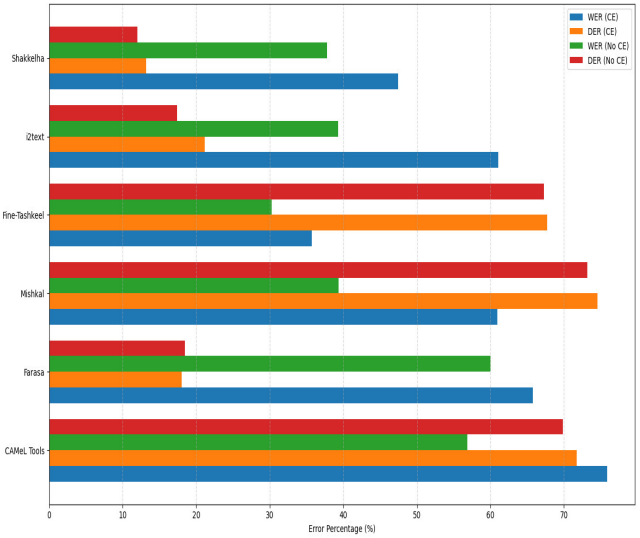


Fig 2. WER and DER by Arabic Diacritization Tool (with/without Case Endings)

Fine-Tashkeel, which leverages a fine-tuned multilingual transformer model (ByT5) with integrated tokenisation, generates fully diacritized text by assigning a diacritic to every character and applying multiple diacritic marks to some characters. This approach results in a lower WER, as words are correctly segmented. However, it leads to a higher DER, since the model introduces unnecessary or additional diacritics, increasing character-level errors. During the rendering of diacritic marks, several issues were introduced (see TABLE IV), including incorrect diacritic substitutions (e.g., fatḥah replaced with kasrah), and the addition of unintended punctuation such as parentheses or colons. These errors highlight the need for enhanced morphological awareness and post-processing techniques to improve the accuracy of diacritic restoration and ensure better alignment with gold-standard references.

TABLE IV.
IDENTIFIED ISSUES IN FINE-TASHKEEL' ATD TOOL

Gold-Standard Reference	Fine-Tashkeel Prediction
اِنْتَقَلَتْ إِلَى رَحْمَةِ اللهِ تَعَالَى	انْتَقَلَتْ إِلَى رَحْمَةِ اللَّهِ تَعَالَى
فَعَالِيَّتَها فِي تَقْلِيلِ نِسَبِ الْإِصابَةِ وَلَيْسَ مَنْعُها بِالْكامِلِ	فَعَالِيَّتَهَا فِي تَقْلِيلِ نَسَبِ الْإِصَابَةِ وَلَيْسَ مَنْعُهَا بِالْكَامِلِ)
النَّجْرِبَةِ مَرَّةً أُخْرَى	النَّجْرِبَةِ مَرَّةً أُخْرَى:

On the other hand, CAMeL Tools tokenises each word independently and applies diacritics using the `mle.disambiguate` function. However, it introduces additional diacritical marks, such as maddah and hamza, that were not present in the original undiacritized input. TABLE V compares the output of CAMeL Tools with the gold-standard diacritized reference. Each row shows a separate sentence from the dataset, highlighting the performance of the tool in restoring Arabic diacritics. It demonstrates inconsistent performance when handling morphological structures and complex verb forms. The model shows orthographic overcorrections, including the unnecessary insertion of the dagger alif, duplication of shaddah, and frequent mismatches in grammatical case endings. As a result, both its WER and DER are high due to the inclusion of these extraneous diacritics, which negatively impact alignment with the reference text.

TABLE V.
IDENTIFIED ISSUES IN CAMEL TOOLS' ATD

Gold-Standard Reference	CAMeLTools Prediction
نَجَحَ مُسْتَشْفَى الْمَلِكِ سَلْمانَ عُضْوٌ تَجَمُّع الرِّياضِ الصِحِّيّ الْأَوَّلِ فِي تَثْبِيتِ كَسْرِ مُضاعَفٍ وَمُهَشَّمٍ لِعَضُدٍ مُصابَةٌ إِثْرَ تَعَرُّضِها لِلسُّقُوطِ	نَجَحَ مُسْتَشْفَى المَلِكُ سَلْمان عُضْوٌ تَجْمَع الرياضِ الصِحِّيّ الأَوَّل فِي تَثْبِيتُ كَسْرِ مُضاعَف وَمُهَشَّم لِعَضُد مُصابَةٌ أَثْرَ تَعَرُّضِها لِلسُّقُوطِ
وَاخْتُتِمَ الْمِرانُ بِمُناوَرَةٍ عَلَى جُزْءٍ مِنَ الْمَلْعَبِ	وَاخْتَتَمَ المَرْآن بِمُناوَرَة عَلَى جُزْء مِن المَلْعَبِ
إِذا طَلَعَتِ الْجَوْزاءُ فَامْلَأِ الْحَوزاءَ	إِذا طَلَعَتْ الجَوْزاء فَأَمْلا الحوزاء
وَأَحالَتْهُمْ إِلَى النِيابَةُ الْعامَّةِ	وَأَحالَتْهُم إِلَى النِيابَةُ العامَّة
غَفَرَ اللهُ لَها وَرَحِمَها	غَفَرَ اللهُ لَها وَرَحِمَها

B. BLEU-Based Evaluation

To provide a more comprehensive evaluation of diacritization quality, beside WER and DER, this study incorporates the Bilingual Evaluation Understudy (BLEU) metric. BLEU offers a corresponding perspective by measuring the degree of n-gram overlap between the predicted and reference outputs.

As represented in TABLE VI and Figure 3, BLEU scores were calculated at both the word level and character level to capture accuracy across different linguistic features. In addition, we employed `camel_tools.tokenizers` to reflect token-based alignment accuracy in Arabic text. This BLEU metrics evaluation enables a more understanding of how effectively each tool restores diacritics in ways that are compatible with downstream token-based NLP tasks. The word-level BLEU scores remain low across all tools, reveal-ing penalties for tokenisation mismatches and sensitivity to n-gram overlap. In contrast, character-level BLEU scores are higher, as this metric treats each Unicode character as a token, making it more suitable for assessing diacritic placement precision at the subword level. Shakkelha performs well in character-level and tokeniser-based BLEU, suggesting good subword precision and compatibility with token-based Arabic NLP frameworks. In contrast, CAMeL Tools shows the lowest scores across all BLEU metrics, consistent with its poor performance in WER/DER due to added and misaligned diacritics within sentence structure.

TABLE VI.
BLEU IN % OF EVALUATION ATD TOOLS

Tool	Diacritized %		Diacritized with camel_tools.tokenizers
	Word-level BLEU	Char-level BLEU	
CAMeL Tools	2.97	66.95	2.94
Farasa	7.94	71.55	7.87
Mishkal	13.01	78.44	13.27
Fine-Tashkeel	39.36	86.80	38.74
i2text	12.98	78.32	13.24
Shakkelha	25.07	81.72	24.69

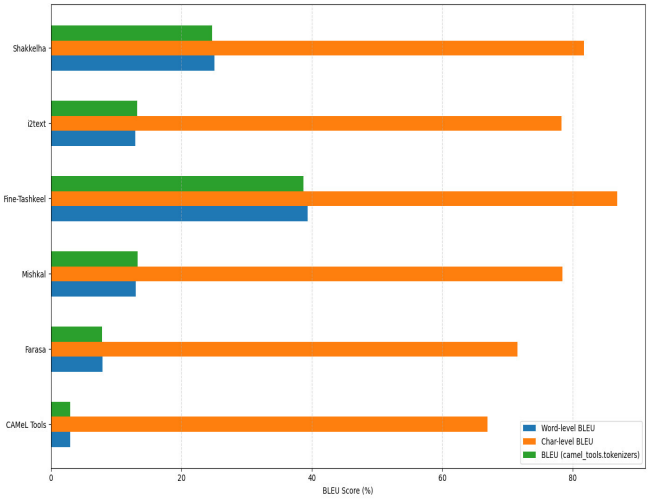


Fig 3. Comparison of BLEU Scores for Arabic Diacritization Tools

We conducted a Unicode-level analysis comparing the predicted outputs against the gold-standard references. As illustrated in TABLE VII. For instance, this is one diacritized pair:

Original no-Diacritic Marks:
"وأكدت الحوسني في إفادتها أن اللقاحات المضادة لفيروس كورونا أثبتت فعاليتها في تقليل نسب الإصابة وليس منعها بالكامل"

Gold-Standard Reference:
"وَأَكَّدَتِ الْحُوسَنِي فِي إِفَادَتِهَا أَنَّ اللِّقَاحَاتِ الْمُضَادَّةَ لِفَيْرُوسِ كُورُونَا أَثْبَتَتْ فَعَالِيَّتَهَا فِي تَقْلِيلِ نِسَبِ الإِصَابَةِ وَلَيْسَ مَنْعُهَا بِالْكَامِلِ"

Predicted by Fine-Tashkeel:
"وَأَكَّدَتِ الْحَوْسَنِيُّ فِي إِفَادَتِهَا أَنَّ اللِّقَاحَاتِ الْمُضَادَّةَ لِفَيْرُوسِ كُورُونَا أَثْبَتَتْ فَعَالِيَّتَهَا فِي تَقْلِيلِ نَسَبِ الإِصَابَةِ وَلَيْسَ مَنْعُهَا بِالْكَامِلِ"

The results showed several differences were identified at the character encoding level that impact the accuracy and evaluability of the tool's output. These irrelevant characters are introduced during token reconstruction and text formatting, and they do not reflect linguistic features. Their presence negatively affects character-level and word-level metrics. Furthermore, the predicted output included duplicated diacritics (e.g., double fathah, parenthesis, dot) and altered character spacing that increased the total Unicode code point count, which was not present in the original input or reference.

TABLE VII.
EXAMPLE OF UNICODE CHARACTER PROPRIETIES

Gold-Standard Reference	Fine-Tashkeel Prediction
وَلَيْسَ مَنْعُها بِالْكامِلِ	وَلَيْسَ مَنْعُهَا بِالْكَامِلِ)
.....	.....
و → U+0648 → ARABIC LETTER WAW	و → U+0648 → ARABIC LETTER WAW
ً → U+064E → ARABIC FATHA	ً → U+064E → ARABIC FATHA
ل → U+0644 → ARABIC LETTER LAM	ل → U+0644 → ARABIC LETTER LAM
ً → U+064E → ARABIC FATHA	ً → U+064E → ARABIC FATHA
ي → U+064A → ARABIC LETTER YEH	ي → U+064A → ARABIC LETTER YEH
ْ → U+0652 → ARABIC SUKUN	ْ → U+0652 → ARABIC SUKUN
س → U+0633 → ARABIC LETTER SEEN	س → U+0633 → ARABIC LETTER SEEN
ً → U+064E → ARABIC FATHA	ً → U+064E → ARABIC FATHA
→ U+0020 → SPACE	→ U+0020 → SPACE
م → U+0645 → ARABIC LETTER MEEM	م → U+0645 → ARABIC LETTER MEEM
ً → U+064E → ARABIC FATHA	ً → U+064E → ARABIC FATHA
ن → U+0646 → ARABIC LETTER NOON	ن → U+0646 → ARABIC LETTER NOON
ْ → U+0652 → ARABIC SUKUN	ْ → U+0652 → ARABIC SUKUN
ع → U+0639 → ARABIC LETTER AIN	ع → U+0639 → ARABIC LETTER AIN
ُ → U+064F → ARABIC DAMMA	ُ → U+064F → ARABIC DAMMA
ه → U+0647 → ARABIC LETTER HEH	ه → U+0647 → ARABIC LETTER HEH
ا → U+0627 → ARABIC LETTER ALEF	ً → U+064E → ARABIC FATHA
→ U+0020 → SPACE	ا → U+0627 → ARABIC LETTER ALEF
ب → U+0628 → ARABIC LETTER BEH	→ U+0020 → SPACE
ِ → U+0650 → ARABIC KASRA	ب → U+0628 → ARABIC LETTER BEH
ا → U+0627 → ARABIC LETTER ALEF	ِ → U+0650 → ARABIC KASRA
ل → U+0644 → ARABIC LETTER LAM	ا → U+0627 → ARABIC LETTER ALEF
ْ → U+0652 → ARABIC SUKUN	ل → U+0644 → ARABIC LETTER LAM
ك → U+0643 → ARABIC LETTER KAF	ْ → U+0652 → ARABIC SUKUN
ا → U+0627 → ARABIC LETTER ALEF	ك → U+0643 → ARABIC LETTER KAF
م → U+0645 → ARABIC LETTER MEEM	ً → U+064E → ARABIC FATHA
ِ → U+0650 → ARABIC KASRA	ا → U+0627 → ARABIC LETTER ALEF
ل → U+0644 → ARABIC LETTER LAM	م → U+0645 → ARABIC LETTER MEEM
ِ → U+0650 → ARABIC KASRA	ِ → U+0650 → ARABIC KASRA
	ل → U+0644 → ARABIC LETTER LAM
	ِ → U+0650 → ARABIC KASRA
	(→ U+0028 → LEFT PARENTHESIS
Number of characters: 177	**Number of characters: 192**

In summary, the Pearson correlation coefficient was used to calculate the correlation for understanding the linear relationship between different error rates across Arabic diacritization tools (see Figure 4). The results indicated that WER and DER measure different aspects of model performance and do not strongly correlate with each other. WER is relevant to downstream tasks such as MT and text-to-speech, where word-level accuracy is critical, while DER provides assessment of diacritization quality, making it suitable for evaluating ATD models.These findings highlight the importance of evaluating both WER and DER separately, and demonstrate how correlation analysis can guide the selection or improvement of diacritization models.

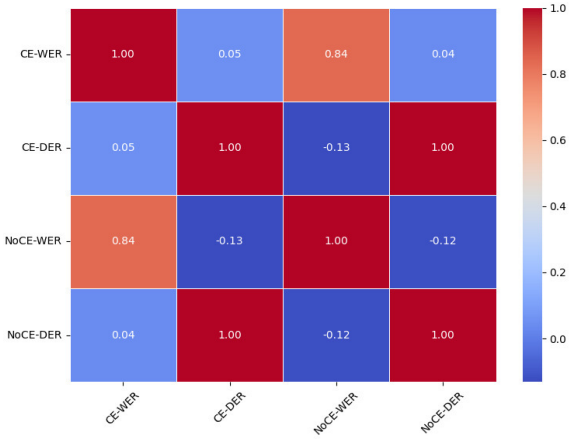


Fig 4. Pearson Correlation Matrix between WER and DER of ATD

V. EMPIRICAL EVALUATION OF FINE-TUNED TRANSFORMER MODELS FOR DIACRITIZED ARABIC-ENGLISH STS

In our proposed framework, a transformer-based approach is employed to effectively address the challenges of cross-lingual STS, with a particular focus on Arabic-English sentence pairs. The Arabic input is diacritized by Shakkelha prior to training, enabling the model to better disambiguate meanings and improve alignment with corresponding English texts. By incorporating fine-tuning into our approach, the transformer encoder is adapted to the specific semantic and linguistic characteristics of the dataset. In order to select the appropriate XLM-R variant for our model, we conducted a comparative evaluation of three XLMR-based transformer encoders by empirical experimentation. The chosen models were tested on an Arabic-English STS dataset, STS-Benchmark has 250 pairs and introduced in 2017, with Arabic sentences diacritized before and after training the models. This STS benchmark contains a variety of semantic similarity levels showing moderate similarity label balance. Then, the score in this dataset was normalized from [0 – 5] to [0 – 1] for fine-tuning. Normalizing STS scores to the [0 – 1] range is standard preparation in sentence embedding models with STS task and is fundamental for correct training and

evaluation. The dataset was randomly divided using the train_test_split function from scikit-learn, with 80% of the sentence pairs used for training and the remaining 20% reserved for development (dev) validation. This split was applied to the pre-processed and normalized STS-labelled CSV data to ensure a clear separation between training and evaluation samples. The results are summarised in TABLE VIII, which presents the Pearson correlation coefficients computed on the development (dev) set before and after fine-tuning. It shows the effect of fine-tuning five different multilingual sentence encoders on a SemEval Semantic Textual Similarity task using Arabic text that has diacritization by ATD tool. The metric used to evaluate the effectiveness of fine-tuning is the Pearson correlation coefficient (r) between the predicted similarity scores and the gold standard labels, along with p-values to test statistical significance. Pearson correlation measures the degree of linear association between the similarity scores produced by the model and the human-annotated gold-standard similarity scores. Evaluation on the dev set, rather than the training set, ensures an unbiased estimation of the model's generalization performance on unseen data and avoid the risk of overfitting or performance inflation due to exposure to training data. In addition to the correlation coefficients, corresponding p-values are reported to assess the statistical significance of the observed correlations. A low p-value (significant if $p < 0.05$) indicates that the correlation is unlikely to have arisen by random chance under the null hypothesis of no association, thereby supporting the reliability and validity of the model's semantic alignment with human judgments. Based on these results, model 1 & 2 are statistically significant because the null hypothesis (Fine-tuning the encoder on diacritized Arabic text does not affect performance on the STS benchmark) is rejected. Therefore, the paraphrase-xlm-r-multilingual-v1 model that achieved the highest Pearson correlation scores both before and after fine-tuning was selected for use in AraXLM. Fine-tuning this encoder model on diacritized Arabic text enhances semantic similarity detection, particularly when applied to paraphrase pretrained multilingual encoder variant.

TABLE VIII.
EXPERIMENTAL RESULTS OF PEARSON CORRELATION PRE- AND POST-FINE-TUNING ON DEV ATD SET

#	Model	Pearson Correlation (Before)	P-Value	Pearson Correlation (After)	P-Value
1	sentence-transformers/paraphrase-xlm-r-multilingual-v1	32.32%	0.0220	37.36%	0.0070
2	sentence-transformers/xlm-r-distilroberta-base-paraphrase-v1	32.32%	0.0220	37.03%	0.0081
3	sentence-transformers/stsb-xlm-r-multilingual	14.66%	0.310	18.96%	0.187
4	xlm-roberta-base	-1.71%	0.90	9.16%	0.53
5	xlm-roberta-large	-24.32%	0.088	5.94%	0.68

VI. COMPARISON WITH OTHER APPROACHES IN STS TRACK2

The SemEval-2017 STS Cross-lingual Arabic-English task provided benchmark reference systems against which new approaches can be evaluated [37]. The ECNU [38] and BIT [39] systems reported Pearson correlation scores of 0.74 and 0.69 , respectively, on the official test set[3]. In comparison, the fine-tuned variant incorporating diacritized Arabic text demonstrated competitive performance. The evaluated textual inputs included:

- The original English sentence from STS (En)
- The original Arabic sentence from STS (Ar)
- The Arabic-to-English translation of Ar (Ar-TrE)
- The diacritized Arabic sentence (ArATD)
- The translation of ArATD into English (ArATD-TrE)

ArATD_TrE_CosineSimilarity and FAISS IP (Inner Product) achieved the highest Pearson correlation scores of 0.78 and 0.77, respectively, outperformed both ECNU and BIT baselines (see TABLE IX). Additionally, Ar_TrE_CosSim exhibited strong performance (Spearman: 0.80, Pearson: 0.77), closely aligning with the top-performing metrics. These results confirm that leveraging translated and diacritized Arabic variants significantly enhances cross-lingual semantic similarity detection.

TABLE IX.
COMPARATIVE EVALUATION BASED ON STS DATASET

Metric	Spearman Correlation	Pearson Correlation
ECNU	NA	0.74
BIT	NA	0.69
IP_Score_FAISS	0.81	0.77
ArabicSentence_CosSim	0.75	0.72
Ar_TrE_CosSim	0.8	0.77
ArATD_CosSim	0.35	0.33
ArATD_TrE_CosSim	0.81	0.78

VII. DISCUSSION OF THE RESULTS AND CONCLUSION

This study aimed to evaluate the performance of six ATD tools for Arabic using established metrics, WER, DER, and BLEU. In response to research sub-question 1: What is the most effective ATD tool for integration into our proposed framework (AraXLM)?

Our experiments demonstrated that Shakkelha, a deep learning-based tool designed for Modern Standard Arabic, outperformed other models across key evaluation metrics. It achieved high character-level BLEU scores and low DER when case endings were excluded, making it the most effective candidate for integration into the AraXLM framework.

In response to research sub-question 2: What issues are identified when conducting our experimental study, particularly for Arabic?

Several challenges were noted, including:

- A significant increase in WER (up to 16 percentage points) when syntactic case endings (CE) were restored.

[3] https://docs.google.com/spreadsheets/d/1a5ZNg5IqKnBLaNKHMyVvmAn2_xc79FqwbXLL5z4ABAk/edit?gid=0#gid=0

- Difficulties in tokenising diacritized Arabic, due to non-standard outputs (e.g., merged diacritics like shaddah with sukun, or irregular representations of tanwīn).
- Variability in how tools apply full diacritization, sometimes leading to redundant or incorrect diacritic marks.

These issues highlight both linguistic and technical limitations in current ATD models when applied to Arabic morphology and syntax.

The results highlight that fully diacritizing every character is not always good and may introduce noise or errors, thereby skewing evaluation metrics and negatively impacting downstream NLP applications. The merging of diacritics (e.g., shaddah with sukun) and inconsistent representations (e.g., tanwīn as repeated fatha) complicate tokenization and evaluation.

These findings emphasize the requirement of standardised evaluation methodologies and tokenization practices that are sensitive to Arabic's unique linguistic structure. The implementation of robust ATD models is vital not only for accurate diacritization but also as a foundational layer in advanced NLP tasks such as semantic search and plagiarism detection across languages. The study also supports the integration of linguistic knowledge with deep learning models, showing the way for adaptable and accurate transformer-based systems in low-resource languages, such as Arabic.

Despite the insights gained from the results, it was limited by the lack of gold-standard diacritized datasets and the narrow range of tools, many of which were trained on specific text domains. Additionally, integrating these tools into Python required adjustments due to the complexity of Arabic script and morphology.

In conclusion, this study highlights the importance of accurate ATD for improving ANLP,and supports the integration of Shakkelha as the most effective tool within the proposed AraXLM framework. The results demonstrate its consistent performance across key evaluation metrics, while exposing limitations in current tokenisation and evaluation methods. These findings underscore the need for linguistically informed, standardised pre and post-processing approaches tailored to the structural complexity of the Arabic language.

In future work, the next phase investigate how diacritization affects cross-lingual semantic similarity in AraXLM. We plan to integrate ATD outputs within our proposed AraXLM framework to enhance performance in cross-lingual semantic similarity tasks with FAISS vector-based. This integration aims to improve accuracy, contextual adaptability, and the robustness of transfer learning systems in Arabic-English applications.

REFERENCES

[1] M. Elyaakoubi and A. Lazrek, "Justify just or just justify," *Journal of Electronic Publishing,* vol. 13, no. 1, 2010. doi: 10.3998/3336451.0013.105.

[2] R. Rjeily, Cultural Connectives: Bridging the Latin and Arabic Alphabets, vol. 1. Brooklyn, NY: Mark Batty Publisher, 2021.

[3] M. Hssini and A. Lazrek, "Design of Arabic Diacritical Marks," *International Journal of Computer Science*, vol. 8, no. 3, May 2011.

[4] M. Maamouri, A. Bies, and S. Kulick, 'Diacritization: A Challenge to Arabic Treebank Annotation and Parsing', *the International Conference on the Challenge of Arabic for NLP/MT* , pp. 35-47, 2006.

[5] S. Alzahrani, "Arabic plagiarism detection using word correlation in N-Grams with K-overlapping approach," Taif, 2015. https://ceur-ws.org/Vol-1587/T5-2.pdf

[6] E. M. B. Nagoudi et al., "2L-APD: A two-level plagiarism detection system for Arabic documents," *Cybernetics and Information Technologies,* vol. 18, no. 1, pp. 124–138, 2018. doi: 10.2478/cait-2018-0011.

[7] B. Akanksha et al., "A survey on plagiarism detection," *International Journal of Computer Applications*, vol. 10, no. 8, pp. 2359–2365, 2017. http://www.ripublication.com

[8] M. F. Akan et al., "An analysis of Arabic-English translation: Problems and prospects," *Advances in Language and Literary Studies,* vol. 10, no. 1, p. 58, Feb. 2019. doi: 10.7575/aiac.alls.v.10n.1p.58.

[9] M. Alshehri, N. Beloff, and M. White, "AraXLM: New XLM-RoBERTa based method for plagiarism detection in Arabic text," *in Intelligent Computing*, K. Arai, Ed., Cham: Springer, 2024, pp. 81–96. doi: 10.1007/978-3-031-62277-9_6

[10] M. M. Elmallah et al., "Arabic diacritization using morphologically informed character-level model," *in Proc. LREC-COLING* 2024, Torino, Italy: ELRA and ICCL, May 2024, pp. 1446–1454. https://aclanthology.org/2024.lrec-main.128/

[11] K. Shaalan and Khaled, "Rule-based approach in Arabic natural language processing," *International Journal on Information and Communication Technologies*, vol. 3, p. 11, May 2010.

[12] A. Chennoufi and A. Mazroui, "Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences," *Journal of King Saud University* - Computer and Information Sciences, vol. 29, no. 2, pp. 156–163, 2017. doi: 10.1016/j.jksuci.2016.06.004.

[13] M. Mézard and J.-P. Nadal, "Learning in feedforward layered networks: the tiling algorithm," *Journal of Physics A*, vol. 22, pp. 2191–2203, 1989. https://api.semanticscholar.org/CorpusID:44826720

[14] W. Almanaseer et al., "A deep belief network classification approach for automatic diacritization of Arabic text," *Applied Sciences,* vol. 11, no. 11, 2021. doi: 10.3390/app11115228.

[15] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[16] Y. Alalawi et al., "A CNN-based Arabic diacritic symbol recognition system using domain adaptation," i*n Proc. 8th Int. Conf. Sustainable Information Engineering and Technology (SIET*), New York, USA: ACM, 2023, pp. 23–32. doi: 10.1145/3626641.3627212.

[17] H. Hewamalage et al., "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021. doi: 10.1016/j.ijforecast.2020.06.008.

[18] Y. Belinkov and J. Glass, "Arabic diacritization with recurrent neural networks," in Proc. *EMNLP 2015, Lisbon, Portugal: ACL*, Sep. 2015, pp. 2281–2285. doi: 10.18653/v1/D15-1274.

[19] A. Vaswani et al., "Attention is all you need," *in Proc. NeurIPS* 2017, 2017. http://arxiv.org/abs/1706.03762

[20] A. Assad et al., "Transformer-based automatic Arabic text diacritization," *Sustainable Engineering and Innovation*, vol. 6, no. 2, pp. 285–296, Nov. 2024. doi: 10.37868/sei.v6i2.id305.

[21] Y. Bengio et al., "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar. 1994. doi: 10.1109/72.279181.

[22] A. Gillioz, J. Casas, E. Mugellini and O. Abou Khaled, "Overview of the Transformer-based Models for NLP Tasks," Proceedings of the Federated Conference on Computer Science and Information Systems, vol. 21, pp. 179–183, 2020, doi: 10.15439/2020F20.

[23] R. Al-Sabri and J. Gao, "LAMAD: A linguistic attentional model for Arabic text diacritization," *in Findings of the Association for Computational Linguistics: EMNLP 202*1, Punta Cana, Dominican Republic: ACL, Nov. 2021, pp. 3757–3764. doi: 10.18653/v1/2021.findings-emnlp.317.

[24] M. Al-Badrashiny et al., "A layered language model based hybrid approach to automatic full diacritization of Arabic," *in Proc. 3rd Arabic*

NLP Workshop, Valencia, Spain: ACL, Apr. 2017, pp. 177–184. doi: 10.18653/v1/W17-1321.

[25] H. Alaqel and K. El Hindi, "Improving diacritical Arabic speech recognition: Transformer-based models with transfer learning and hybrid data augmentation," *Information, vol.* 16, no. 3, 2025. doi: 10.3390/info16030161.

[26] O. Obeid et al., "CAMeL Tools: An open source Python toolkit for Arabic NLP," 2020. http://qatsdemo.cloudapp.net/farasa/

[27] A. Abdelali et al., "Farasa: A fast and furious segmenter for Arabic," *in Proc. NAACL Demonstrations*, San Diego, CA: ACL, Jun. 2016, pp. 11–16. doi: 10.18653/v1/N16-3003.

[28] F. Alasmary et al., "CATT: Character-based Arabic Tashkeel Transformer," *arXiv preprint*, vol. abs/2407.03236, 2024. https://api.semanticscholar.org/CorpusID:270924323

[29] B. Al-Rfooh et al., "Fine-Tashkeel: Fine-tuning byte-level models for accurate Arabic text diacritization," *in Proc. IEEE JEEIT 2023*, pp. 199–204, 2023. https://api.semanticscholar.org/CorpusID:257767345

[30] B. M. King, "Analysis of variance," *in International Encyclopedia of Education,* 3rd ed., Jan. 2009, pp. 32–36. doi: 10.1016/B978-0-08-044894-7.01306-3.

[31] A. Lazrek, "Arabic mathematical notation," *National Institute of Standards and Technology,* USA, 2006. https://www.w3.org/TR/2006/NOTE-arabic-math-20060131/

[32] K. Darwish et al., "Arabic diacritization: Stats, rules, and hacks," *in Proc. 3rd Arabic NLP Workshop,* Valencia, Spain: ACL, Apr. 2017, pp. 9–17. doi: 10.18653/v1/W17-1302.

[33] A. Fadel et al., "Neural Arabic text diacritization: State of the art results and a novel approach for Arabic NLP downstream tasks," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 21, no. 1, Jan. 2022. doi: 10.1145/3470849.

[34] O. Obeid et al., "CAMeL Tools: An open source Python toolkit for Arabic NLP," *in Proc. LREC 2020, Marseille, France: ELRA,* May 2020, pp. 7022–7032. https://aclanthology.org/2020.lrec-1.868/

[35] T. Zerrouki, "Towards an open platform for Arabic language processing," 2020. doi: 10.13140/RG.2.2.29882.82881.

[36] A. Fadel et al., "Neural Arabic text diacritization: State of the art results and a novel approach for machine translation," *in Proc. 6th Workshop on Asian Translation,* Hong Kong, China: ACL, Nov. 2019, pp. 215–225. doi: 10.18653/v1/D19-5229.

[37] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity – Multilingual and Cross-lingual Focused Evaluation," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, Aug. 2017, pp. 1–14.

[38] Y. Tian, Y. Song, H. Xia, Y. Li, and Q. Zhang, "ECNU at SemEval-2017 Task 1: Leverage Kernel-Based Traditional NLP Features and Distributed Word Representations for Semantic Textual Similarity Estimation," in Proc. 11th Int. Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, Aug. 2017, pp. 125–131. https://aclanthology.org/S17-2015

[39] H. Wu, H. Huang, P. Jian, Y. Guo, and C. Su, "BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity," in *Proc. 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Aug. 2017, pp. 77–84. https://aclanthology.org/S17-2007

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 461–470
DOI: 10.15439/2025F4310
ISSN 2300-5963 ACSIS, Vol. 43

Hybrid Approaches for Pneumonia Detection in X-rays: Combining CNNs and ML Classifiers

Gabriele S. Araújo
0000-0003-1143-507X
Center for Technological Sciences,
State University of Maranhão,
São Luís/Maranhão, Brazil
gabriele.20231002966@aluno.uema.br

Omar Andres C. Cortes
0000-0002-5805-2490
Department of Computing,
Federal Institute of Maranhão,
São Luís/Maranhão, Brazil
omar@ifma.edu.br

Olaf Reinhold
0000-0003-1977-1641
University of Cooperative Education Saxony
Social CRM Research Center,
Riesa/Leipzig, Germany
olaf.reinhold@dhsn.de

Fabio M. F. Lobato
Orcid: 0000-0002-6282-0368
Institute of Engineering and Geosciences,
Federal University of Western Pará,
Santarém/Pará, Brazil
fabio.lobato@ufopa.edu.br

***Abstract*—Pneumonia is a disease that impacts millions of people worldwide, and X-ray image detection is one of the primary diagnostic tools used. This study presents a hybrid diagnostic approach combining Convolutional Neural Networks (CNNs) with traditional classifiers, namely Random Forest (RF) and Support Vector Machine (SVM), to detect pneumonia from chest X-ray images. Features were extracted from MobileNetV2, VGG16, and EfficientNetB0 and used to train RF and SVM models with hyperparameter tuning via GridSearchCV. Ensemble models were also explored: (i) CNNs + RF, (ii) CNNs + SVM, and (iii) RF + SVM. Experiments were conducted using a public pediatric dataset (5.856 X-rays) with stratified k-fold cross-validation and data augmentation. CNNs + RF achieved the highest 0.977 AUC and 91.9% accuracy, while individual models like VGG16 showed competitive performance (91.8% accuracy, 0.969 AUC). Results were statistically validated and showed strong potential for clinical support, particularly in settings with limited resources. In future work, we propose extending the approach to multiclass classification and refining model optimization strategies.**

***Index Terms*—Deep Learning, Ensemble Learning, Medical Imaging**

I. Introduction

PNEUMONIA is one of the leading causes of death globally. According to the World Health Organization (WHO)[1], over 808,000 children under the age of five died from pneumonia in 2017, which accounted for 15% of all deaths in that age group. Additionally, data from the 2024 Global Burden of Disease study conducted by the Institute for Health Metrics and Evaluation[2] revealed that the highest mortality rates from pneumonia in 2019 were among individuals aged 70 and older. In Brazil, prolonged droughts and recent Amazon wildfires have worsened air quality and increased respiratory illnesses, especially among children, highlighting the need for urgent public health action, as confirmed by a 2024 UNICEF report[3] [1].

X-ray imaging has become a standard method for detecting pneumonia. However, manual interpretation of these images requires time and expertise from health professionals [2]. In addition, there has been a significant increase in the costs associated with diagnostic imaging, especially in universal healthcare systems such as the Brazilian or the British ones, where resources are often limited and unevenly distributed [3]. In this context, deep learning methods have shown potential for optimizing the diagnostic process, reducing costs, and increasing efficiency and accessibility [4].

Sharma et al. [4] systematically review pneumonia detection studies using chest X-ray images, presenting models based on Convolutional Neural Networks (CNNs), pre-trained, and Ensemble models. The latter, which combines the predictions of different architectures, has outperformed individual models in metrics such as accuracy, sensitivity, and specificity. Each implementation has advantages and disadvantages, depending on the data characteristics and the number of samples. Models based on CNNs can extract relevant features directly from X-ray images, but face challenges related to the need for large volumes of data to avoid overfitting [5]. Pre-trained models have shown high efficiency due to transfer learning, reducing training time, and using prior knowledge. Ensemble models excelled, achieving an accuracy of up to 99.61% in some studies [6], due to the integration of complementary architectures [4], [7].

Despite these advances, few studies have explored how traditional classifiers, such as Random Forest (RF) or Support

[1]https://www.who.int/health-topics/pneumonia
[2]https://ourworldindata.org/pneumonia
[3]https://bit.ly/4jzNPXy

Vector Machines (SVM), perform with features extracted from modern CNNs. This combination may offer benefits in scenarios with limited data or noisy inputs [8], [9]. Moreover, many works do not assess the statistical significance of model comparisons or provide publicly available code to support reproducibility, issues increasingly emphasized in recent reviews [10].

Inspired by recent advances in deep learning and hybrid modeling, this study proposes a classification pipeline that combines CNN-based feature extraction (*e.g.*, MobileNetV2, EfficientNetB0, and VGG16) with traditional machine learning classifiers, namely RF and SVM. Additionally, ensemble models that aggregate predictions from multiple classifiers are evaluated. The objective is to assess whether conventional classifiers can outperform end-to-end deep models when leveraging CNN-derived features and whether combining predictions improves diagnostic accuracy. The approach is applied to the publicly available pneumonia X-ray dataset by [11], with experiments incorporating hyperparameter tuning, data augmentation, and statistical validation to ensure reproducibility.

The results demonstrate the potential of hybrid pipelines in improving diagnostic performance, particularly their applicability in low-resource clinical settings. By integrating deep learning with lightweight classifiers, the approach seeks to balance diagnostic efficacy with computational efficiency, contributing to scalable and affordable diagnostic tools for real-world healthcare settings.

The remainder of this paper is organized as follows: The related works are discussed in Section II. The proposed method is described in Section III, and results are presented in Section IV. Finally, conclusions, limitations, and directions for future work are given in Section V.

II. Related Works

General studies in the field of medical diagnostics show that image analysis, such as X-rays, is one of the main tools for detecting diseases, including pneumonia [12], [6], [7], [13], [9]. According to [4], the early diagnosis of pneumonia via chest images is widely studied, with approaches ranging from CNNs created from scratch to pre-trained models (*e.g.*, VGG16, VGG19, DenseNet, ResNet, and Inception) and ensemble techniques. The study highlights the benefits of deep learning for feature extraction and classification, emphasizing the relevance of combining deep feature extractors with traditional machine learning classifiers such as SVM, KNN, and RF. Additionally, the authors discuss challenges such as limited datasets, the need for data augmentation, the absence of standardized comparisons across models, and insufficient studies on hyperparameter tuning. Although the reviewed models have shown promising results, they suggest that further improvements are possible, primarily through more diverse datasets and integrating multiple architectures and classification strategies.

Diniz et al. [12] describes a novel algorithm for diagnosing breast cancer in ultrasound images, called EfficientNet Ensemble, evaluating many image preprocessing methods and data augmentation. The proposed method can extract Regions of Interest (ROIs), apply guided and median filters, and train three EfficientNet variants (B0, B1, B2) combined in an ensemble with majority voting. After adjusting the hyperparameters (learning rate, batch size, and epochs), the method achieved 96.67% accuracy, outperforming the individual EfficientNetB0 (86.57%) and several other related studies. The architecture demonstrated the benefits of integrating preprocessing with CNN ensembles.

Munzlinger, Yepes, and Rieder [14] presented a detector of COVID-19, pneumonia, and tuberculosis in chest X-rays. The system is based on the ResNet-50 architecture, employing transfer learning. After model adjustments, custom dense layers were added, and the last five layers were trained while retaining the pre-trained weights. The model was trained on a dataset of 7,097 images collected from multiple open repositories (Kaggle, Vindr-CXR, FIPS), achieving an accuracy of 89%, slightly below the predefined goal of 90%, due to computational constraints (Google Colab limit of 100 epochs).

Toğaçar et al. [6] proposed a hybrid classification model combining CNN feature extraction (VGG16, VGG19, and AlexNet) with traditional machine learning classifiers such as SVM and LDA. The features extracted from the final fully connected layers were reduced using the mRMR (minimum Redundancy Maximum Relevance) method to 100 per model, and then concatenated. Their approach achieved an accuracy of up to 99.41%, showing that combining CNN-derived features with classical classifiers can yield highly competitive results. This highlights the potential of hybrid pipelines and reinforces the motivation for exploring CNN + ML classifier combinations in this study.

Varshni et al. [9], pre-trained CNN models such as DenseNet-169 are evaluated for diagnosing pneumonia in chest X-rays, using feature extraction and supervised classifiers, especially SVM (kernel RBF). The ChestX-ray14 database contained 2,862 images balanced between normal and pneumonia. The method included image resizing and feature extraction with DenseNet-169, which showed superior performance with an AUC of 0.8002 after hyperparameter optimization. The proposed model outperforms previous studies, demonstrating the effectiveness of DenseNet-169 combined with SVM for accurate and efficient diagnosis in medical images.

Akgundogdu [8], the model based on combining the 2D Discrete Wavelet Transform (2D DWT) for extracting features from chest X-ray images and using RF as a classifier for detecting pneumonia, is introduced. The dataset contained 5,856 images (4,273 with pneumonia and 1,583 normal). The method uses 24 features extracted from the wavelet sub-bands (LL, LH, HL, HH), including minimum, maximum, mean, standard deviation, variance, and third-order moment values. Their results showed that the model achieved an accuracy of 97.11%, sensitivity of 91.79%, and specificity of 99.09%, with an area under the curve (AUC) of 0.99. The study highlights the efficiency of the proposed model in terms of speed and accuracy, making it a promising and less complex alternative

to deep learning methods for diagnosing pneumonia. In future work, the authors mentioned that deep learning methods should be incorporated for comparison and improvement.

Mabrouk and Dias Redondo [7] proposed an Ensemble Learning (EL) model, combining DenseNet169, MobileNetV2, and Vision Transformer (VIT), adjusted with transfer learning on the ImageNet database and trained on pneumonia chest X-rays. The method outperformed the state-of-the-art, achieving 93.91% accuracy and 93.88% F1-score. The approach combines the features extracted by the models with pooling, normalization, and regularization techniques, reducing the risk of overfitting. In addition, the integration of Vision Transformer stands out for its ability to capture relationships between image patches, complementing the limitations of traditional CNNs. Despite the success, challenges such as the hyperparameter tuning and more significant variation in the data are mentioned. Moreover, strategies for weight assignments are interesting to investigate.

Recent advances in deep learning for medical images have also explored the integration of attention mechanisms and Transformer-based architectures. These approaches, by allowing models to focus on the most salient regions of an image, have shown promising results in tasks such as chest X-ray classification with improved interpretability and performance by dynamically weighting the importance of features. For example, [15] proposed an attention-based deep learning model for medical image classification, demonstrating its effectiveness on various medical datasets. [16] presented an attention-driven Spatial Transformer Network (STERN) for abnormality detection in chest X-ray images, highlighting its ability to dynamically scale and align images to maximize classifier performance by selecting the thorax and eliminating artifacts, achieving a mean AUC of 85.67% on the CheXpert dataset. Similarly, [17] introduced ResfEANet, a novel architecture combining ResNet with an external attention mechanism for tuberculosis diagnosis from chest X-rays, achieving high accuracy (97.59%) and sensitivity (100%) even without pre-training on a multi-source TB dataset. While this study focuses on established CNN backbones and hybrid classifiers, the evolving landscape of deep learning presents further avenues for investigation.

The reviewed literature demonstrates that combining CNN-based models [9], [14], traditional classifiers such as RF and SVM [8], [6], and ensemble strategies [7] can enhance pneumonia detection from chest X-rays. Despite high reported accuracies, many studies rely on limited datasets, lack external validation, underexplored hyperparameter tuning, and often restrict classification to single-view X-rays. Additionally, some works are constrained by limited computational resources, which affects model complexity and training duration. While DWT + RF [8] offer simplicity and high accuracy (97.11%), they depend on handcrafted features and lack scalability. This motivates exploring deep feature extractors such as VGG16, widely used in medical imaging tasks due to its simplicity and effectiveness [11], alongside more modern CNNs like MobileNetV2 and EfficientNet. A systematic hyperparameter search (GridSearchCV), k-fold cross-validation, and data augmentation techniques are employed to improve robustness and generalization. Similar strategies have proven valuable in other domains, such as image captioning, where the effectiveness of visual representation is highly dependent on the backbone CNN used [18]. In line with open science practices, the experimental framework and results are publicly available to support reproducibility and enable adaptation to other medical imaging modalities.

III. Material and Methods

This work proposes a systematic approach for diagnosing pneumonia in chest X-rays using different machine-learning models selected from existing literature. The experimental framework includes the definition of a computational environment for the experiments, the use of a dataset widely referenced in the literature, the application of pre-processing techniques to improve data quality, the implementation of a classifier architecture based on deep and ensemble learning models, and a validation process using K-Fold Cross-Validation. The following subsections detail the experimental framework, the dataset used, the pre-processing procedures applied, the model architectures, and the training and evaluation strategies.

A. Setup

The experiments used Google Colab Pro+, leveraging a Google Compute Engine backend with an NVIDIA T4 GPU. The system had 51 GB of RAM and 15 GB of GPU memory and ran Python scripts in version 3. This cloud-based environment provided efficient computational resources, enabling faster training and experimentation.

The main libraries and tools used include: `TensorFlow`, `NumPy`, `Scikit-learn`, `Matplotlib`, `Seaborn` and `Pandas`. The pre-trained CNNs were tuned using the TensorFlow framework, while the RF and SVM model was implemented with Scikit-learn. In addition, data augmentation techniques were applied using `ImageDataGenerator` from TensorFlow/Keras, and `EarlyStopping` callbacks were employed to prevent overfitting during CNN training [19], [4]. The corresponding subsections detail the hyperparameter tuning, including specific configurations for the CNN networks, classifiers, and the ensemble methods.

B. Dataset

The dataset used in this work consists of chest X-ray images, made publicly available[4] and presented initially by [11], [20]. This dataset is widely used in research to detect pneumonia [4]. It contains 5,857 images, which are distributed into two main sets: a training set with 5,233 images (1,349 Normal and 3,884 Pneumonia), and a fixed test set with 624 images (234 Normal and 390 Pneumonia). Figure 1 illustrates representative examples of both classes. A chest X-ray of a healthy patient shows clear lungs with no areas of abnormal opacification. In contrast, pneumonia usually shows focal lobar

[4]https://data.mendeley.com/datasets/rscbjbr9sj/3

consolidation (bacterial pneumonia) or manifests with a more diffuse "interstitial" pattern in both lungs (viral pneumonia), as described for [11].

The images are in JPEG format and have a resolution suitable for analysis by deep learning models. All X-rays were selected from retrospective cohorts of pediatric patients aged between one and five years from the Guangzhou Women and Children's Medical Center in Guangzhou, China. The X-ray examinations were performed as part of routine clinical care, and the image diagnoses were evaluated by two medical specialists, with the test set evaluation subsequently reviewed by a third specialist to mitigate possible classification errors [11].

C. Pre-Processing

To ensure the reliability of the experiments and improve model generalization, a 5-fold Stratified Cross-Validation was used. This technique preserves the class distribution across folds, allowing each sample to be used for training and validation in different iterations while maintaining a consistent proportion of `NORMAL` and `PNEUMONIA` cases [8]. The test set, comprising 624 images (234 Normal and 390 Pneumonia), was held out from the total dataset and remained fixed throughout the process to enable standardized performance evaluation across models. [4].

All images were resized to 224x224 pixels, a standard input format for pre-trained architectures like MobileNetV2, EfficientNetB0, and VGG16. This resolution balances computational efficiency and preserves key visual patterns critical for diagnosis. Pixel values were normalized to the range [0,1] to facilitate training convergence.

For the training set, *data augmentation* techniques were applied to mitigate overfitting, increase data diversity, and improve the generalization capacity of the models [19]. These augmentations simulate real-world variations observed in X-ray images, such as patient positioning, lighting, and equipment differences. These transformations help models generalize better to opacities or consolidations located in various lung regions [21], [13], [6]. The selected augmentations are summarized in Table I and follow recommendations from recent literature [6], [13], [4].

TABLE I
DATA AUGMENTATION TECHNIQUES APPLIED TO THE TRAINING SET

Transformation	**Values**
Rotations (`rotation_range`)	30°
Horizontal Shifts (`width_shift_range`)	30%
Vertical Shifts (`height_shift_range`)	30%
Zoom (`zoom_range`)	20%
Brightness Adjustments (`brightness_range`)	[0.8, 1.2]
Shear (`shear_range`)	20°
Horizontal Flip (`horizontal_flip`)	`True`

As presented in Table I, these transformations which include rotation, shear, displacement, brightness adjustments, and horizontal flipping, are clinically relevant because they improve the model's ability to detect pneumonia manifestations that may appear in different lung zones or vary due to patient orientation or image quality [4].

D. Model Architecture and Training

This study proposes a hybrid architecture that combines deep feature extraction using CNNs with traditional machine learning classifiers, represented by the state-of-the-art discussed in Section II. The overall architecture of the proposed method is summarized in Figure 2.

Six classification strategies were tested: three end-to-end CNNs (MobileNetV2, EfficientNetB0, VGG16), two traditional classifiers (RF, SVM) trained on CNN-derived features, and ensemble configurations. The goal is to evaluate whether classical classifiers can outperform deep models when provided with expressive representations and whether combining several models yields additional performance gains, considering computational limitations.The CNNs were implemented using *transfer learning*, a technique shown to be efficient in medical image classification tasks [4].

MobileNetV2: The first model is an efficient architecture for devices with limited resources. It employs separable convolutions (*depthwise* and *pointwise*) to reduce computational cost and model size, as described by [22]. The MobileNetV2 is optimized with layers of batch normalization (*ReLU*), activation (*ReLU*), and global pooling (*pooling global*) before the fully connected layer, which makes it suitable for *transfer learning* in specific tasks [4], [7].

EfficientNetB0: This model features a scalable *design*, combining efficiency and superior performance by automatically adjusting the network's depth, width, and resolution [23]. It uses transfer learning techniques and is designed to maximize accuracy while reducing the computational resources required, as suggested by [13].

VGG16: Developed by the *Visual Geometry Group* in partnership with DeepMind, expands the AlexNet architecture by using smaller convolutional layers (3x3) and *pooling* layers (2x2) to increase the network's depth [20]. This allows for more detailed feature extraction, excelling in visual classification and transfer learning tasks [4].

The parameter settings for the CNNs were primarily based on the literature reviewed in the previous sections. All models were initialized with ImageNet weights and personalized using transfer learning. The last 20 layers of each base CNN were unfrozen and fine-tuned on the dataset to allow the networks to specialize in domain-specific features. They share the same general training settings, so the architecture was extended with a global average pooling layer, followed by batch normalization, a dense layer with 128 neurons (with *ReLU activation*), and a 50% dropout layer to deal with overfitting. Then, a second dense layer with 64 neurons and *ReLU activation* (called *"feature_dense"*) was added for feature extraction, which feeds into a final sigmoid output neuron for binary classification. This 64-dimensional output was chosen as a balanced representation that retains important information while significantly reducing dimensionality for classical classifiers, thereby managing computational complexity and mitigating

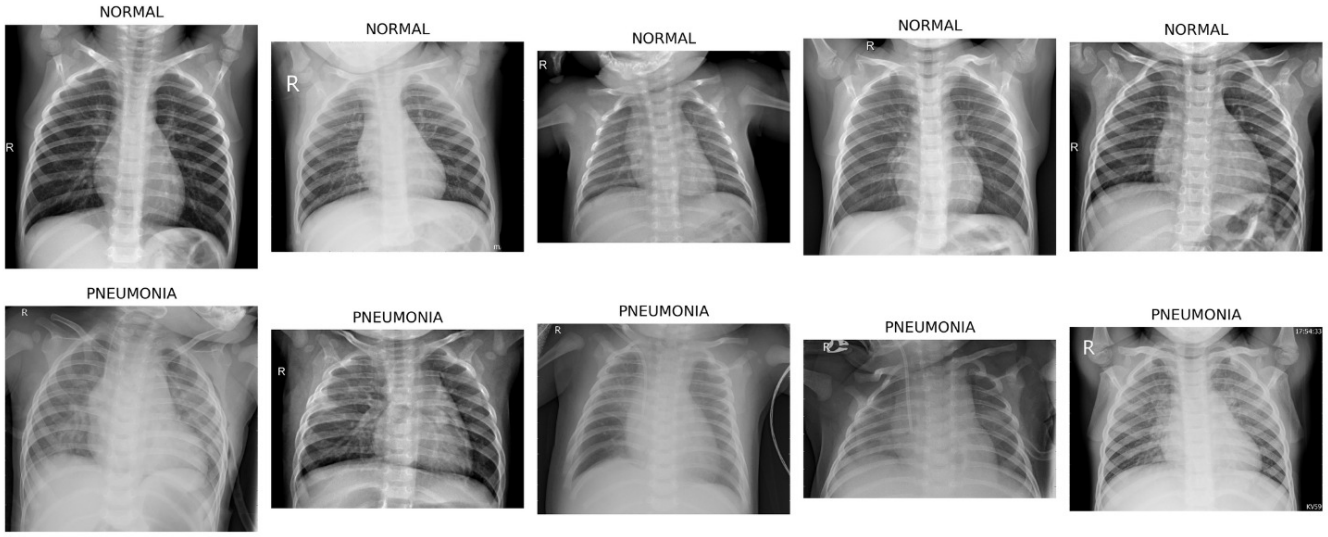


Fig. 1. Examples of chest X-ray images, including normal cases and patients diagnosed with pneumonia

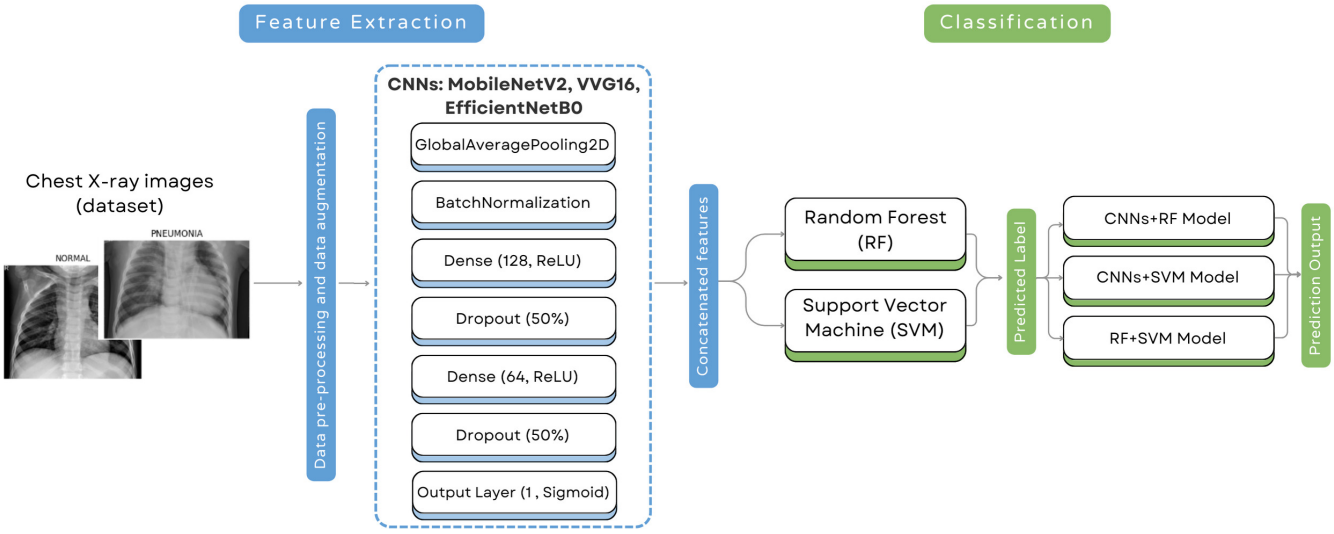


Fig. 2. Proposed hybrid architecture. Features extracted from MobileNetV2, VGG16, and EfficientNetB0 are concatenated and used to train classical classifiers (RF and SVM). The predicted labels are combined into ensemble models: CNNs+RF, CNNs+SVM, and RF+SVM.

the risks of overfitting for subsequent tasks [9], [6]. The models were trained using the Adam optimizer with a learning rate of 1×10^{-5} and the loss function *binary crossentropy*.

The 64-dimensional outputs from the *"feature_dense"* layer were extracted for all images in the training and test sets. These features were concatenated across the CNNs to form comprehensive input vectors for the classical classifiers. This hybrid approach combines the expressive power of deep neural networks with the interpretability and efficiency of traditional machine learning algorithms. The extracted features were subsequently used to train the classifiers described below.

RF: was selected due to its robustness to noise, ability to handle redundant features, and relatively low computational cost. According to previous studies, it can be trained and show significant results using features extracted from the CNNs [8], [4], [9]. This study was implemented with 200 trees (`n_estimators=100`), a maximum depth of 15, and class balancing enabled. It was also evaluated using 5-fold cross-validation with ROC AUC as the scoring function. The importance of the features was also extracted and analyzed.

SVM: is a well-established algorithm known for its strong performance in high-dimensional spaces, especially when dealing with limited data samples [9], [13], [4]. Its hyperparameters were optimized using a GridSearch strategy, and the kernel types were tested, including linear and Radial Basis Functions (RBF). Key parameters such as the regularization term (C) and kernel coefficient (γ) were tuned based on AUC performance, with the following search space:

$C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{\texttt{scale}, \texttt{auto}, 0.01, 0.1\}$. The best parameters were `kernel=poly`, `C=1`, `gamma=scale`, which were evaluated using 5-fold cross-validation, using metrics such as accuracy, balanced accuracy, F1-score (macro and weighted), and AUC were computed per fold.

The ensemble prediction was generated by averaging the predicted probabilities from the individual models. This method aligns with prior studies emphasizing the benefits of combining heterogeneous architectures to increase robustness and compensate for weaknesses of individual models [7], [6]. It is worth mentioning that four test scenarios were conducted during the experiments: two Ensemble CNNs + RF/SVM, created by combining the three proposed CNNs using either RF or SVM; and Ensemble RF + SVM; finally, Ensemble MN + VGG + RF generated by combining MobileNetV2, VGG16, and RF, only for testing due to the performance of EfficientNetB0.

E. Evaluation

For model evaluation, multiple metrics were considered to know: i) *accuracy* - represents the overall proportion of correct predictions and serves as a general indicator of model effectiveness. However, in clinical settings with imbalanced datasets, it may be misleading; ii) *precision* - quantifies the proportion of predicted positive cases that are truly positive. Clinically, a high precision reduces the risk of incorrectly diagnosing healthy individuals with pneumonia; iii) *recall* (also known as sensitivity)- represents the model's ability to identify all true positives correctly. This metric is particularly critical in the medical context, as a low recall could lead to undiagnosed pneumonia cases, potentially resulting in delayed or missed treatment; iv) *F1-score* - the harmonic mean between precision and recall, providing a balanced metric handy for imbalanced data, handy when both false positives and false negatives carry significant clinical implications; and v) the *Area Under the ROC Curve (AUC-ROC)*, which assesses the model's ability to distinguish between positive and negative classes across different decision thresholds - indicates strong discriminative capacity, essential when defining clinically appropriate cut-off points. [24].

The CNNs were evaluated using 5-fold cross-validation. The resulting models were consolidated by averaging their weights, and final predictions were made on the test set. Classical classifiers (RF and SVM) and their ensembles were evaluated using the same metrics on the test set.

The Kruskal–Wallis test was applied to statistically compare model performances. This non-parametric test evaluates whether significant differences exist between multiple classifiers across different folds by analyzing the mean ranks of different groups [25]. It quantifies the dissimilarities using a single metric (the p-value), testing the null hypothesis (H_0) that the observed differences between group medians can be attributed to random sampling, meaning the groups may originate from the same population [26], [27].

If the Kruskal–Wallis test indicated statistically significant differences ($p < 0.05$), post-hoc pairwise comparisons were conducted using Dunn's test to determine which models presented significantly different performances [28]. This test is widely used in diagnostic and fault detection studies where normality and homoscedasticity assumptions do not hold [29], [10]. Applying these statistical tests ensures a more reliable model performance evaluation by minimizing biases arising from single-test comparisons.

IV. Results

This section discusses the comparative performance of all tested classification models, including individual CNNs, ensemble strategies, and baseline approaches from the literature. All metrics reported consider class imbalance, following recommendations by [4]. Table II presents the performance evaluated through multiple metrics.

Among the individual CNN architectures, VGG16 achieved the highest performance, with an accuracy of 91.8% and an AUC of 0.969. These results are consistent with prior findings from [11], who reported competitive accuracy using the same architecture for pediatric pneumonia detection. The stability and convergence behavior of VGG16, despite its relative architectural simplicity, support its continued use in medical imaging tasks where model transparency and reliability are valued.

MobileNetV2 obtained 88.9% accuracy, making it a viable candidate for low-resource applications. EfficientNetB0, though theoretically promising, underperformed slightly (89.6% accuracy and 0.896 F1-score), suggesting potential sensitivity to hyperparameters or training instability, as previously noted in [13].

When applied to CNN-extracted features, Random Forest achieved 87.8% accuracy, validating the idea of combining handcrafted or statistical features with decision tree-based models, as done by [8].

Ensemble models outperformed individual classifiers across most metrics. The CNNs + RF ensemble achieved the highest overall accuracy (91.9%) and F1-score (0.918), demonstrating a balanced and consistent prediction capability. The CNNs + SVM ensemble yielded the best recall (0.979), a critical factor in clinical diagnosis scenarios where false negatives must be minimized. Despite lower precision (0.862), its F1-score (0.917) and AUC (0.963) confirm its practical effectiveness, aligning with results from [13] using hybrid CNN-SVM methods.

In contrast, the RF + SVM ensemble presented the weakest performance (accuracy = 72.4%, F1 = 0.703), indicating that naïve classifier fusion strategies may not effectively capture model complementarity. Despite being from different families, both classifiers operated on identical CNN-extracted features, which may have limited their diversity and led to overlapping decision boundaries, since the effectiveness of the ensemble depends not only on accuracy but also on the diversity among the basic learners [30].

Compared to the literature, the proposed ensembles generalize well over the same *Kermany dataset* [20]. The CNNs + SVM ensemble, for example, achieved recall comparable to

TABLE II
PERFORMANCE METRICS FOR INDIVIDUAL AND ENSEMBLE MODELS (PROPOSED AND LITERATURE)

Model	Accuracy	Precision	Recall (Sens.)	F1-score	F1-macro	AUC
Literature Models						
[6] – VGG16 + SVM	0.967	0.966	**0.968**	0.967	–	–
[8] – 2D DWT + RF	**0.971**	0.990	0.917	**0.980**	**0.926**	**0.990**
[8] – 2D DWT + SVM	0.934	0.946	0.852	0.875	0.830	0.908
[13] – Hybrid (EffNetB0 + SVM)	0.970	**1.000**	0.958	0.979	–	0.980
[13] – EfficientNetB0 (sigmoid)	0.967	0.999	0.956	0.977	–	0.976
[11] – InceptionV3 + TL	0.928	0.901	0.932	0.928	–	0.968
Proposed Models						
MobileNetV2	0.889	0.893	0.889	0.887	0.877	0.954
VGG16	0.918	0.921	0.918	0.917	0.911	0.969
EfficientNetB0	0.896	0.898	0.896	0.896	0.890	0.963
Random Forest	0.878	0.882	0.878	0.875	0.864	0.945
Ensemble CNNs + RF	**0.919**	**0.921**	**0.919**	**0.918**	**0.911**	**0.977**
Ensemble MN + VGG + RF	0.911	0.915	0.911	0.909	0.901	0.974
Ensemble CNNs + SVM	0.897	0.862	0.979	0.917	0.889	0.963
Ensemble RF + SVM	0.724	0.725	0.724	0.703	0.670	0.802

the *EffNetB0 + SVM* hybrid from [13] (0.979 vs. 0.958), while offering better balance in precision (0.862 vs. 1.000). Additionally, the proposed methods rival traditional pipelines such as *2D DWT + RF* and *2D DWT + SVM* [8], despite relying on raw image features rather than handcrafted descriptors.

This balance is also reflected in the confusion matrices (Figure 3), where ensemble models, particularly CNNs + SVM and CNNs + RF, demonstrate reduced false negatives and improved sensitivity in detecting pneumonia cases, confirming their clinical relevance.

Figure 3 presents the confusion matrices obtained from the best-performing fold of each model. The results indicate model efficacy in distinguishing Normal and Pneumonia cases, with ensembles achieving more balanced predictions. Ensemble methods, especially CNNs + SVM and CNNs + RF, resulted in fewer false negatives, confirming their higher recall and F1-scores. CNNs + SVM had the lowest number of missed pneumonia cases (only 8), while CNNs + RF maintained a strong balance across both classes (only 15 false negatives and 29 false positives). Among individual models, VGG16 demonstrated strong overall performance with low false positive and false negative counts (18 and 27, respectively), outperforming EfficientNetB0 and MobileNetV2, which had higher error rates. Random Forest showed the highest number of false positives (48), suggesting a tendency to over-predict the Pneumonia class, which aligns with its slightly lower precision.

To assess the statistical significance of performance differences across models, the Kruskal-Wallis test was applied to all metrics. As shown in Table III the test revealed significant differences ($p < 0.05$) for all metrics, including accuracy, precision, recall, F1-score, and AUC (only statistically significant pairwise comparisons - Accuracy, are shown).

Dunn's post-hoc test confirmed that Random Forest was significantly outperformed by VGG16 and the CNNs + RF ensemble, with p-values of 0.0078 and 0.0068, respectively. No statistically significant differences were observed between the remaining models, suggesting comparable performance among top classifiers.

TABLE III
KRUSKAL-WALLIS AND DUNN'S TEST RESULTS COMPARING MODELS USING 5-FOLD CROSS-VALIDATION.

Kruskal-Wallis test results		
Metric	**H-statistic**	**p-Value**
Accuracy	22.3256	0.0005
Balanced Accuracy	18.2603	0.0026
Precision	23.0593	0.0003
Recall	22.3256	0.0005
F1-score	22.1094	0.0005
F1-macro	20.7827	0.0009
AUC	25.6245	0.0001
Dunn's Post-Hoc test results (Accuracy)		
Comparison		**p-Value**
Ensemble CNNs + RF *vs* Random Forest		**0.0068**
Random Forest *vs* VGG16		**0.0078**

These findings confirm that ensemble approaches, especially those integrating multiple CNNs with RF or SVM, are practical for pneumonia classification. The models achieved results comparable to or better than prior studies while maintaining practical implementation advantages through lightweight, image-only pipelines.

V. CONCLUSION

This study investigated hybrid and ensemble strategies for detecting pneumonia in chest X-ray images, combining CNN-based feature extraction with classical machine learning classifiers. The results confirm that ensemble models, especially

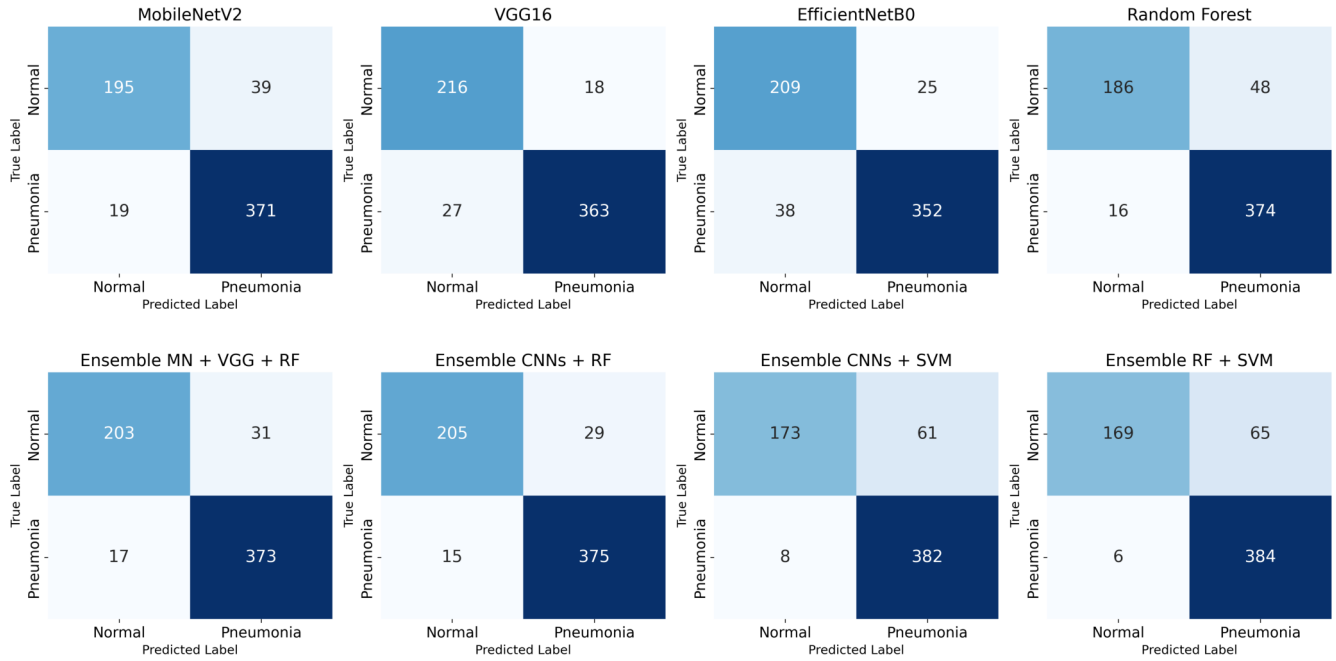


Fig. 3. Confusion matrices representative of the best fold's performance for each individual and ensemble model, evaluated on the fixed test set (624 samples). Each matrix displays (top-left) True Negatives, (top-right) False Positives, (bottom-left) False Negatives, and (bottom-right) True Positives.

those integrating CNNs with RF or SVM, are promising approaches compared to individual models in several evaluation metrics. In particular, the CNNs + RF ensemble obtained the highest accuracy (91.9%) and AUC (0.977), while CNNs + SVM maximized recovery (0.979), an essential metric for minimizing false negatives in clinical diagnosis. Nevertheless, indicator models such as VGG16 also showed significant individual performance (91.8% accuracy), reinforcing their continued relevance in medical imaging tasks, as seen in other studies.

The RF performance suggests that traditional classifiers combined with deep feature extraction remain competitive in medical applications, especially in data-limited scenarios. This can be attributed to RF's tolerance of redundant features and outliers, an aspect worth exploring in future feature importance analyses. In contrast, naive fusion models such as RF + SVM underperformed, highlighting the need for more ensemble strategies and further investigation.

In addition to reinforcing the utility of ensemble strategies, this study offers a refined pipeline for pneumonia detection in X-ray images. The proposed methodology is generalizable to other imaging modalities, such as ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI), particularly when datasets are limited or noisy. A key contribution is the full public release of the experimental code and data pipeline, adhering to open science principles to support reproducibility and community-driven development [31], [32].

In addition to empirical performance, the results also present relevant theoretical and practical implications. Theoretically, this work corroborates the growing evidence that deep feature extractors combined with classical classifiers can be a viable alternative to purely end-to-end models, particularly when training data is limited, noisy, or imbalanced. The practical implications of this study are significant, especially for low-resource healthcare environments where access to specialized radiologists or advanced computational infrastructure may be limited. The hybrid approach provides a cost-effective and computationally efficient solution for automated pneumonia screening, potentially facilitating rapid diagnosis and early intervention by offering scalable and affordable diagnostic tools for real-world healthcare settings. Using lightweight CNNs and interpretable classifiers, such as RF and SVM, allows for deployment in real-world scenarios with constrained computational resources. These characteristics are especially relevant for diagnostic support in underserved healthcare settings, where scalable and efficient solutions must be considered.

Some limitations of this study include computational constraints, the exclusive focus on a pediatric dataset (aged between one and five years), which limits direct generalizability to adult populations, and the focus on binary classification. Although clinically justified for early pneumonia screening, the expansion to multiclass scenarios should be evaluated and applied for future work, particularly in the differential diagnosis of conditions such as tuberculosis, COVID-19, and other lung diseases with larger classes.

Given this, future directions include hyperparameter optimization for underperforming models such as EfficientNetB0, experimentation with advanced ensemble fusion strategies, such as weighted averaging schemes, stacking, or boosting

techniques that can leverage the complementary strengths of diverse hybrid models, and incorporation of explainability techniques to support clinical decision-making and model transparency. Furthermore, exploring more advanced architectures, such as those integrating attention mechanisms or Transformer-based models, as demonstrated by recent studies like [17], [16] and [15], could yield further performance improvements and enhance interpretability for critical clinical tasks. Broader validation on external datasets is also recommended to assess generalizability in real-world clinical scenarios, especially across diverse age groups and multi-institutional cohorts to enhance clinical translatability.

The code developed is available publicly[5], promoting the reproduction and extension of the experiments carried out.

ACKNOWLEDGMENTS

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9, POSDOC- 101057/2024-5; and by Acordo de Cooperação Técnica N.° 02/2021 (N.° 38328/2020-TJ/MA).

REFERENCES

[1] H. F. M. de Sousa, L. S. V. da Silva, and F. N. da Costa, "Efeitos das queimadas na saúde da população com foco para as doenças pulmonares," *Revista Ibero-Americana de Humanidades, Ciências e Educação*, vol. 10, no. 5, pp. 3126–3150, 2024. doi: 10.51891/rease.v10i5.14016

[2] F. Caobelli, "Artificial intelligence in medical imaging: Game over for radiologists?" *European journal of radiology*, vol. 126, 2020. doi: 10.1016/j.ejrad.2020.108940

[3] G. Liebel, P. V. Dias, I. J. C. Schneider, A. R. d. Sá Junior, A. Hentz, C. d. S. Ferreira, and A. Chaoubah, "Analysis of expenses with diagnostic imaging in brazil," *Cadernos Saúde Coletiva*, vol. 29, pp. 453–463, 2021. doi: 10.1590/1414–462X202129030397

[4] S. Sharma and K. Guleria, "A systematic literature review on deep learning approaches for pneumonia detection using chest x-ray images," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24 101–24 151, 2024. doi: 10.1007/s11042-023-16419-1

[5] W. Liawrungrueang, I. Han, W. Cholamjiak, P. Sarasombath, and K. D. Riew, "Artificial intelligence detection of cervical spine fractures using convolutional neural network models," *Neurospine*, vol. 21, no. 3, p. 833, 2024. doi: 10.14245/ns.2448580.290

[6] M. Toğaçar, B. Ergen, Z. Cömert, and F. Özyurt, "A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models," *Irbm*, vol. 41, no. 4, pp. 212–222, 2020. doi: 10.1016/j.irbm.2019.10.006

[7] A. Mabrouk, R. P. Diaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest x-ray images using ensemble of deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 13, p. 6448, 2022. doi: 10.3390/app12136448

[8] A. Akgundogdu, "Detection of pneumonia in chest x-ray images by using 2d discrete wavelet feature extraction with random forest," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 82–93, 2021. doi: 10.1002/ima.22501

[9] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia detection using cnn based feature extraction," in *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2019. doi: 10.1109/ICECCT.2019.8869364 pp. 1–7.

[10] F. G. da Silva, L. P. Ramos, B. G. Palm, and R. Machado, "Assessment of machine learning techniques for oil rig classification in c-band sar images," *Remote Sensing*, vol. 14, no. 13, p. 2966, 2022. doi: 10.3390/rs14132966

[11] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018. doi: 10.1016/j.cell.2018.02.010

[12] J. O. Diniz, D. A. Dias Jr, L. B. da Cruz, D. L. Gomes Jr, O. A. Cortês, and A. O. de Carvalho Filho, "Efficientensemble: Diagnóstico de câncer de mama em imagens de ultrassom utilizando processamento de imagens e ensemble de efficientnets," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. SBC, 2024. doi: 10.5753/sbcas.2024.2155

[13] O. M. El Zein, M. M. Soliman, A. Elkholy, and N. I. Ghali, "Transfer learning based model for pneumonia detection in chest x-ray images," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 5, pp. 56–66, 2021. doi: 10.22266/ijies2021.1031.06

[14] C. Munzlinger, I. Yepes, and R. Rieder, "Uso de uma rede neural convolucional para análise de exames de radiografia de pulmao com detecçao de covid-19, pneumonia e tuberculose," in *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*. SBC, 2023. doi: 10.5753/sbcas_estendido.2023.229629 pp. 25–30.

[15] Y. Wu, N. Japkowicz, S. Gilbert, and R. Corizzo, "Attention-based medical knowledge injection in deep image classification models," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[16] J. Rocha, S. C. Pereira, J. Pedrosa, A. Campilho, and A. M. Mendonça, "Stern: Attention-driven spatial transformer network for abnormality detection in chest x-ray images," *Artificial Intelligence in Medicine*, vol. 147, p. 102737, 2024.

[17] C. J. Ejiyi, Z. Qin, A. O. Nnani, F. Deng, T. U. Ejiyi, M. B. Ejiyi, V. K. Agbesi, and O. Bamisile, "Resfeanet: Resnet-fused external attention network for tuberculosis diagnosis using chest x-ray images," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100133, 2024.

[18] M. Bartosiewicz, M. Iwanowski, M. Wiszniewska, K. Frączak, and P. Leśnowolski, "On combining image features and word embeddings for image captioning," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023. doi: 10.15439/2023F997 pp. 355–365.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[20] D. Kermany, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, 2018. doi: 10.17632/rscbjbr9sj.2

[21] O. O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, and A. Abayomi-Alli, "Bilstm with data augmentation using interpolation methods to improve early detection of parkinson disease," in *2020 15th conference on computer science and information systems (FedCSIS)*. IEEE, 2020. doi: 10.15439/2020F188 pp. 371–380.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. doi: 10.1109/CVPR.2018.00474 pp. 4510–4520.

[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[25] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. doi: 10.1080/01621459.1952.10483441

[26] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010. doi: 10.1002/9780470479216.corpsy0491

[27] E. Ostertagova, O. Ostertag, and J. Kováč, "Methodology and application of the kruskal-wallis test," *Applied mechanics and materials*, vol. 611, pp. 115–120, 2014. doi: 10.4028/www.scientific.net/AMM.611.115

[28] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[29] M. A. Jamil and S. Khanam, "Influence of one-way anova and kruskal–wallis based feature ranking on the performance of ml classifiers for bearing fault diagnosis," *Journal of Vibration Engineering & Technologies*, vol. 12, no. 3, pp. 3101–3132, 2024. doi: 10.1007/s42417-023-01036-x

[5]https://github.com/GabrieleAraujo/pneumonia_detection_cnn-ml.git

[30] Y. Yang, H. Lv, and N. Chen, "A survey on ensemble learning under the era of deep learning," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5545–5589, 2023.

[31] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. Ioannidis, "A manifesto for reproducible science," *Nature human behaviour*, vol. 1, no. 1, p. 0021, 2017. doi: 10.1038/s41562-016-0021

[32] D. Donoho, "50 years of data science," *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 745–766, 2017. doi: 10.1080/10618600.2017.1384734

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 471–478
DOI: 10.15439/2025F1103
ISSN 2300-5963 ACSIS, Vol. 43

Treating OCR Output as a Language (TOOL) – Improving OCR Output with Seq2Seq Translation

Thomas Asselborn*†, Magnus Bender‡, Ralf Möller*, Sylvia Melzer*†
* University of Hamburg, Institute for Humanities-Centered AI, Warburgstraße 28, 20354 Hamburg, Germany
Email: {thomas.asselborn, ralf.moeller, sylvia.melzer}@uni-hamburg.de
https://orcid.org/0009-0005-3011-7626, https://orcid.org/0000-0002-1174-3323, https://orcid.org/0000-0002-0144-5429
† University of Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany
‡ Aarhus University, Department of Management, Fuglesangs Allé 4, 8210 Aarhus V, Denmark
Email: magnus@mgmt.au.dk, https://orcid.org/0000-0002-1854-225X

***Abstract*—Optical Character Recognition (OCR) systems are frequently used to digitise text, but often produce noisy results, especially with historical, poor-quality or multilingual data. Despite advances in OCR technology, post-processing remains a significant bottleneck. We propose TOOL (Treating OCR Output as a Language), a new approach that understands OCR correction as a machine translation task. By treating noisy OCR text as a language in its own right, TOOL employs sequence-to-sequence models like Marian to translate it into clean, standardised text. This method is scalable, model-independent and language-flexible. We demonstrate this approach by translating "OCR German" to Standard German from around 1871 to the present day, improving accuracy at the token level by using matched training pairs of OCR output and base text.**

I. Introduction

OCR (**O**ptical **C**haracter **R**ecognition) systems are widely used to digitise printed and handwritten documents. However, their output is often noisy, particularly when dealing with low-quality scans, historical materials or multilingual content. While recent advancements in computer vision and language modelling have enhanced OCR (**O**ptical **C**haracter **R**ecognition) accuracy, post-processing remains a critical bottleneck, especially in error-prone scenarios. OCR text recognition errors tend to follow systematic, learnable patterns, similar to those found in low-resource language translation or noisy text correction.

TOOL (**T**reating **O**CR **O**utput as a **L**anguage) introduces a novel paradigm: treating OCR output as a textual language that can be translated into clean text (free from errors) using Seq2Seq (**Seq**uence-**to**-**Seq**uence) models. By leveraging translation models such as BART (**B**idirectional and **A**uto-**R**egressive **T**ransformers) [14], Marian [13] or other LLMs (**L**arge **L**anguage **M**odels), TOOL frames OCR correction as a machine translation task. This enables robust, model-agnostic and domain-adaptable post-processing. The approach offers a scalable and language-independent framework for improving OCR results without requiring any modifications to the OCR engine itself.

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796.

There is a need for such a flexible, general-purpose solution that can learn to correct OCR errors in a data-driven manner. This approach addresses the problem by training Seq2Seq models on aligned pairs of noisy OCR output and the corresponding GT (**G**round **T**ruth) text. In this paper, we specifically address the task of translating from "OCR German" to standard German from 1871 to the present day to improve token accuracy.

II. Related Work

Seq2Seq modelling plays a foundational role in tasks such as machine translation, text summarisation, and now, OCR post-correction. The original Seq2Seq framework was introduced by Sutskever et al. (2014) [24], who proposed using RNs (**R**ecurrent **N**etwork**s**) with LSTM (**L**ong **S**hort-**T**erm **M**emory) units to encode an input sequence into a fixed-length vector and decode it into an output sequence. This approach demonstrated promising results for translations but had difficulties with long sequences due to information bottlenecks in the fixed-size context vector. To address this limitation, Bahdanau et al. (2015) [3] introduced the *attention mechanism*, allowing the decoder to access different parts of the input sequence dynamically. This innovation significantly improved translation quality and became a core component of subsequent Seq2Seq models.

The Transformer architecture by Vaswani et al. (2017) [26] replaced recurrence entirely with self-attention mechanisms, enabling more efficient training and better performance on long sequences. Transformers rapidly became the standard architecture for Seq2Seq tasks. Building on the Transformer, Lewis et al. (2020) [14] introduced BART, a denoising autoencoder for Seq2Seq generation. BART combines a bidirectional encoder (similar to BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) [9]) with an autoregressive decoder (similar to GPT (**G**enerative **P**re-trained **T**ransformer) [21]), making it well-suited for text generation, correction and translation tasks; particularly when input text is noisy, such as in OCR outputs.

Around the same time, Marian [13] was developed as a highly efficient, production-grade translation system optimised for multilingual environments. As shown by Junczys-

Dowmunt et al. (2018) [12], Marian is particularly effective in low-resource settings, making it a strong candidate for OCR correction across diverse languages.

III. Use Case

This section begins by outlining the background of education, followed by an examination of the data availability concerning VET (**V**ocational **E**ducation and **T**raining), which serves as the primary data source and use case for the proposed TOOL approach.

A. Education Stages

Education comprises a sequence of structured stages, each contributing uniquely to an individual's personal and intellectual development [19]. The stages usually begin with ECEC (**E**arly **C**hildhood **E**ducation and **C**are), which includes programmes and services for children from birth to five years of age. This period is vital for the development of foundational cognitive, social and emotional skills. Evidence suggests that high-quality ECEC programs can have long-term positive effects on academic performance and social integration [17], [25].

Following ECEC, school education is generally categorised into primary and secondary levels. Primary education lays the groundwork for essential academic skills such as reading, writing, mathematics and science while fostering critical thinking and communication. Secondary education expands on this base, introducing more complex subjects and sometimes offering specialised or vocational tracks to prepare students for higher education or employment [10].

Higher education encompasses academic institutions like universities and colleges that confer undergraduate and postgraduate qualifications. These institutions provide in-depth disciplinary expertise and facilitate the development of research, innovation and professional competencies. Higher education is thus a significant driver of socio-economic growth and individual career advancement [1].

VET is a parallel pathway focusing on practical and occupational skills, tailored to specific sectors such as healthcare, manufacturing, hospitality and technology. VET programs often involve work-based learning, such as apprenticeships, to ensure graduates meet industry needs. These programs also support lifelong learning by offering reskilling and upskilling opportunities, thereby enhancing workforce adaptability [7], [8].

CVET (**C**ontinuing **V**ocational **E**ducation and **T**raining) plays a crucial role in supporting ongoing employability and national competitiveness. Many European countries, including Germany, have implemented policies to support lifelong learning and professional development through national strategies [22]. The German system differentiates between initial vocational training (German: *Ausbildung*), retraining (German: *Umschulung*), advanced continuing training (German: *Weiterbildung*) and upgrading training (German: *Fortbildung*). Upgrading training is often governed by federal regulations such as the Vocational Training Act (German: *Berufsbildungsgesetz*, BBiG) [6] and the Crafts Code (German: *Handwerksordnung*, HwO) [5]. In contrast, other forms of CVET are typically more decentralised and less regulated, though equally important for continuous skill development.

B. OCR Processing of Documents for VET

Fig. 1 shows a document which is part of a collection of historical VET. It is just one example of many similar materials that outline training regulations, exam requirements and pedagogical standards for skilled trades in early 20th-century Germany. These documents were typically issued or supervised by state institutions such as the Imperial Institute for Vocational Training in Trade and Industry (German: *Reichsinstitut für Berufsausbildung in Handel und Gewerbe*), which played a central role in standardising vocational training across the German Reich. The aim is to process such

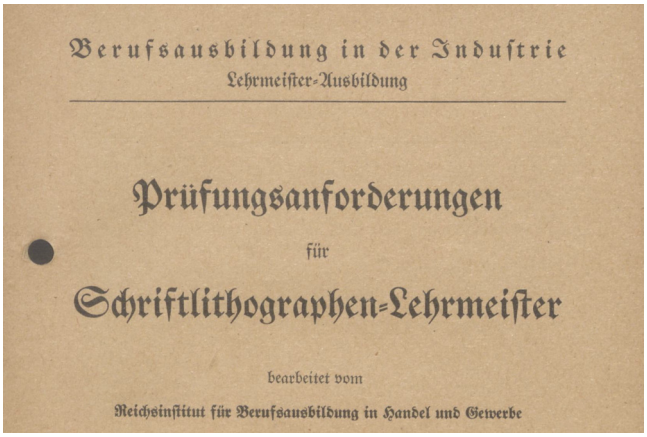


Fig. 1. Exam requirements for a type lithographer teacher (orig. in German: *Prüfungsanforderungen für Schriftlithographen-Lehrmeister*).

documents using OCR to make them digitally searchable and analysable. By converting these historical printed documents into machine-readable text, we aim to simplify research into the development of vocational training systems, qualification frameworks and occupation-specific pedagogical methods. However, processing these materials presents significant technical challenges.

An important problem is the typographical style of the original prints: Many documents from the early 20th century use Fraktur or broken Antiqua fonts, which are very different from modern Latin script. These fonts have unusual letter forms (e.g., long s "ſ" versus short s "s") and dense ligatures that are poorly recognised by standard OCR engines. The result is a high rate of recognition errors, such as character confusion, word fragmentation or misreadings that distort key terms (e.g., the German words: *Lehrmeister*, *Prüfungsanforderung*, etc.). In addition, the print quality of many originals contributes to noisy OCR output with smudges, uneven contrast and marginal notes that are interpreted as content. Therefore, in the following, we show a new approach to dealing with these problems without having to change the OCR process itself.

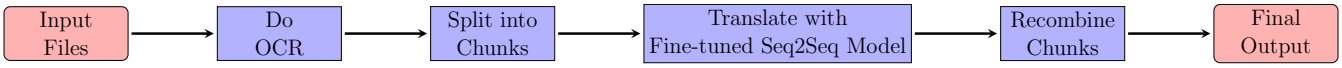


Fig. 2. Process flow for the TOOL-based OCR translation pipeline. *Input files* are first passed through the *Do OCR* process. Then the following processes are executed: *Split into Chunks*, *Translate with Fine-tuned Seq2Seq Model*, and *Recombine Chunks*. Together, these steps produce the *Final Output*.

IV. Method

To introduce TOOL, we begin by analysing the broader context. Specifically, we consider a comprehensive dataset comprising scans of VET documents, which is a collection of image files representing a set of text-based records. The primary objective is to generate high-quality transcriptions of these documents in a manner that is as automated as possible. Additionally, a limited set of manually transcribed documents is available, which can be used for training purposes. Generally, this situation is a common use case for OCR, but as stated before, the OCR results do not fulfil our needs. Furthermore, the available set of transcribed documents is too small to be used for training of a full custom OCR model. However, we recognise that the not-satisfying OCR output follows its own characteristics. Typical and recurring errors are, e.g., (i) the Umlaut "ü" in a scanned document often results in the sequence "ii" in the recognized text or (ii) the long s "ſ" is recognized as "f".

We observe that the OCR output is not correct German, but some type of inherent *language*. Hence, the output can be treated as a language, which leads us to the approach of TOOL: Translating from the OCR output *language* to a correct natural language, German in our case. Thus, TOOL consists of two steps: the OCR step and the *translate* step.

The OCR step does not require any further training of models and generalises well for different OCR models and languages. Even though the *translate* step still requires an individually trained model, this training requires only a small amount of training data and resources.

State-of-the-art translation models are based on LLMs, just like the translation step of TOOL. There are many different methods for solving the OCR problem described above, several of which we briefly outline here. These include a) Encoder, b) Instruction LLM, and c) Seq2Seq.

a) Encoder: One potential approach involves the use of an encoder-only model such as BERT. BERT processes an input sequence of words and outputs a corresponding sequence of contextualised embedding vectors. The idea is to utilise the vector space of the embedding vectors of a default BERT and train a custom BERT in a way that it generates for the OCR output embeddings in the same vector space as a default BERT. The big advantage over the instruction LLMs, such as Llama and GPT, is that in contrast only a small BERT needs to be fine-tuned for TOOL and a shared embedding vector space is used, i.e., techniques like RAG (**R**etrieval **A**ugmented **G**eneration) for retrieving documents from a collection can be applied directly. However, there is also a need for the correct German transcriptions of the scans and text generation from BERT embeddings is not straightforward. Additionally, there is no GT for the vector embeddings as our training data contains transcribed documents and not embeddings.

b) Instruction LLM: Another possibility is to apply an instruction LLM, like Llama or GPT, similar to [2]. Instead of training a model for TOOL's translation step, the instruction LLM gets the OCR output and is prompted to *correct* it. This approach works well, but there are disadvantages. While requiring no training is an advantage, the downside is that the process does not speed up once training is complete. An instruction LLM is a pretty large model, and running it requires special hardware and a lot of resources.

c) Seq2Seq: After all, TOOL's translation step should use a small transformer based LLM, i.e., a model taking a sequence of words as input and returning a transformed, e.g., translated, version of the words. A popular choice for such Seq2Seq models is Marian [13]. Using Marian for TOOL's translation step solves both issues stated before: First, there is no need for sharing an embedding vector space as required when using BERT. Second, Marian is a comparatively small model and can be fine-tuned in a reasonable amount of time. Additionally, after fine-tuning, the model can be used with standard hardware. Following a thorough evaluation, we selected the Seq2Seq model Marian, or more specifically, its Python implementation called MarianMT[1], to perform the translation step within TOOL.

The full TOOL pipeline is shown in Fig. 2. On the left-hand side, the red box depicts the input files, i.e., the scanned VET documents. The first two purple boxes depict TOOL's first step, i.e., the OCR. The remaining three purple boxes represent TOOL's second step, the translation using a custom fine-tuned MarianMT. As the input length of MarianMT is limited, the OCR output is split into chunks before it is translated by MarianMT. Finally, the translated chunks are concatenated, and the results are ready to be stored (the red box on the right-hand side).

Overall, the big advantages of the TOOL approach are:

- Any OCR tool can be used in the first step.
- For the second step, a pre-trained Seq2Seq LLM for translation, like variants of MarianMT, can be used.
- It does not require much training in terms of fine-tuning.
- The resulting pipeline, OCR and translation, is quite fast and resource efficient, especially compared to applying an instruction LLM [2].

Next, we describe the technical details of TOOL, our evaluation and present the results.

[1]https://huggingface.co/docs/transformers/model_doc/marian

V. Experiments and Results

In order to verify the method described in this paper, experiments were conducted using the texts described in Section III. First, the process of fine-tuning with its subprocesses a) OCR, b) Preprocessing for Fine-Tuning, and c) Fine-Tuning is described.

A. Preprocessing and Fine-Tuning

A process flow diagram for fine-tuning the model can be seen in Fig. 3. The experimentation setup does not serve as a mean to find out the best possible models and parameters to use but rather it aims to show that the method works in practical scenarios.

a) OCR: As a first step, OCR is performed on all files from our dataset. We have decided to use Tesseract [23] as our OCR engine of choice as it is open source and widely adopted, with libraries available for many programming languages like Python[2].

b) Preprocessing for Fine-Tuning: The files are randomly split into a dataset for training and validation, and another dataset for testing after training. The second testing dataset is not used during the training phase, but to compute the metrics later on to verify how well the model generalises. For this experiment, we have split the dataset to have around 50% of the text files in the training and validation dataset and the other 50% in the testing dataset. Thus, each dataset has around 145 text files in total.

As a next step, the texts in each file are split into separate sentences using the NLTK (**N**atural **L**anguage **T**ool**k**it) [16] punkt tokeniser. This is done for the texts from OCR and the GT texts to have texts that can be used for fine-tuning the MarianMT model.

For the final preprocessing step, the corresponding sentences from the GT (see Fig. 2) and after the OCR process need to be aligned in order to have the corresponding sentence pairs. To do so, we have used a LaBSE (**L**anguage-**a**gnostic **B**ERT **S**entence **E**mbedding) model [11] which is based on the BERT architecture. This language-agnostic model allows us to encode texts from different languages into the same vector space. Using a distance measure, like the cosine similarity, we can pair sentences that are close in this vector space. The cosine similarity, as used, is defined as

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}, \text{ where} \tag{1}$$

- $\mathbf{A}$ and $\mathbf{B}$ are two n-dimensional vectors, $n \in \mathbb{N}$,
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product between two vectors $\mathbf{A}$ and $\mathbf{B}$ and
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ being the Euclidean norms of $\mathbf{A}$ and $\mathbf{B}$.

The vectors $\mathbf{A}$ and $\mathbf{B}$ are the embeddings of the GT texts and after the OCR process. The resulting value is in the range of $[-1, 1]$, with -1 denoting exactly opposite vectors, 0 denoting orthogonal vectors and 1 denoting exactly the same vectors.

[2]https://pypi.org/project/pytesseract/

Since our goal is to have sentence pairs that are highly likely to correspond, yet they may differ in a few aspects, we have set a minimum cosine similarity between the pairs to 0.7. In the end, we obtained approximately 10000 sentence pairs for training and verification during training and 8000 sentence pairs for testing afterwards.

c) Fine-Tuning: After generating the sentence pairs, fine-tuning is possible. We have chosen to use a MarianMT model [13] because this class of models is relatively small and lightweight. This not only ensures that fine-tuning is possible on moderate hardware but also makes inference possible on modern consumer computers.

Based on the specific language(s) needed, the user must choose the appropriate base model. Since we are dealing with texts written in German, we have used the "Helsinki-NLP/opus-mt-de-de"[3] model as our basis for fine-tuning. This model was pre-trained to translate from German into German, with our goal after fine-tuning being to translate from "OCR German" into correct German. For the experiments here, the following hyperparameters have been used:

- number of epochs: 30,
- learning rate: 5×10^{-5} and
- weight decay: 0.05.

The hyperparameters were chosen based on past experiences and are thus not necessarily the ones that provide the best results. Instead, the goal is to show that the method does work in general. Based on the specific use case, the hyperparameters need to be found, e.g., using Optuna[4]. For the fine-tuning, 90% of the training dataset has been used for training while the remaining 10% has been used for validation.

Fine-tuning was performed on a single Nvidia DGX A100 with 80GB of video RAM[5]. After around 1.5 hours, fine-tuning was finished, and the resulting model was stored to be used in the translation pipeline.

B. Translation

The translation pipeline follows the same process flow described in Fig. 2. In the first step, standard OCR needs to be performed on the text corpus. This output then needs to be split up into chunks of appropriate length. The reason behind splitting up is that most Seq2Seq models are only able to process a certain number of tokens. This value needs to be researched by the user and is set to 512 tokens for MarianMT. One possible way to generate these chunks is to split the input into sentences. Since the experiments performed for this article already did both steps for the pipeline presented prior, OCR and splitting up were not done in the translation pipeline, where it would have been typically done in a real-world scenario. The next step is then the translation, which is done with the fine-tuned MarianMT model. Additionally, for comparison in this work, translation was also done using the base MarianMT that is pre-trained for translating German into

[3]https://huggingface.co/Helsinki-NLP/opus-mt-de-de

[4]https://optuna.org/

[5]The datasheet can be found here: https://images.nvidia.com/aem-dam/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf

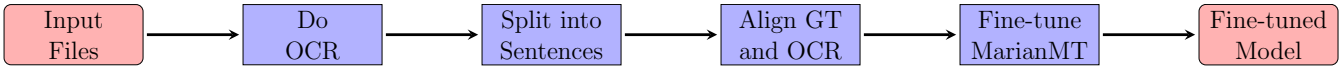


Fig. 3. Process flow for OCR to MarianMT fine-tuning pipeline. *Input files* are first passed through the *Do OCR* process. Then the following processes are executed: *Split into Sentences*, *Align GT and OCR*, and *Fine-tune MarianMT*. Together, these steps produce the *Fine-tuned Model*.

German, but not with the special cases we have in our corpus, like the ſ. After translating the text sentence by sentence, the results need to be recombined to get the full text back again.

This translation step was performed on a standard MacBook Pro 14 inch with M3 processor and 16GB of RAM[6]. For all the examples from the testing dataset, translation took around 1.5 hours.

C. Performance Metrics

Before discussing the results of our experiments, a brief introduction to the performance metrics used is given.

a) Hunspell Errors: One of the goals of TOOL is to have valid German afterwards. Thus, we have used Hunspell to verify that. Hunspell is an open-source spell checker and morphological analyser that is used in a few popular open-source tools like LibreOffice or the Firefox web browser, among others [18]. In order for Hunspell to do its work, a dictionary needs to be used. For our task, we have used a German dictionary called "German (de-De frami)"[7]. The number of errors given by Hunspell based on this dictionary are counted. To have a value that then also accounts for the length of the text, as longer texts are more likely to have a higher number of errors produced, the resulting value is then divided by the total number of words in the GT text.

$$\text{Hunspell value (text)} = \frac{E}{N}, \quad \text{where} \tag{2}$$

- E is the number of Hunspell-detected spelling errors and
- N is the number of words in the GT text.

b) (Normalised) Levenshtein Distance: The Levenshtein distance is a measure used to calculate the distance between two sequences. In our case, we calculate the distance between two sequences of characters. To calculate the Levenshtein distance, the minimum number of single-character edits to transform a provided sequence A, e.g., the texts after the TOOL pipeline, into the other sequence B, here the GT texts, is counted. These single-character edits can be insertions, deletions or substitutions. Similar to the case with the Hunspell errors, a longer text will more likely need more edits to be transformed into the other text, and thus, we have normalised the Levenshtein distance by dividing the results by the number of characters in the GT text.

c) Word Error Rate (WER): The WER (**W**ord **E**rror and **R**ate) is calculated in a similar way to the Levenshtein distance. In contrast to the Levenshtein distance, though, the WER counts the minimum number of words needed to change a text A into another text B. As an example, "I like a sandwich" and "I like this sandwich" have three character errors but only one word error, which is viewed at for the WER. For computing the WER, the number of word substitutions, deletions and insertions is counted. The resulting value is then divided by the total number of words in the GT text.

$$WER = \frac{S + D + I}{N}, \text{where} \tag{3}$$

- S being the number of substitutions,
- D being the number of deletions,
- I being the number of insertions and
- N being the total number of words in the GT text.

d) Jaccard Similarity: The Jaccard similarity is a set-based similarity measure between two sets A and B. Because sets do not have an order, the similarity measure does not take the order of the words in the text into account. However, this measure can still provide interesting insights into how well a user searching for a specific term may find this term in the results. Formally, it is defined as:

$$\text{Jaccard}(A, B) = \left| \frac{A \cap B}{A \cup B} \right|. \tag{4}$$

The resulting value will be between 0 and 1, with 1 denoting identical sets and 0 meaning no overlaps at all. In order to still preserve some form of ordering, we have measured the Jaccard similarity with regards to word bigrams, i.e., the sets consist of combinations of two words as entries. This takes immediate context into account and makes the measure stricter.

e) BLEU(-4): The BLEU (**Bil**ingual **E**valuation **U**nderstudy) score is a value between 0 and 1, and it evaluates the quality of one text compared to another. If the resulting value is getting closer to one, both texts are getting more similar. Initially, it was designed for machine translation tasks, but it can also be applied to other similar tasks. In essence, BLEU calculates how many words or n-grams in the candidate text, i.e., the text after TOOL, match those of the reference text. Detailed information can be found in its original paper by Papineni et al. [20].

We have decided to use the BLEU-4 score version that considers n-grams up to 4-grams, i.e., sequences of one up to four words in length. The precision for each n-gram is calculated, and then the geometric mean over all scores is taken.

$$\text{BLEU-4} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n \right), \text{where} \tag{5}$$

- BP is the brevity penalty,
- p_n the precision for n-gram of size $n \in \mathbb{N}$ and

[6]For technical specifications, visit https://web.archive.org/web/20250518115318/https://support.apple.com/en-gb/117735

[7]https://extensions.openoffice.org/en/project/german-de-de-frami-dictionaries

- w_n the weight for each n-gram precision (in our case, we have set it uniformly to $\frac{1}{4}$).

The brevity penalty is used to penalise translations that are too short when compared to the reference.

$$\text{BP} = \begin{cases} 1 & \text{if } t > g \\ \exp\left(1 - \frac{g}{t}\right) & \text{if } t \leq g \end{cases}, \text{where} \tag{6}$$

- t is the total word length of the translation and
- g is the total word length of the GT text.

Thus, if the translation is longer than the GT text, nothing happens, but if the translation is shorter, the BLEU score gets penalised exponentially.

f) METEOR: Unlike BLEU, the METEOR (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **Or**dering) score is not only looking at exact matches, but it is extended to also include stem matching (e.g., "read" and "reading") and synonym matching (e.g., "bike" and "bicycle"). More details can be found in its original paper by Banerjee and Lavie [4]. It is calculated as follows:

$$\text{METEOR} = F_{mean} \cdot (1 - Penalty). \tag{7}$$

The individual components of this equation are given as:

$$Precision = \frac{m}{w_t}, \quad Recall = \frac{m}{w_{gt}}, \text{where} \tag{8}$$

- m is the number of matched unigrams,
- w_t the number of words in the translation and
- w_{gt} the number of words in the GT texts.

Using precision and recall, the F_{mean} score can be calculated.

$$F_{mean} = \frac{(1+\alpha) \cdot Precision \cdot Recall}{\alpha \cdot Precision + Recall}, \text{where} \tag{9}$$

- $\alpha \in (0, 1)$ being a parameter giving weight to recall over precision.

By default, α is set to 0.9, meaning that recall is weighted slightly above precision.

Finally, the penalty term is computed as:

$$Penalty = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}, \text{where} \tag{10}$$

- ch is the number of chunks,
- m the number of matches and
- β and γ being tuning parameters.

Also for the penalty term, the parameters β and γ have been kept their default values ($\beta = 3$, $\gamma = 0.5$).

g) ROUGE(-2): The ROUGE (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation) score was first described by Lin [15] in 2004. In contrast to BLEU, ROUGE is generally a measure of recall. However, it can also be F1-score based, which is the version we have used in our evaluation. Additionally, we have calculated the score using text bigrams to capture more of the context of the words. This F1-score version using bigrams is calculated as follows:

$$\text{ROUGE-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{11}$$

Precision and recall are as defined in Equation 8 but with word bigrams instead of unigrams.

D. Results

The evaluation for TOOL was performed on the previously held-back testing dataset. This dataset contains around 8000 sentence pairs consisting of the texts after OCR, base MarianMT or TOOL respectively, and the corresponding GT sentences. All the text used for this dataset was chosen at random from our text corpus. Using that dataset, eight performance metrics were calculated. First, they were calculated for each sentence individually and then combined using both the mean and the median. Table I shows all the metrics for all the test datasets, all eight performance metrics and both the mean and median. Row "Tesseract only" shows the performances for only using the Tesseract OCR engine. In row "Base MarianMT" are the results shown for using the base MarianMT German into German translation model after the Tesseract OCR. This is done to compare the performance of the pre-trained MarianMT with both Tesseract alone and also TOOL. Finally, the row "TOOL" contains the performances for the approach presented in this article. The best values for both mean and median for each performance metric are highlighted in boldface.

Figure 4 shows the same data also in bar chart form. Orange bars are for the experiments with Tesseract alone, blue bars are the experiments after using the pre-trained MarianMT, and magenta bars are for TOOL. Solid bars are used for the mean, while hatched bars are for the median.

The performance metrics are calculated as described in Subsection V-C. For the columns "Hunspell Errors", "Normalised Levenshtein Distance" (abbreviated to "Norm. Lev. Dist."), and the "Word Error Rate" ("WER" in the table), a smaller value denotes a better performance. In contrast, for the "Jaccard Bigrams", "BLEU-4", "METEOR" and "ROUGE-2", a higher value denotes a better performance. For brevity, in this section we will refer to the BLEU-4 score as just BLEU and the ROUGE-2 score as just ROUGE.

At first, we look at the mean of each performance metric. Overall, it can be seen that the TOOL pipeline improves performance for all performance metrics. The number of errors identified using Hunspell was, on average, almost halved. Also, the Jaccard similarity for word bigrams as well as the BLEU and ROUGE scores doubled compared to the baseline Tesseract only. The normalised Levenshtein distance also got reduced. All of this indicates that the resulting text after TOOL gets closer to the GT, showing that the pipeline works.

Comparing the results after Tesseract alone and TOOL to the base pre-trained MarianMT was done in order to judge how much of the ability to correct the OCR errors is already inherent to the base model and how much was added via fine-tuning. Additionally, this serves as a baseline to judge whether the still relatively small dataset for fine-tuning has an impact on performance or not.

The MarianMT base model is consistently outperformed by our TOOL pipeline, demonstrating that fine-tuning sig-

TABLE I
EVALUATION METRICS FOR DIFFERENT MODELS (MEAN, MEDIAN).

Metric / Model	Hunspell Errors		Norm. Lev. Dist.		WER		Jaccard Bigrams		BLEU-4		METEOR		ROUGE-2	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Tesseract only	0.3670	0.4568	0.2275	0.1834	0.5357	0.5347	0.2079	0.1435	0.3047	0.2559	0.5977	0.5710	0.2913	0.2270
Base MarianMT	0.3455	0.4311	0.3387	0.3090	0.6252	0.5975	0.1975	0.1383	0.2338	0.1749	0.5083	0.4737	0.2755	0.2150
TOOL	**0.1963**	**0.1925**	**0.2175**	**0.1587**	**0.3626**	**0.2716**	**0.4723**	**0.4954**	**0.5695**	**0.6211**	**0.7761**	**0.8052**	**0.5781**	**0.6053**

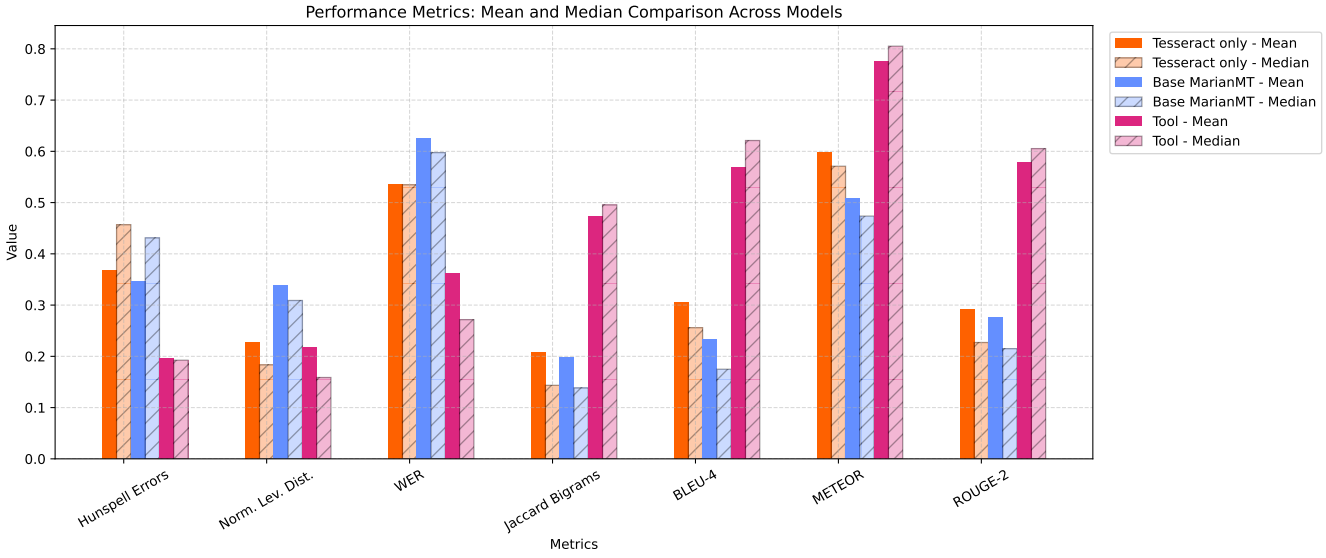


Fig. 4. Evaluation metrics shown as bar charts. Solid bars indicate the mean, while hatched bars are the median.

nificantly improves the results, even with a relatively small dataset. Moreover, the base MarianMT is also almost always worse than just Tesseract alone. It only outperforms Tesseract alone in the number of Hunspell errors, which is an indicator that the language produced after the translation is closer to German than the result after OCR alone. However, in correcting these mistakes, the results diverge from the GT leading to overall worse performance. Thus, using the base MarianMT without first fine-tuning to this specific OCR task is not helpful and even degrades results.

In addition to the mean, we also calculated the median value of each performance metric over all results. This was done in order to judge whether a few examples are skewing the results in one or the other direction. Looking at all performance metrics, the median is mostly worse than the mean for Tesseract only, but also for the base MarianMT model. The two exceptions are the normalised Levenshtein distance and the word error rate, where the results are getting slightly better. This is an indicator that there are a few outliers skewing the results slightly in a mostly worse direction. For our TOOL pipeline, the value is, however, almost always very close in both mean and median. The trend here is the same, with base MarianMT producing the worst results, while Tesseract alone performs slightly better, apart from the Hunspell errors, as already described in the mean value, and TOOL is the best.

VI. CONCLUSION AND OUTLOOK

In this article, we introduced TOOL, a novel method that reconceptualises OCR correction as a translation task from noisy to clean text. We demonstrated TOOL, which uses the Seq2Seq model MarianMT to translate the "OCR German" output into Standard German from around 1871 to the present day. We implemented TOOL with this lightweight model that runs efficiently on consumer-grade hardware. While the results are promising and show performance improvements, the current study relies on a relatively small dataset and serves as a proof of concept.

Future work will focus on scaling to larger datasets, which is expected to enhance performance and robustness. Extending this approach to other languages also represents a promising direction for future research.

REFERENCES

[1] Altbach, P.G., Reisberg, L., Rumbley, L.E.: Trends in Global Higher Education: Tracking an Academic Revolution. UNESCO Publishing, Paris (2009), https://unesdoc.unesco.org/ark:/48223/pf0000183219, a report prepared for the UNESCO 2009 World Conference on Higher Education

[2] Asselborn, T., Dörpinghaus, J., Kausar, F., Möller, R., Melzer, S.: Enhancing Text Recognition of Damaged Documents through Synergistic OCR and Large Language Models, pp. 29–36. Polish Information Processing Society (Sep 2024). https://doi.org/10.15439/2024F7400

[3] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate (2016), https://arxiv.org/abs/1409.0473

[4] Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909/

[5] Bundesregierung der Bundesrepublik Deutschland: Handwerksordnung (HwO) in der Neufassung vom 24. September 1998. Bundesgesetzblatt Teil I Nr. 67, Bonn (1998), http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl198s3074.pdf

[6] Bundesregierung der Bundesrepublik Deutschland: Berufsbildungsgesetz (BBiG) in der Neufassung vom 23. März 2005. Bundesgesetzblatt Teil I Nr. 20, Bonn (2005), https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl105s0931.pdf

[7] Busemeyer, M.R., Trampusch, C.: The Political Economy of Collective Skill Formation. Oxford University Press (11 2011). https://doi.org/10.1093/acprof:oso/9780199599431.001.0001

[8] Cedefop: Vocational education and training in Europe, 1995–2035: Scenarios for European vocational education and training in the 21st century. No. 114 in Cedefop reference series, Publications Office of the European Union, Luxembourg (2020). https://doi.org/10.2801/794471

[9] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805

[10] European Commission: Education Levels (nd), https://education.ec.europa.eu/education-levels, accessed: 2025-05-09

[11] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT Sentence Embedding (2022), https://arxiv.org/abs/2007.01852

[12] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast Neural Machine Translation in C++. In: Liu, F., Solorio, T. (eds.) Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-4020

[13] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast Neural Machine Translation in C++ (2018), https://arxiv.org/abs/1804.00344

[14] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703

[15] Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/

[16] Loper, E., Bird, S.: NLTK: the Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. p. 63–70. ETMTNLP '02, Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1118108.1118117

[17] Melhuish, E.C., Ereky-Stevens, K., Petrogiannis, K., Aricescu, A.M., Penderi, E., Rentzou, K., Tawell, A., Slot, P., Broekhuizen, M., Leseman, P.: A review of research on the effects of Early Childhood Education and Care (ECEC) upon child development. Technical report, European Commission (2015), https://eprints.bbk.ac.uk/id/eprint/16443/, cARE project; Curriculum Quality Analysis and Impact Review of European Early Childhood Education and Care (ECEC)

[18] Németh, L., Foundation, F.: Hunspell, https://hunspell.github.io/, accessed: 2025-02-15

[19] OECD: Education at a Glance 2021: OECD Indicators. OECD Publishing, Paris (2021). https://doi.org/10.1787/b35a14e5-en

[20] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135

[21] Radford, A., Narasimhan, K.: Improving Language Understanding by Generative Pre-Training. In: Computer Science, Linguistics (2018), https://api.semanticscholar.org/CorpusID:49313245

[22] Schiersmann, C.: Weiterbildungsberatung im Kontext der Nationalen Weiterbildungsstrategie: Finanzielle und strukturelle Aspekte. Hessische Blätter für Volksbildung **72**(1), 43–53 (2022)

[23] Smith, R.: An Overview of the Tesseract OCR Engine. In: ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition. pp. 629–633. IEEE Computer Society, Washington, DC, USA (2007), https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf

[24] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks (2014), https://arxiv.org/abs/1409.3215

[25] UNESCO, UNICEF: Global Report on Early Childhood Care and Education: The Right to a Strong Foundation. UNESCO and UNICEF, Paris (2024). https://doi.org/10.54675/FWQA2113

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2023), https://arxiv.org/abs/1706.03762

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 479–488
DOI: 10.15439/2025F8569
ISSN 2300-5963 ACSIS, Vol. 43

Blackcurrant Plantations Monitoring Using Drones

Jānis Bičevskis, Reinis Odītis,
Ivo Odītis, Zane Bičevska
0000-0001-5298-9859, 0009-0003-5488-1608
0000-0003-2354-3780, 0000-0002-5252-7336
University of Latvia
Raina Blvd. 19, Riga, Latvia
Email: janis.bicevskis@lu.lv, ro17020@students.lu.lv,
ivo.oditis@lu.lv, zane.bicevska@lu.lv

***Abstract*—The work is dedicated to the study of drone use in horticulture, focusing on an example of blackcurrant cultivation. The research aims to use drones to monitor vegetation in plantations and to maintain the technological environment of plants, using traditional agrotechnical field care methods. The concept offers mapping and instance segmentation followed by multi-label classification operations, taking into account the specifics of blackcurrant plantations. The mapping operation creates blackcurrant plantation maps from images taken by drones at low altitudes. This ensures the acquisition of high-quality maps of large areas with the help of simple image photography cameras. Instance segmentation is intended for extracting singular leaf instances from mapped images, which are analyzed using classification methods to detect blackcurrant diseases, pest spread, nutrient and moisture deficiencies, and other plant vegetation-related parameters. Classification employs machine learning techniques and is specific to the cultivation of a particular plants – blackcurrants. The proposed technology, with appropriate adjustments, can also be applied to the vegetation monitoring of other horticultural plants.**

I. Introduction

TODAY, drone manufacturing costs and production volumes have reached a level that allows their application in various fields, including precision agriculture. The scientific reviews available in the literature analyze the latest advances in drone technology used in precision farming. For example, [1] compiles 184 publications using Google Scholar and SCOPUS data, while [2] reviews 164 articles focusing on the application of AI methods in agriculture.

The authors of these studies unanimously conclude that drones can be used effectively to monitor plant vegetation, while agricultural operations should be carried out using ground-based equipment. During a flight, a drone can capture images of designated field areas and transmit them to computers for further analysis. Using machine learning methods, this enables the detection of vegetation conditions and issues that determine necessary agronomic actions. Thus, drones allow for rapid inspection of large agricultural areas and the collection of crucial data on crop health and required maintenance tasks.

This study adopts a practical approach with the primary objective of developing a technology that integrates drones and artificial intelligence techniques to monitor plant vegetation. The system is designed to provide early warnings to farmers regarding plant diseases, pest infestations, nutrient deficiencies, and other agronomic issues. A key contribution of this work is the development of a robust image analysis system capable of processing leaf images to reliably detect plant (particularly blackcurrant) diseases, thereby enhancing the system's applicability in real-world agricultural settings.

The following chapters provide a description of the technology designed to support blackcurrant cultivation using basic drones and imaging cameras. The proposed technology was tested on a blackcurrant farm in Latvia, confirming the validity of the chosen approach.

The structure of this study is as follows: Chapter 1 provides an overview of drone usage in precision agriculture worldwide. Chapter 2 focuses specifically on drone applications in horticulture. Chapter 3 presents the authors' proposed methodology for assessing blackcurrant plantations. Chapter 4 offers a visualization of blackcurrant plantation conditions. Chapter 5 discusses the results obtained and presents the conclusions.

II. Drone Application in Horticulture

To feed the rapidly growing global population, agricultural enterprises must produce more food without increasing the areas of cultivated land. This can be achieved by applying advanced farming technologies. Some of these technologies are still in development, while others are already offered by commercial companies. Today, farms can utilize a variety of advanced tools, such as satellite data, drones, autonomous platforms for agricultural operations, sensors, and robots, to obtain detailed information about crop and soil conditions and to perform specific agronomic tasks. However, in many countries, including Latvia, the adoption of drone technology in agriculture is still in its early stages. Among the various scientific and technological challenges that are being addressed to achieve the goals of sustainable development, the use of new technologies and methodologies in agriculture has attracted the interest of the engineering research community ([3]). The objective is to develop technologies suitable for precision agriculture that improve the long-term profitability and efficiency of agricultural production ([4]).

A. General Overview of Drone Applications

According to [1] and [2], data from the Food and Agriculture Organization (FAO) indicate that global food production

must increase by 70% by 2050 to support the growing world population. However, in the European Union, the number of people employed in agriculture has decreased by 35% over the past decade, and the expansion of agricultural land is largely unfeasible.

These factors have driven an increase in interest in advanced agricultural technologies, including sensors, robots, drones, digitalization, and artificial intelligence (AI). AI and machine learning are considered highly promising for detecting agricultural problems, monitoring crop health, forecasting yields and prices, mapping harvests, and optimizing pesticide and fertilizer use.

There are various research directions that discuss the use of modern technologies in agriculture: Internet of Things (IoT) technologies in agriculture ([5]), bibliometric analysis of drone use in agriculture ([6]), deep learning methods for controlled environment agriculture ([7]), robotic harvesting technologies (Mail et al., 2023), machine vision applications in agricultural robot navigation ([8]), AI in agriculture ([9]), Agriculture 4.0 ([10], [11]).

Depending on the specific task, drones can offer similar capabilities to satellite image analysis but with higher precision and flexibility. They can perform tasks such as soil analysis ([12], [13], [14]), monitoring sowing density and crop development ([15], [16]), weed and pest detection and classification ([17], [18], [19]) and yield prediction and maturity assessment ([20], [21], [22]).

In rare cases, drones can also be used for harvesting, precision fertilization ([23], [24], [25]), pesticide spraying ([26], [27], [28]) and even mechanical pest eradication. IoT and sensor technologies provide farmers with real-time data on soil parameters, temperature, atmospheric gases, weather conditions, and many other variables, often processed in cloud-based IT infrastructures for further analysis and forecasting ([29], [30], [31]).

B. Scope of Drone Applications

The use of artificial intelligence and cloud technology in drones has brought significant improvements to smart agriculture. These new technologies can capture high-resolution images, aerial maps, and thermal images, which can be utilized in various agricultural applications, including:

- Soil analysis: Drones can be used for soil sampling, analyzing soil moisture levels, and assessing soil quality, helping farmers optimize fertilization and irrigation processes,
- Planting: Drones can be used for precise seed sowing and/or seedling planting, reducing labour and planting material costs,
- Crop spraying: Drones equipped with spraying systems can be used for the precise distribution of pesticides, herbicides, and fertilizers, minimizing environmental impact while saving time and financial resources,
- Irrigation management: Drones equipped with thermal sensors and infrared cameras can identify areas needing irrigation, helping to optimize water use and reduce waste,
- Yield mapping: Drones can generate yield maps, assisting farmers in optimizing crop management and increasing overall production,
- Livestock monitoring: Drones equipped with cameras can be used to monitor livestock health and behaviour, as well as track animal locations,
- Crop monitoring: Drones equipped with sensors and cameras can collect real-time data on crop health, growth, and yield, creating crop health maps,
- Field mapping: Drones can create high-resolution field maps, providing data on soil structure, topography, and plant populations, which can be used for informed decision-making regarding planting, fertilization, and other crop management practices,
- Pest and disease control: Drones can help detect and map the spread of pests and diseases in crops, helping farmers take timely action.

These drone applications have gained significant research attention over the past five years. Studies provide evidence of the potential of drones in agriculture. However, these results remain a future vision that is not yet accessible to practitioners. Implementing such technologies requires the involvement of highly qualified specialists and the establishment of modern infrastructure.

III. Applications of Machine Learning in Plant Disease Diagnosis

A. Overview and Summary of Existing Solutions

The authors have reviewed several solutions for diagnosing plant diseases using images of plant leaves. The works discussed in this subsection will not be described individually, but rather grouped according to the nature of the solutions reviewed, and common features will be highlighted to indicate current trends in addressing such problems.

The solutions include methods such as the use of convolutional neural networks, the application of transfer learning, the use of traditional classification methods, the application of segmentation, object detection, and data augmentation.

Convolutional neural networks dominate plant leaf image classification in the context of the reviewed studies. In most of the research reviewed by the authors, convolutional neural networks are either trained from scratch on specific plant data or used as part of transfer learning. For example, in the work by Khalid M. Hosny and co-authors titled **"Multi-Class Classification of Plant Leaf Diseases Using Feature Fusion of Deep Convolutional Neural Network and Local Binary Pattern"** [32], several examples of convolutional neural network architectures - AlexNet, VGG16, and GoogLeNet - are analyzed. These are trained on the PlantVillage dataset [**datasetPlantVillage**] and enhanced with the use of local binary patterns to incorporate texture information. Another good example of convolutional neural network use is described in the work by Prabhjot Kaur and co-authors titled

"A novel hybrid CNN methodology for automated leaf disease detection and classification" [33], where a convolutional neural network (ResNet50) is combined with a support vector machine (replacing the fully connected layers of the convolutional neural network). An interesting application of convolutional neural networks is found in the publication by Rasika Gajendra Patil and Ajit More titled **" Grape Leaf Disease Diagnosis System Using Fused Deep Learning Features Based System"** [34], where the proposed method fuses the results obtained from the deep feature layers of two selected convolutional neural networks.

Similar to convolutional neural networks, transfer learning approaches are also commonly found in the reviewed solutions. They are widely used because they allow the utilization of pre-trained networks. An example is the study developed by Vinay Gautam and co-authors titled **A Transfer Learning-Based Artificial Intelligence Model for Leaf Disease Assessment** [35], where transfer learning is applied to models such as: AlexNet, VGG16, VGG19, InceptionV4, SqueezeNet, and Xception. The use of transfer learning is particularly important in situations where data collection is expensive or limited, such as in the case of rare diseases. A similar situation is observed in the study by Yan Guo, Jin Zhang, and co-authors titled **" Plant Disease Identification Based on Deep Learning Algorithm in Smart Farming"** [36], where the solution is based on pre-trained VGG16 and ResNet101 models. These models were fine-tuned for specific plant recognition using a dataset of 1000 images from the China Plant Image Bank.

Although deep learning models dominate the reviewed solutions, some of them still employ classical supervised learning methods. Techniques such as support vector machines, the k-nearest neighbors algorithm, and the random forest classifier are often used in combination with traditional feature extraction—such as histogram analysis, color moments, and shape descriptors [37],[38]. For example, in the previously mentioned work by Rasika Gajendra Patil and Ajit More [34], a convolutional neural network was initially applied to the dataset items, but the final classification was determined using a support vector machine classifier. Such methods are often easier to interpret and less demanding in terms of computational resources.

Similar to the use of convolutional neural networks and transfer learning, a commonly observed trend in solutions is the use of semantic or instance segmentation prior to classification in order to achieve better results. In the publication by Gangadevi Ganesan and Jayakumar Chinnappan titled **"Hybridization of ResNet with YOLO classifier for automated paddy leaf disease recognition: An optimized model"** [39], segmentation of diseased leaf regions was performed using K-means clustering. Similarly, in the publications by Muhammad Sharif and co-authors and by Vinay Gautam and co-authors [40], [35], semantic segmentation of the diseased regions was applied. This indicates that, in plant leaf disease diagnosis, it is valuable to use only the regions showing anomalies for classification, rather than the entire leaf instance. However, such an approach is effective only in solutions where the processed dataset contains images in which each image depicts only a single leaf instance. In the solution described in the publication by Marcelo Vassallo-Barco and co-authors titled **"Automatic Detection of Nutritional Deficiencies In Coffee Tree Leaves Through Shape And Texture Descriptors"** [41], the Otsu thresholding method was used for instance segmentation, however, the publication does not specify the exact accuracy achieved by the proposed method, making it difficult to draw significant conclusions from this study. Instance segmentation is also found in the solution mentioned in the publication by F. Khan and co-authors titled **"A mobile-based system for maize plant leaf disease detection and classification using deep learning"** [37], where the YOLO model family is used. In this context, instance segmentation is applied to identify and classify leaf instances, including their masks, since the dataset images depict multiple leaves. This approach provides high data granularity and allows for the analysis of individual leaves.

Additionally, it should be noted that the analyzed solutions also include image preprocessing methods and various data augmentation techniques. Given the frequent lack of data, different data augmentation methods are used - such as image rotation, zooming, cropping, and others. These methods help reduce the risk of overfitting and improve the generalization ability of the models.[36], [41]

B. Conclusions for the Selection of a Practical Solution

After analyzing and comparing various existing solutions, the author chose to apply convolutional neural networks (CNN) for leaf classification and instance segmentation to separate leaf instances prior to classification in the development of a practical solution. Since the previously mentioned work using YOLO models demonstrated high precision [37], the author selected the YOLO model family for the instance segmentation task and, based on the strong performance of ResNet models [34], chose ResNet models for the classification task. At the same time, it can be concluded that for each plant species, a specifically trained model must be developed because plants differ visually and various diseases may manifest differently. Therefore, the prototype of the study's solution focuses on only one plant species - blackcurrants - on which the proposed solution is tested.

IV. Proposed Solution of Technology

This section presents the authors' proposed technology for monitoring blackcurrant vegetation using simple drones and artificial intelligence methods. The proposed solution is applied in a blackcurrant farming operation in Latvia.

A. Informal Description of Solution

The technology is offered to blackcurrant growers to monitor plantations using drones, allowing detection of healthy blackcurrant plants, fungal diseases, and nutritional deficiencies. The system consists of two main functions:

- Mapping: Prepares maps of blackcurrant plantations with the required precision (scale), links them to GPS (Global



Fig. 1. Blackcurrant plantation photo from 5 meters height (row spacing – 4 meters).

Positioning System) coordinates, records drone flight routes, and specifies operations/photography to be performed during flights.

- Analyze and classification: Uses trained neural networks to extract blackcurrant leaf clusters from the mapped images. The instance segmentation then identifies individual blackcurrant leaves, which are passed to the classification stage for further analysis. As a result, three types of leaves are identified: healthy leaves, leaves affected by fungal diseases, and leaves indicating nutrient deficiencies. As a result, the study provides a blackcurrant plantation analysis solution that gathers information on plantation conditions and visualizes it for growers, aiding decision-making on necessary interventions.

B. Mapping and Planning of Image Capturing Routes for of Horticultural Areas

The area captured by drones for image analysis is usually significantly smaller than the cultivated field area in horticulture. This is determined by the technical parameters of the drone camera and the scale of the captured images Fig. (1). To obtain complete information about the field, the authors took multiple photographs every 10 meters. Although the images do not fully cover the entire field (each image captures an area of approximately 8×5 meters), they still provide a sufficient overview of the field as a whole, which typically ranges in size from 2 to 20 hectares.

To obtain the necessary images, a drone flight route with a photography plan must be created for each monitored field. The route can be created using the Mission Planner program (web-03). The overall image route is calculated on the basis of the drone's flight altitude and camera parameters (e.g., the angle the camera captures). As a result, a route will be obtained for the drone to follow, an example of which can be seen in the image prepared by the authors in Fig. 2.

The automatic flight is provided by the Litchi program. It is available on both the computer's website (web-04) to create the route and on mobile phones (web-05) to fly the route.

During the experiments, *Dji Mavic Air 2S* with a 4k camera was used. All pictures were taken from a height of 5m. To



Fig. 2. Drone mission plan: route and photography points

carry out the mission plan shown in Fig. 2 (which covered 2.5 hectares), approximately 15 minutes were required (time may be affected by wind speed), and one drone battery was used (i.e., the flight was continuous).

C. Image Analyzes and Classification

For the analysis of blackcurrant leaf images, the authors have chosen a solution that combines instance segmentation and multi-label classification. Instance segmentation ensures the separation of leaf instances from the background, while multi-label classification assigns each leaf instance to one or more of the predefined classes. For instance segmentation, the authors trained two models from the YOLO family — YOLOv8n-seg and YOLOv9c-seg. For multi-label classification, the authors trained three models from the ResNet family — ResNet-50, ResNet-101, and ResNet-152. Based on the training results, the authors selected one model for each task (instance segmentation, multi-label classification) and integrated the segmentation and classification models to form a unified solution for blackcurrant leaf analysis.

1) Image Acquisition: The authors independently performed the acquisition of specific images used to train the models applied in the proposed solution. All the images used in the training, validation, and test datasets, as described in subsection IV-C3, were captured using a *Nikon D3300 DSLR* camera equipped with a 24.2 MP DX-format CMOS sensor. The images (sample image in Fig. 3) were taken from the same blackcurrant field, capturing photographs from various angles and under different lighting conditions to ensure dataset diversity. This diversity can enhance model training results and improve the robustness of the trained models in non-standard scenarios.

2) Image Annotation: Image annotation for instance segmentation and multi-class classification was performed using



Fig. 3. Image of blackcurrant leaves

various methods to support the formats required by the specific machine learning models.

For the instance segmentation datasets, each image was manually annotated by outlining the contour of each individual blackcurrant leaf instance. Annotation was performed using the open source image and video annotation tool *Computer Vision Annotation Tool (CVAT)*, developed by *Intel*. The annotations were then exported as point coordinates in the *YOLOv8 segmentation 1.0* format. This type of data annotation is time-consuming as images of blackcurrant bushes contain a large number of leaf instances that must be precisely annotated. Such accuracy is essential for the model to correctly detect and isolate leaves from the background and overlapping objects, which is critical for accurate instance segmentation. One of resulting images is available in Fig. 4)

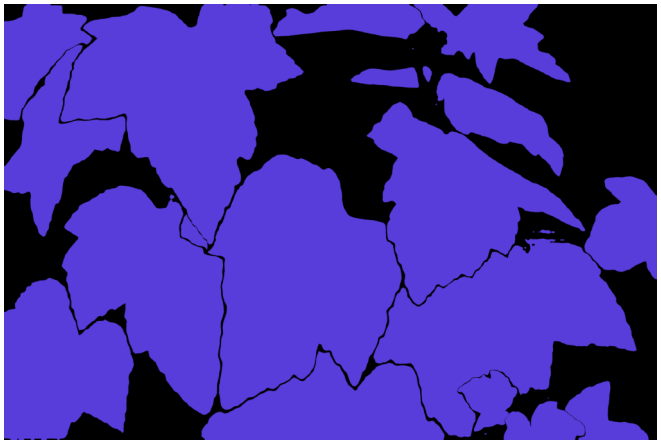


Fig. 4. Instance segmentation annotation example

For multi-label classification, annotation was performed by isolating individual instances from blackcurrant leaf images using the *Adobe Photoshop* graphic editing software, creating images where each blackcurrant leaf instance was placed against a black background (sample Fig. 5). Each resulting image was then assigned a label indicating its correspondence to one or more of the predefined blackcurrant health condition classes. The specific classification scheme — which includes the classes: healthy leaves (HL), leaves with nutrient deficiency (NDL), and leaves affected by *Mycosphaerella ribis* (MRL) — was developed in collaboration with an industry expert, based on an assessment of the blackcurrant field available to the author. A more detailed overview of the classes can be found in subsection IV-C3. All image labels were compiled into a structured file containing the image names and their corresponding class assignments.



Fig. 5. Multi-label classification dataset example: leaf with nutrient deficiency

3) Datasets overview: This work utilizes two main datasets to support different stages of the proposed method: one for multi-label classification and another for instance segmentation. The multi-label classification dataset contains 287 images, each featuring a single blackcurrant leaf against a black background. These images are divided into training (229), validation (49), and test (49) subsets. Each leaf is labeled according to a classification scheme developed in collaboration with an industry expert, categorizing them as healthy leaves (HL), leaves with nutrient deficiency (NDL), or leaves affected by Mycosphaerella ribis (MRL). Since a single leaf may exhibit symptoms of multiple classes, a multi-label classification approach is used. The class distribution is provided in a Table I, showing that the total number of labeled instances exceeds the number of images due to overlapping class characteristics.

The instance segmentation dataset comprises 87 images, split into 71 for training and 16 for validation. Each image includes only blackcurrant leaves, and each leaf has been manually annotated with precise contours. This detailed annotation process is resource-intensive but essential, as it enables the

TABLE I
DISTRIBUTION OF CLASS OCCURRENCES ACROSS TRAINING, VALIDATION, AND TEST SETS

Class	Training (229 images)	Validation (49 images)	Test (49 images)
HL	92	21	19
NDL	85	19	16
MRL	80	14	18
Total Labels	257	54	53

segmentation model to accurately detect and isolate individual leaves from complex backgrounds.

4) Model training overview: For the instance segmentation training process, the following models were used: YOLOv8n-seg and YOLOv9c-seg [42]. These were trained using the open-source Ultralytics package in Python [43], which provides an efficient workflow for YOLO models, including automated hyperparameter configuration such as learning rate adjustment based on gradient changes, default use of the Adam optimizer, and automatic batch size selection based on available GPU memory — set to 8 in this case. Each model was trained with three different image resolutions: 256x256, 512x512, and 1024x1024. This approach allowed the author to evaluate model behavior across varying resolutions and assess the impact of image size on instance segmentation accuracy. Based on the training loss analysis, 70 epochs were chosen for the final training.

For the multiclass classification training process, the following models were used: ResNet-50, ResNet-101, and ResNet-152. These models were implemented using the PyTorch machine learning library in Python and initialized with default pretrained weights to leverage feature representations learned from large-scale datasets. The final fully connected layer of each ResNet model was replaced to enable prediction of three target classes, making them suitable for the classification of disease instances on blackcurrant leaves. Input images were resized to 256×256 pixels and transformed into tensors for model processing. The models were trained using the BCEWithLogitsLoss loss function, combining Sigmoid activation with binary cross-entropy for numerical stability in multiclass tasks. The Adam optimizer was chosen for its ability to dynamically adjust learning rates and incorporate momentum. To identify optimal hyperparameters, a grid search was conducted over the number of epochs, learning rate, and batch size for each ResNet variant. The best-performing configurations were: 70 epochs, a learning rate of 0.0001, and batch sizes of 16 (ResNet-50 and ResNet-152) or 8 (ResNet-101).

5) Model training results: The training results of the instance segmentation models YOLOv8n-seg and YOLOv9c-seg are summarized in Tables II and III, showing their performance across different image resolutions (256×256, 512×512, 1024×1024). Both box (B) and mask (M) performance metrics were evaluated, including mAP50, mAP50–95, and precision (Pre), which together assess the models' ability to localize and delineate object instances at the pixel level. The results indicate that YOLOv9c-seg consistently outperformed YOLOv8n-seg.

across all tested resolutions and metrics. The highest performance was observed at the highest resolution (1024×1024), where YOLOv9c-seg achieved mAP50 values of 0.72130 for boxes and 0.69930 for masks. Additionally, an increase in input image resolution correlates with improved model performance for both architectures, emphasizing the significance of higher resolution imagery in enhancing segmentation precision. This is particularly important within the context of this study, where precise delineation of blackcurrant leaf instances from the background is essential for subsequent classification tasks. Nonetheless, the results are not without limitations, likely attributable to the relatively small size of the training dataset employed.

TABLE II
YOLOV8N INSTANCE SEGMENTATION RESULTS

Model	YOLOv8n-seg		
Image size	256x256	512x512	1024x1024
mAP50 (B)	0.59164	0.65656	0.67211
mAP50-95 (B)	0.38272	0.49653	0.52423
Pre (B)	0.67455	0.73628	0.74736
mAP50 (M)	0.58450	0.65931	0.66956
mAP50-95 (M)	0.34303	0.47840	0.51023
Pre (M)	0.74249	0.76106	0.75789

TABLE III
YOLOV9C-SEG INSTANCE SEGMENTATION RESULTS

Model	YOLOv9c-seg		
Image size	256x256	512x512	**1024x1024**
mAP50 (B)	0.64149	0.68866	**0.72130**
mAP50-95 (B)	0.46194	0.54670	**0.68643**
Pre (B)	0.74662	0.87470	**0.88182**
mAP50 (M)	0.65124	0.69849	**0.69930**
mAP50-95 (M)	0.47340	0.51147	**0.53121**
Pre (M)	0.72102	0.79023	**0.81230**

Table IV summarizes the training performance of three ResNet architectures (ResNet-50, ResNet-101, and ResNet-152) on a multi-label classification task. Among the evaluated models, *ResNet-152* achieved the highest overall performance, with accuracy, sensitivity, specificity, and precision reaching 94.0%, 96.7%, 98.8%, and 98.2% respectively. *ResNet-101* also demonstrated strong results, attaining 93.1% accuracy and 96.5% precision, while *ResNet-50* showed slightly lower accuracy (89.6%) but competitive sensitivity and precision,

and even outperformed *ResNet-101* in specificity. The results suggest that increasing model complexity yields diminishing returns, as the simpler *ResNet-50* architecture delivered satisfactory performance. The similar metric values across architectures imply that dataset characteristics, such as size and class distribution, may exert greater influence on model efficacy than network depth. Consequently, *ResNet-50* offers a favorable balance between accuracy and computational efficiency, making it a practical choice for this classification task.

TABLE IV
MULTI-LABEL CLASSIFICATION RESULTS TABLE

Model	ResNet-50	ResNet-101	ResNet-152
Acc (%)	89.6	93.1	94.0
Sen (%)	91.3	94.8	96.7
Spe (%)	98.2	98.1	98.8
Pre (%)	91.3	96.5	98.2
Epochs	70	70	70
Batch size	16	8	16
Lr	0.0001	0.0001	0.0001

Summarizing the instance segmentation and multilabel classification model training results, the authors selected YOLOv9c-seg for instance segmentation and ResNet-50 for multilabel classification as the final practical solution.

V. HISTORY OF FIELDS

For every farm, it is beneficial to maintain a field record journal that logs all activities within a specific agricultural area, including completed horticultural operations, the use of crop materials, and plant protection products. Several information systems already exist to support such functionalities.

However, this study focuses on collecting field images, offering a different perspective on historical data, both visually and through insights derived from image analysis. The goal of our research is to develop a solution that allows visual tracking of field changes over time while also providing timely detection of plant health issues identified through image analysis.

To ensure field history tracking:

- identify the surveyed fields (this information is used to plan drone flight and photography routes),
- conduct field surveys using drones to capture images and link them to specific geolocations,
- analyze the images to identify plant health issues,
- visualize the extracted information on a map for better interpretation and decision-making.

The results of the image analysis conducted on a single day are presented in Fig. 6. As shown, within the selected location, 6.51% of the identified leaves were classified as diseased, 0.38% exhibited signs of nutrient deficiency, and 93.10% were determined to be healthy.

This type of solution is designed for horticulturists. Their main interest is tracking long-term changes in fields, especially in crop cultivation involving perennial plants such as



Fig. 6. Blackcurrant field condition analysis result

blackcurrants. In such cases, even historical images taken from the same vantage point can provide valuable insights into field conditions, moisture levels, pest infestations, disease development, and more.

Although the functionalities may seem simple, this type of solution comes with certain technical challenges. Even if the captured images do not cover the entire field area, their number is sufficiently large, and storing historical images can be space-intensive, considering that blackcurrant plantations typically cover between 2 to 20 hectares. Moreover, analyzing such a vast number of images is time-consuming as it involves identifying and classifying thousands of leaves.

The problem can be partially solved by increasing the decreasing of the image capture grid, e.g., from 10 to 20 meters. A similar approach is used in the evaluation of soil fertility, where the sampling is carried out systematically at predetermined intervals. In this case of 20 meters, evaluating one hectare would require only 36 images, significantly reducing storage and analysis demands. This approach would still provide sufficiently representative information about the field's condition and the spread of potential issues.

VI. DISCUSSION

The study results demonstrate new technologies for monitoring blackcurrant vegetation. Instead of traditional visual assessments by horticulturists, an automated system is proposed, offering several advantages: precise plant evaluation using AI methods, applicability to large blackcurrant plantations, economic efficiency, and ease of implementation.

However, these achievements should be considered a first step toward precision agriculture in blackcurrant cultivation, requiring further development. The proposed approach relies on blackcurrant leaf analysis, which means it can only detect diseases affecting leaves, while issues affecting roots and stems - such as blackcurrant clearwing moth (*Synanthedon tipuliformis*) and blackcurrant bud mite (*Eriophyes ribis*) - remain undetected. Similarly, yield prediction requires an

alternative approach, possibly analyzing entire blackcurrant bushes rather than just leaves. Pest infestations can also only be partially identified through leaf analysis. Nevertheless, other blackcurrant cultivation challenges, such as crown rust, can be identified using similar methods by analyzing different plant parts, segmenting them in images, and classifying them based on the specific problem being addressed.

A. Additional Applications of the Method

A diligent horticulturist monitors not only the spread of blackcurrant diseases but also frost damage, flowering progress, yield ripening time and volume predictions, and other vegetation-related events. Although these aspects were not the primary focus of this study, they could be addressed by modifying the proposed method—for example, by segmenting and classifying flower buds and berry clusters accordingly.

Beyond plant vegetation monitoring, the method can also be applied to optimizing agricultural operations. By identifying disease-affected field areas, maintenance tasks such as targeted spraying can be carried out only in infected regions. This would lead to significant savings in materials and labor resources.

A promising direction for further development is integrating the method into dynamic robotic management. By transferring real-time data from the blackcurrant plant analysis module to an agricultural operations execution robot, it would be possible to perform precise interventions only where necessary, further increasing efficiency and sustainability.

B. Limitations of the Method's Application

When analyzing the benefits of the proposed method, it is also important to highlight its limitations.

One key limitation is the lack of precision in determining nutrient deficiencies. While the method can detect a deficiency, it cannot specify which particular nutrient — potassium, phosphorus, or nitrogen — is lacking. Currently, this type of analysis is performed using soil and plant agrochemical testing, which involves manually collecting soil and leaf samples. This process requires significant labor resources.

A more advanced approach involves spectral analysis of plants, which can provide more precise diagnosis of nutrient deficiency. However, this requires more complex imaging cameras.

In addition, leaves are only one indicator of plant health, but they do not reveal all potential problems. For example, pest infestations, such as the blackcurrant clearwing moth, which primarily affects the stems rather than the leaves, cannot be detected using this method.

VII. Conclusions

Key conclusions of the conducted research and its application results:

1) Blackcurrant plantation monitoring is recommended to be carried out in two stages: (1) field mapping – capturing plantation field images with the required precision; (2) image instance segmentation and multi-label classification – an instance segmentation model extracts individual leaf instances from the mapped images, and a multi-label classification model employs a trained neural network to identify fungal diseases and nutrient deficiencies by assigning one or more class labels to each leaf.
2) Mapping functions can be implemented using standard solutions available in commercial drone systems or by adapting open source solutions for blackcurrant cultivation. Segmentation solutions must be developed individually in collaboration with industry experts, in the case of this project, blackcurrant growers. This includes training a neural network for leaf recognition and transmitting the identified leaf information for classification. The classification task involves training a neural network to recognize specific characteristics of blackcurrant leaves, which is the final goal of the project and is of interest to horticulturists.
3) The final result consists of blackcurrant field images that visually distinguish between healthy areas, fungal disease-affected zones, and nutrient-deficient regions. The field history application provides horticulturists with an intuitive tool to monitor blackcurrant crop development over time
4) Scientific literature and practical experience confirm that drone-based plant condition monitoring is effective. Collected crop image datasets have significant potential for machine learning applications.

Following the identification of fungal diseases and nutrient deficiencies, future potential applications include prediction of yield, detection of pests, and identification of other plant diseases. A key challenge remains in improving the speed of image analysis, as computational demands can exceed the capabilities of simple and inexpensive hardware.

VIII. Acknowledgement

This work has been conducted within the research project "Competence Centre of Information and Communication Technologies" of The Recovery and Resilience Facility, contract No. 5.1.1.2.i.0/1/22/A/CFLA/008 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 1.8 “Innovative Use of Drones in Horticulture".

IX. References

References

[1] A. Botta, P. Cavallone, L. Baglieri, G. Colucci, L. Tagliavini, and G. Quaglia, “A review of robots, perception, and tasks in precision agriculture”, Applied Mechanics **3**, 830–854 (2022).

[2] A. Uzhinskiy, “Advanced technologies and artificial intelligence in agriculture”, AppliedMath **3**, 799–813 (2023).

[3] J. D. B. Gil, P. Reidsma, K. Giller, L. Todman, A. Whitmore, and M. van Ittersum, "Sustainable development goal 2: improved targets and indicators for agriculture and food security", Ambio **48**, 685–698 (2019).

[4] P. De Castro, P. P. Miglietta, Y. Vecchio, et al., "The common agricultural policy 2021-2027: a new history for european agriculture", Rivista Di Economia Agraria **75**, 5–12 (2020).

[5] J. Xu, B. Gu, and G. Tian, "Review of agricultural iot technology", Artificial Intelligence in Agriculture **6**, 10–22 (2022).

[6] A. Rejeb, A. Abdollahi, K. Rejeb, and H. Treiblmaier, "Drones in agriculture: a review and bibliometric analysis", Computers and Electronics in Agriculture **198**, 107017 (2022).

[7] M. O. Ojo and A. Zahid, "Deep learning in controlled environment agriculture: a review of recent advancements, challenges and prospects", Sensors **22**, 10.3390/s22207965 (2022).

[8] T. Wang, B. Chen, Z. Zhang, H. Li, and M. Zhang, "Applications of machine vision in agricultural robot navigation: a review", Computers and Electronics in Agriculture **198**, 107085 (2022).

[9] R. C. d. Oliveira and R. D. d. S. e. Silva, "Artificial intelligence in agriculture: benefits, challenges, and trends", Applied Sciences **13**, 10.3390/app13137405 (2023).

[10] M. A. Dayıoğlu and U. Turker, "Digital transformation for sustainable future - agriculture 4.0: a review", Journal of Agricultural Sciences **27**, 373–399 (2021).

[11] R. Abbasi, P. Martinez, and R. Ahmad, "The digitization of agricultural industry – a systematic literature review on agriculture 4.0", Smart Agricultural Technology **2**, 100042 (2022).

[12] J. Huuskonen and T. Oksanen, "Soil sampling with drones and augmented reality in precision agriculture", Computers and Electronics in Agriculture **154**, 25–35 (2018).

[13] J. Zhou, Y. Xu, X. Gu, T. Chen, Q. Sun, S. Zhang, and Y. Pan, "High-precision mapping of soil organic matter based on uav imagery using machine learning algorithms", Drones **7**, 10.3390/drones7050290 (2023).

[14] L. Bertalan, I. Holb, A. Pataki, G. Négyesi, G. Szabó, A. Kupásné Szalóki, and S. Szabó, "Uav-based multispectral and thermal cameras to predict soil water content – a machine learning approach", Computers and Electronics in Agriculture **200**, 107262 (2022).

[15] "Assessment of plant density for barley and wheat using uav multispectral imagery for high-throughput field phenotyping", Computers and Electronics in Agriculture **189**, 106380 (2021).

[16] J. C. O. Koh, M. Hayden, H. Daetwyler, and S. Kant, "Estimation of crop plant density at early mixed growth stages using uav imagery", Plant Methods **15**, 64 (2019).

[17] P. Ong, K. S. Teo, and C. K. Sia, "Uav-based weed detection in chinese cabbage using deep learning", Smart Agricultural Technology **4**, 100181 (2023).

[18] E. C. Tetila, B. B. Machado, G. Astolfi, N. A. de Souza Belete, W. P. Amorim, A. R. Roel, and H. Pistori, "Detection and classification of soybean pests using deep learning with uav images", Computers and Electronics in Agriculture **179**, 105836 (2020).

[19] N. A. Mohidem, N. N. Che'Ya, A. S. Juraimi, W. F. Fazlil Ilahi, M. H. Mohd Roslim, N. Sulaiman, M. Saberioon, and N. Mohd Noor, "How can unmanned aerial vehicles be used for detecting weeds in agricultural fields?", Agriculture **11**, 10.3390/agriculture11101004 (2021).

[20] C. Kumar, P. Mubvumba, Y. Huang, J. Dhillon, and K. Reddy, "Multi-stage corn yield prediction using high-resolution uav multispectral data and machine learning models", Agronomy **13**, 10.3390/agronomy13051277 (2023).

[21] L. Zeng, G. Peng, R. Meng, J. Man, W. Li, B. Xu, Z. Lv, and R. Sun, "Wheat yield prediction based on unmanned aerial vehicles-collected red–green–blue imagery", Remote Sensing **13**, 10.3390/rs13152937 (2021).

[22] T. B. Shahi, C.-Y. Xu, A. Neupane, D. B. Fleischfresser, D. J. O'Connor, G. C. Wright, and W. Guo, "Peanut yield prediction with uav multispectral imagery using a cooperative machine learning approach", Electronic Research Archive **31**, 3343–3361 (2023).

[23] P. Chen, F. Ouyang, Y. Zhang, and Y. Lan, "Preliminary evaluation of spraying quality of multi-unmanned aerial vehicle (uav) close formation spraying", Agriculture **12**, 10.3390/agriculture12081149 (2022).

[24] C. Song, L. Liu, G. Wang, J. Han, T. Zhang, and Y. Lan, "Particle deposition distribution of multi-rotor uav-based fertilizer spreader under different height and speed parameters", Drones **7**, 10.3390/drones7070425 (2023).

[25] D. Su, W. Yao, F. Yu, Y. Liu, Z. Zheng, Y. Wang, T. Xu, and C. Chen, "Single-neuron pid uav variable fertilizer application control system based on a weighted coefficient learning correction", Agriculture **12**, 10.3390/agriculture12071019 (2022).

[26] K. Anand and R. Goutam, "An autonomous uav for pesticide spraying", Int. J. Trend Sci. Res. Dev **3**, 986–990 (2019).

[27] S. Ivić, A. Andrejčuk, and S. Družeta, "Autonomous control for multi-agent non-uniform spraying", Applied Soft Computing **80**, 742–760 (2019).

[28] J. Sinha, "Aerial robot for smart farming and enhancing farmers' net benefit", in (ICAR, 2020).

[29] M. Dhanaraju, P. Chenniappan, K. Ramalingam, S. Pazhanivelan, and R. Kaliaperumal, "Smart farming: internet of things (iot)-based sustainable agriculture", Agriculture **12**, 10.3390/agriculture12101745 (2022).

[30] A. Madushanki, M. Halgamuge, W. Wirasagoda, and A. Syed, "Adoption of the internet of things (iot) in agriculture and smart farming towards urban greening: a review", English, International Journal of Advanced Computer Science and Applications **10**, 11–28 (2019).

[31] G. Bilotta, E. Genovese, R. Citroni, F. Cotroneo, G. M. Meduri, and V. Barrile, "Integration of an innovative atmospheric forecasting simulator and remote sensing data into a geographical information system in the frame of agriculture 4.0 concept", AgriEngineering **5**, 1280–1301 (2023).

[32] K. M. Hosny, W. M. El-Hady, F. M. Samy, E. Vrochidou, and G. A. Papakostas, "Multi-class classification of plant leaf diseases using feature fusion of deep convolutional neural network and local binary pattern", IEEE Access **11**, 62307–62317 (2023).

[33] P. Kaur, A. M. Mishra, N. Goyal, S. K. Gupta, A. Shankar, and W. Viriyasitavat, "A novel hybrid cnn methodology for automated leaf disease detection and classification", Expert Systems **41**, e13543 (2024).

[34] R. Patil and A. More, "Grape leaf disease diagnosis system using fused deep learning features based system", Procedia Computer Science **235**, 372–382 (2024).

[35] V. Gautam, N. K. Trivedi, A. Singh, H. G. Mohamed, I. D. Noya, P. Kaur, and N. Goyal, "A transfer learning-based artificial intelligence model for leaf disease assessment", Sustainability **14**, 10 . 3390 / su142013610 (2022).

[36] Y. Guo, J. Zhang, C. Yin, X. Hu, Y. Zou, Z. Xue, and W. Wang, "Plant disease identification based on deep learning algorithm in smart farming", Discrete Dynamics in Nature and Society **2020**, 1–11 (2020).

[37] F. Khan, N. Zafar, M. N. Tahir, M. Aqib, H. Waheed, and Z. Haroon, "A mobile-based system for maize plant leaf disease detection and classification using deep learning", Frontiers in Plant Science **14**, 10.3389/fpls.2023.1079366 (2023).

[38] Y. Ai, C. Sun, J. Tie, and X. Cai, "Research on recognition model of crop diseases and insect pests based on deep learning in harsh environments", IEEE Access **8**, 171686–171693 (2020).

[39] G. Ganesan and J. Chinnappan, "Hybridization of resnet with yolo classifier for automated paddy leaf disease recognition: an optimized model", Journal of Field Robotics **39**, 1085–1109 (2022).

[40] M. Sharif, M. A. Khan, Z. Iqbal, M. F. Azam, M. I. U. Lali, and M. Y. Javed, "Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection", Computers and Electronics in Agriculture **150**, 220–234 (2018).

[41] M. Vassallo-Barco, L. Vives, V. Tuesta-Monteza, H. Mejia, and R. Yera Toledo, "Automatic detection of nutritional deficiencies in coffee tree leaves through shape and texture descriptors", Journal of Digital Information Management **15**, 7–18 (2017).

[42] S. A. A. Qadri, N.-F. Huang, T. M. Wani, and S. A. Bhat, "Plant disease detection and segmentation using end-to-end yolov8: a comprehensive approach", in 2023 ieee 13th international conference on control system, computing and engineering (iccsce) (2023), pp. 155–160.

[43] U. LLC, *Ultralytics yolo: real-time object detection and segmentation*, 2023.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 489–497
DOI: 10.15439/2025F6951
ISSN 2300-5963 ACSIS, Vol. 43

A Serious Game for Teaching Entrepreneurship and Strategic Management

Boyan Bontchev
0000-0002-8554-2188
Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, Sofia 1164, Bulgaria
Email: bbontchev@fmi.uni-sofia.bg

Sia Tsolova
0000-0002-8199-6465
Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 5, J. Bourchier Blvd, Sofia 1164, Bulgaria
Email: siyat@fmi.uni-sofia.bg

Dessislava Vassileva
0000-0003-3670-8599
Scientific Research Center, Sofia University "St. Kliment Ohridski", 8, Dragan Tzankov Blvd, Sofia 1164, Bulgaria
Email: ddessy@gmail.com

***Abstract*—Educational video games have demonstrated significant potential as a modern complement to traditional e-learning in higher education. However, the scarcity of free tools and platforms for creating and customizing educational games impedes their widespread adoption in teaching. This paper introduces an educational video game focused on entrepreneurship, developed using a custom-designed software platform for generating and personalizing learning-based video games through 3D mazes. This platform empowers educators to effortlessly construct hierarchical, personalized 3D mazes for interactive presentation of multimedia educational content. Learners engage with new ideas, concepts, and theories while navigating the 3D maze, and can participate in various embedded mini-games designed to enhance visual and spatial reasoning, psychomotor skills, and context-based reasoning—all of which are vital for successful entrepreneurship. The results obtained from the study are highly promising, particularly regarding the usability and effectiveness of the created 3D video game.**

***Index Terms*—serious game, maze, entrepreneurship, strategic management, GEQ.**

I. Introduction

SINCE the 1970s, computer games have continued to grow, evolving from simple text-based formats into complex, visually rich experiences. Modern computer games make extensive use of two-dimensional (2D) and three-dimensional (3D) video technologies, along with embedded multimedia content, to create immersive and interactive environments. Various genres of video games can now be identified [1], all falling into two main categories: entertainment and serious games. Entertainment games are designed primarily to entertain players [2]. In contrast, serious games (also known as applied games) are created with goals beyond entertainment, integrating entertainment with education, business, healthcare, public order, culture, and other areas [3]. These games have been successfully implemented as educational tools [4], learning environments for social reintegration of marginalized youth [5], rehabilitation platforms for different patient groups [6], and systems for improving cognitive and social activities in the elderly, as well as for telerehabilitation and prevention [7]. The increasing sophistication of serious games demonstrates their potential to engage users in meaningful ways, making learning more interactive, accessible, and effective across disciplines. This sophistication is supported by advanced research in areas such as game AI [8], underlying game theory [9], and even security considerations in game design [10].

This study focuses on innovative approaches to game-based learning in higher education, especially in entrepreneurship education. The target group was chosen due to the growing importance of entrepreneurship education as a key driver of economic growth and social progress [11]. Despite the importance of entrepreneurship education, the availability of educational games in this area remains limited. Furthermore, there is a lack of free software platforms that allow educators without technical expertise to create engaging educational games. This problem is why it is difficult to integrate game-based learning into entrepreneurship courses, as developing such games requires specialized IT knowledge.

If this challenge is met, game-based learning can enhance entrepreneurship courses by helping students develop domain-specific skills, such as critical thinking, problem solving, and innovation. This paper presents STRAMAG, an innovative educational video game designed to bridge this gap. Developed using an open-source, adaptive platform, STRAMAG uses 3D mazes to deliver personalized and hierarchical strategic management training within an entrepreneurship curriculum. Unlike generic educational games, STRAMAG focuses on developing specific skills necessary for entrepreneurial success. Learners navigate the 3D maze, engaging with multimedia content and participating in embedded mini-games designed to improve visual-spatial thinking, psychomotor skills, and context-based decision-making skills directly applicable to real-world entrepreneurial challenges.

The STRAMAG game was created as a customizable video maze using Brainstorm eStudio & API[1] and enabling educators with no IT skills to create engaging, hands-on learning experiences. An experiment with this game was conducted,

[1] https://www.brainstorm3d.com/products/estudio/

and the preliminary results of this study demonstrate the potential of this platform and game.

This paper has two primary objectives:

1. To introduce and describe STRAMAG, a novel adaptive 3D educational game designed specifically for entrepreneurship education;
2. To evaluate the game experience of students playing different mini-games with style-based adaptation of learning content inside a maze using the Game Experience Questionnaire (GEQ).

These objectives lead to the following research question: *How do different video mini-games that adapt educational content to learning styles affect student engagement*?

The paper is organized as follows: Section II reviews relevant literature on serious games, 3D maze games, entrepreneurship education, and game experience evaluation. Section III details the STRAMAG game's design, structure, and content. Section IV describes the experimental setup. Section V presents the results from the game experience evaluation. Section VI discusses the implications of these findings, and Section VII concludes with limitations and directions for future research.

II. Background

A. Serious games for education

Serious games are defined as tools that add entertainment elements to learning. These games are considered an alternative to entertainment games, although there is evidence supporting the acquisition of various skills and abilities through the latter, such as spatial thinking, strategic skills, and memory development [12]. Recent meta-analyses indicate that serious games can increase knowledge retention compared to traditional instructional methods [13], with particularly strong effects in applied fields like business and management [14].

The scenarios of serious video games involve problem-solving and decision-making activities, usually through content presented in three-dimensional form. One of the most intriguing challenges in designing serious learning games is how the player can make meaningful and beneficial choices in a serious context. One key approach to address this is through adaptation in these games. Adaptation in serious games operates at multiple levels: content adaptation (adjusting material based on prior knowledge), difficulty adaptation (modifying challenge levels based on performance), and path adaptation (altering the sequence of learning activities [15]. However, Mayer et al. [16] identified significant limitations in current adaptation mechanisms, noting that different player types cannot fully utilize their qualification levels, playing styles, and cognitive abilities when making key decisions. This limitation is particularly relevant to entrepreneurship education, where individual differences in risk tolerance, creativity, and decision-making styles significantly impact learning outcomes [17]. In other words, the personalization and adaptation of games remains an unresolved problem.

A key advantage of serious games in educational contexts is their ability to track a wide range of performance metrics beyond traditional assessments. As Rahimi and Shute [18] observe, these games not only test current knowledge and practical skills but also assess preparation for future learning by tracking collaboration, innovation, design thinking, and entrepreneurial abilities. Recent advances in learning analytics have enhanced this capability, allowing for more sophisticated assessment of complex competencies [19].

B. Adaptive 3D Maze Games for Learning

Many of the latest approaches to learning through games rely on the use of serious maze games for educational purposes [20]. Mazes are widely used in entertainment games and can be represented by graphs, making them very suitable for interactive content presentation, where the player chooses one of several possible actions. Typically, a maze presents part of a course or module, where each position (i.e., a node in the maze's graph) contains information about a concept or situation. For a given position, there are several possible choices (actions) through which the player can continue in the maze. After selecting the desired action, the player moves to the next position (node) of the transition graph and explores the information in that node to decide how to proceed further. In this way, he/she can navigate the entire maze, following his/her strategy for choosing nodes in the transition graph and familiarizing himself/herself with the learning material presented in it. In this way, mazes enhance engagement and promote active learning are a useful tool for educational games.

The 3D virtual environments offer significant potential for engaging and immersive learning, and this extends to maze-based educational tools. While platforms like EcoMUVE [21], which allows students to explore complex ecological systems, and Second Life [22] have demonstrated the capacity of 3D environments for experiential learning, challenges remain in adapting these tools for specific pedagogical needs. Specifically, the authoring and modification of 3D learning content often demand considerable technical expertise from educators, hindering widespread adoption in mainstream education [23]. In contrast, simpler 2D maze creation tools, such as those created with Scratch [24] for basic navigation skills, provide greater accessibility but typically lack the immersive qualities and capacity for sophisticated, scenario-based decision-making that more complex 3D environments can offer [25].

The primary limitations of current maze-based learning tools include:

1. Limited educator control over content without programming knowledge
2. Insufficient adaptation to different learning styles and paces
3. Weak connection between maze structure and pedagogical objectives
4. Inadequate assessment mechanisms for complex competencies.

These limitations are particularly problematic for entrepreneurship education, where rapidly evolving business landscapes require frequent content updates, and diverse student backgrounds necessitate personalized learning paths [17].

C. Teaching technological entrepreneurship through serious games

The application of games in strategic education has a rich history, tracing back to 17th-century war games that simulated complex military decisions [26]. Contemporary computer-based simulations have become increasingly prevalent in entrepreneurship education, with numerous studies demonstrating their effectiveness in developing specific entrepreneurial competencies [27]. The competencies targeted with these games include opportunity recognition and evaluation, resource acquisition and allocation, risk assessment and management [13], business model development and pivoting, and market analysis and competitive positioning. These competencies are crucial parts of all stages of the strategic management process, and their proper implementation increases the chances of success of strategic management.

Neck and Greene [11] argue that entrepreneurship is best taught as a method rather than a process, requiring active practice of entrepreneurial thinking rather than passive learning of concepts. Serious games align perfectly with this pedagogical approach by allowing students to experiment with entrepreneurial decisions in low-risk environments [28].

Several initiatives have emerged to support the development and implementation of serious games in entrepreneurship education:

- The Serious Games Initiative [29] focuses on addressing management and leadership challenges through game-based learning
- The eSG Project [30] aims to stimulate innovation and entrepreneurial mindsets through tailored video games
- ENTRExplorer [31] emphasizes entrepreneurial motivation and risk-taking through gamified learning experiences.

A diverse array of techniques and methodologies for applying serious games in technology entrepreneurship exists [32], including the use of customizable video mazes to teach effectuation through specialized software platforms [33]. However, the potential of style-based adaptive video games remains largely unexplored [12]. Preliminary research among master's students in entrepreneurship indicates a positive reception to these games, particularly regarding their ability to capture student interest [12]. Further investigation, incorporating student interest indicators and adaptive game implementation, alongside detailed student feedback, is essential to comprehensively understand the motivational impact of style-based adaptive video games.

D. The Game Experience Questionnaire (GEQ)

The Game Experience Questionnaire (GEQ) [34] is a widely validated and standardized instrument specifically designed to assess the multifaceted subjective experience of players during gameplay. Beyond simply measuring enjoyment, the GEQ delves into key aspects of engagement that are particularly relevant to the effectiveness of serious games for learning. These aspects include:

- Immersion - the degree to which players feel absorbed in and connected to the game world, a crucial factor in maintaining attention and facilitating engagement with learning content.
- Flow - a state of optimal experience characterized by deep concentration, a sense of control, and intrinsic enjoyment, which can significantly enhance learning and motivation.
- Competence - the player's feeling of mastery and effectiveness within the game, which can positively impact self-efficacy and the willingness to tackle challenging learning tasks.
- Tension/Challenge - The level of pressure and difficulty experienced, which, when appropriately balanced, can drive engagement and critical thinking.
- Negative Affect - feelings such as frustration, boredom, or anxiety, which can hinder learning and should be minimized in effective educational games.
- Positive Affect - emotions like enjoyment, fun, and satisfaction, which can foster a more positive learning attitude and improve retention.

The GEQ provides valuable insights into how players interact with and perceive a game, highlighting elements that contribute to engagement and overall experience. Understanding these aspects is crucial in the context of serious games, as a positive and engaging experience can enhance learning outcomes. By quantifying these subjective dimensions, the GEQ allows researchers to gain a deeper understanding of how game design choices impact player engagement and overall experience, providing essential data for the design and evaluation of serious games aimed at achieving specific learning outcomes.

E. Synthesis and Research Gap

According to the literature review, each of the domains - 3D maze serious games, entrepreneurship education and motivational design - contributes to effective entrepreneurship education. 3D maze serious games provide opportunities for engagement and assessment and offer solution-based learning pathways. Entrepreneurship education defines critical competencies. Understanding the game experience, as measured by instruments like the GEQ, provides a framework for evaluating the effectiveness of these games in engaging learners.

However, significant gaps remain in the effective integration of these elements:

- Most existing entrepreneurship games lack sophisticated adaptation mechanisms to suit individual learning styles and entrepreneurial orientations
- Current 3D learning environments require technical expertise to be modified, which limits their usefulness for entrepreneurship educators
- Few platforms combine immersive 3D environments with teacher-friendly content management systems

The STRAMAG game addresses these gaps by providing an accessible system for creating adaptive 3D maze learning

environments specifically designed for entrepreneurship education. Unlike previous approaches that prioritize either technical complexity or ease of use, STRAMAG offers both and enables non-technical educators to create serious games. By focusing on creating an engaging game experience (assessed using the GEQ) and mapping maze structures to entrepreneurial decision-making processes, STRAMAG represents a new synthesis of the research areas discussed above.

III. THE STRAMAG GAME

A recent research of game-based learning indicated that most serious game studies focus on learning conceptually, with concepts such as general reasoning, creativity, system understanding, and decision making, which do not demand special knowledge of subject areas [2]. This study, however, investigates the use of style-based adaptive video games to enhance learning motivation within a strategic management module of an entrepreneurship program.

The game design and learning content were developed adhering to the Essential Design Criteria for Educational Games [32], which emphasize:

1. Winning should be based on knowledge or skills, not random factors.
2. The game should focus on core curriculum content, not trivial information.
3. The dynamics of the game should be easy to understand and interesting for the players but not obstruct or distort learning.
4. Students should not lose points for wrong answers.
5. Games should not be zero-sum exercises.

One of the biggest challenges in developing learning content for games is the risk that the content in these games is too general and inappropriate for fulfilling existing curricula [3], which leads to an inability to accurately evaluate the applicability of serious games in education. The learning content of this experiment was developed following the educational curricula of the “Strategic management” course within the master's degree program “Technology Entrepreneurship and Innovations in IT” at Sofia University “St. Kliment Ohridski” (Bulgaria). More precisely, the didactic content in the maze game is based on the first two modules of this course, namely “Introduction to Strategic Management” and “Strategic Management Analysis”.

A. Structuring of the learning content

The implementation of a style-based adaptive video game in the educational process of Strategic management required preparation of a specific learning content for the game. This content was designed to:

- Encompass all key syllabus elements;
- Offer well-structured and precise information for each of the elements in the syllabus;
- Offer textual and graphic learning content in format and size, which is supported by the game's design;
- Include learning content suitable for the learning styles of all four basic types - theorists, reflectors, pragmatists, and activists.

The learning content was structured into learning units [11] based on the strategic management process [31], covering the initial stages: Strategic Philosophical Values and Strategic Analysis. An overview of the overall strategic management process is also included at the beginning of the learning content, as it is included in the educational curricula.

The learning content in the game incorporated the following elements:

- Overview of the Strategic Management Process and Stages;
- Strategic Philosophical Values (company values, vision, mission, strategic goals);
- Strategic Analysis (SWOT Analysis, PEST Analysis, Porter’s Five Forces Analysis, VRIO Analysis, USP Analysis).

The learning units, included in the 3D serious educational game STRAMAG, cover the introduction to the Strategic Management Stages. Following the initial structural knowledge on the process stages, a thorough view of the strategic philosophical basics is introduced. These include Strategic Philosophical Values of the company, Vision, Mission, and Strategic goals. These competencies are of crucial importance for the stable strategic management of the company. After introducing the competencies for forming the foundation of strategic management of the company, an introduction to Strategic Analysis is performed. The Strategic Analysis tools included in the game start with SWOT Analysis. This basic micro-environment analysis on the internal strengths and weaknesses, and external opportunities and threats is focused on analyzing all specific factors for the company impacting its strategic choices. This is done through dividing the factors by two scale dimensions – source type (internal/external) and impact type (positive/negative). This tool is applicable in every occasion of changes in the micro-environment and is one of the most widely used strategic management tools. The second Strategic Analysis Tool implemented in the game is PEST Analysis. It is evaluating external political, economic, social, and technological macro-environment factors influencing the company, and it is of very high importance for strategic management when changing the external macro-environment, starting operations in a new market for the company, or facing large changes in the current market. [35] The third Strategic Analysis Tool included in the game is Porter’s Five Forces Analysis. Its goal is to analyze the competitive intensity and attractiveness of an industry, and it is achieving its goal through analyzing the forces that shape them – Competitive Rivals, Potential for New Entrants in an Industry, Supplier Power, Customer Power, and Threat of Substitutes. The fourth Strategic Analysis Tool, which is included in the game, is VRIO Analysis. It is focused on evaluating the availability, type and strength of the company’s competitive advantage. The fifth Strategic Analysis Tool included in the STRAMAG game is USP Analysis. The Unique

Selling Proposition Analysis is a detailed approach to identifying the specific competitive advantages in the company's offers. USP Analysis is a crucial step in defining a strategic advantage for the company's products on the market. It is supporting the development of products and services with higher value for the customers with great precision due to the evaluation of each feature and its level of importance for the customers. [36]

Each learning unit developed in STRAMAG game includes: (1) textual learning content, (2) graphical learning content, (3) additional information (video, textual, graphical, etc.), (4) additional learning tasks and homework. The learning tasks and homework were grouped based on their appropriateness for a specific learning style of Honey and Mumford [37]. Therefore, four groups of tasks were created, appropriate for the learning style of activists, reflectors, theorists, and pragmatists as follows:

- Activists' tasks – include implementation of different activities, aiming to increase the learning motivation of activist players
- Reflectors' tasks – include materials, tasks and puzzles, giving them opportunity for commenting, overviewing and reflecting on the educational materials and practices
- Theorists' tasks – include reading texts and further materials, creating structures and solving logically complex problems
- Pragmatists' tasks – include directions for direct application of the learned theoretic materials, best practices and examples.

B. Game design and structure

Effective implementation methodologies are crucial for the success of serious games in education [5, 6]. Recent applications have demonstrated the efficacy of maze structures for presenting learning material [7], which are also employed in the "STRAMAG" 3D educational game. Students navigate between learning units situated within "rooms" (nodes) by selecting from connecting "tunnels" (pathways).

Each learning unit is situated within a distinct "room." The "tunnels" connecting these "rooms" incorporate tasks and mini-games tailored to individual student learning styles. Figure 1 illustrates the organization of learning units and their placement within these "rooms," with numerical identifiers (0-9) indicated in red circles.

Both "rooms" and "tunnels" contain learning materials and tasks designed for different learning styles. The game's adaptive algorithm ensures students are presented only with content relevant to their specific style.

The primary objective is to address the unique learning needs of each student and enhance their motivation through the use of a style-based adaptive video game. To illustrate the specific learning content delivered through this maze structure, the following section details the content within each of the ten rooms (or halls) of the maze representing the "STRAMAG" game.

C. Description of the maze rooms with learning content

The maze contains ten rooms (interconnected via tunnels), as follows below.

Room 1: "Strategic Management - Basic stages" – includes information about the basic stages in the strategic management process: Philosophical values - Strategic Analysis - Strategy Formulation - Strategy Implementation - Strategy Execution - Strategy Control and Evaluation. It includes textual material, graphics and diagrams clarifying the concepts, as well as links for further reading.

Room 2: "Strategic Management – Vision" – provides the textual learning content connected with company's vision - definition, basic elements, major directions when formulating the vision and most important questions, which the vision should answer. The learning content also includes practical examples of visions from real companies and links for further reading on the topic.

Room 3: "Strategic Management – Mission" – contains textual content on company mission: definition, differentiation from vision, key questions, examples, video materials, and links for further reading.

Room 4: "Strategic Management - Strategic goals" – focuses on strategic goals and requirements, explaining SMART goals, key questions, graphics, and links for further reading.

Room 5: "Strategic Management - Strategic analysis introduction" – covers the importance and place of strategic analysis in the strategic management process, basic analysis directions, the distinction between strategic analysis in the internal and external environment, basic tools of strategic analysis, examples, and links for further reading.

Room 6: "Strategic Management -Strategic analysis - SWOT Analysis" – focuses on SWOT analysis. It includes textual and graphical learning content describing the tool, all of its elements, the basic questions, and the direction for analysis in each of its sectors, as well as providing a thorough step-by-step guide for implementing the analysis, suitable also for pragmatists and activist learners.

Room 7: "Strategic Management – Strategic analysis - PEST Analysis" – dedicated to PEST analysis. The learning content includes textual and graphical information about the analysis, its application, elements, basic questions, and directions in the implementation of each element of the analysis, as well as an explanation of the other variations of the PEST analysis - PESTLE/PESTEL/PESTLED/STEEPLE/etc.

Room 8: "Strategic Management - Strategic analysis - P5F Analysis" – covers Porter's Five Forces Analysis, its definition, application, content, directions in implementation, and application. It describes all the basic powers in the analysis, their specifics, and influence on the others, as well as information on the key practical points in applying the P5F Analysis in the Strategic Management Process.

Room 9: "Strategic Management - Strategic analysis - VRIO Analysis" – covers VRIO analysis, its definition, textual and graphical presentation, its importance, and a graphical presentation of the methodology of evaluating the type

and durability of the competitive advantage of a company by providing a step-by-step evaluation model for practical application of the analysis.

Room 10: “Strategic Management - Strategic analysis - USP Analysis” – focuses on USP analysis and its importance in the strategic management process for entrepreneurial companies. The room contains the definition of the analysis, its application, and a practical graphical and textual presentation of an example in the application of the analysis.

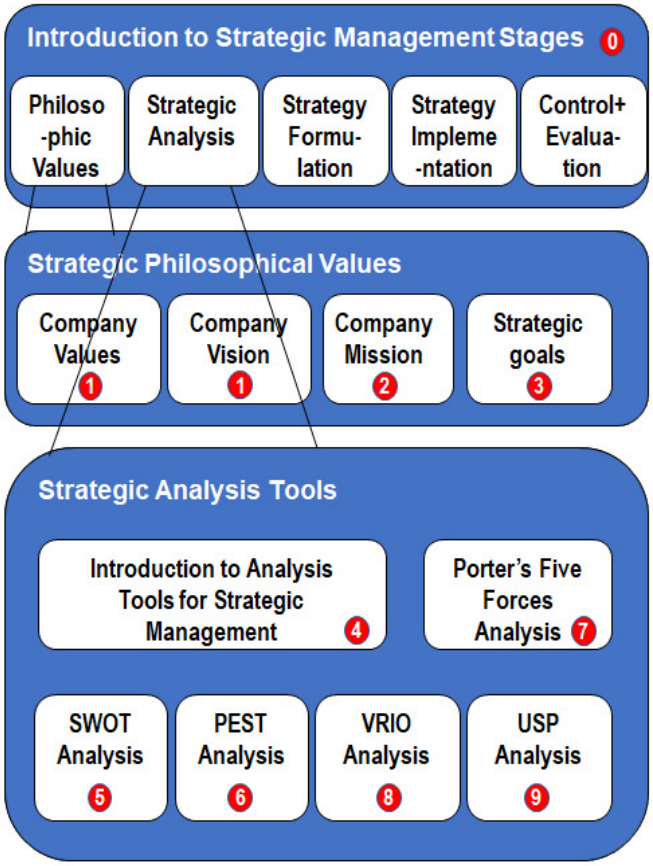


Fig 1. Content distribution across learning units and game rooms

Each room is accessible by "tunnels" and contains tasks and puzzles tailored to individual learning styles, determined during initial gameplay. The "tunnels" feature tasks and activities designed for activist, reflector, theorist, and pragmatist learning styles.

The learning content in every room and the tunnels includes materials and tasks suitable for all learning styles, but all players are shown only the elements suitable for them, according to the learning-style adaptation of the game. The goal is for the learning content to respond to the learning needs of every student in the game, about his/her learning style, and to foster learning motivation during the educational process by using a style-based adaptive video game.

Other mini-games include arranging contextually related images and assembling 3D geometric bodies, which develop associative thinking and spatial imagination. Teachers can customize both the quiz questions and the content of the other mini-games.



Fig. 2. Start screen of the STRAMAG game

The educational content presented in the maze rooms (Fig. 2) encompasses the theory of effectuation, developed by Dr. Saras Sarasvathy [38]. This theory, which has become popular in the last decade, focuses on strategies for achieving entrepreneurial goals under conditions of uncertainty [39].

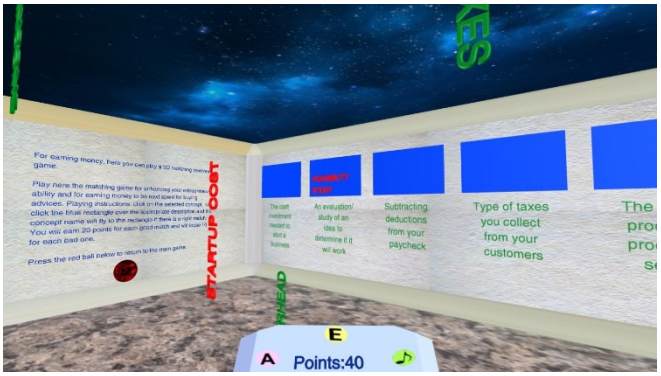


Fig. 3. Inside one of the matching games in the quiz room

Each tunnel passage requires students to use a certain number of points, which adds an element of strategy and planning. To increase their points, students must complete additional tasks integrated into the game.



Fig. 4. Screen of the 3D Zoom game

To prevent monotony and increase the effectiveness of learning, a variety of mini-games are included in the maze. These mini-games are designed to develop fine motor skills, visual-spatial thinking, and context-based decision-making abilities, encouraging entrepreneurial activity and creativity. The mini-games range from entertaining to those with a clear didactic purpose, such as 3D quizzes and games with goal-oriented tasks. For example, a dynamic 3D quiz (Fig. 3) requires students to select from moving text objects containing answers to questions. The speed of movement and rotation of the objects can be adjusted to vary the difficulty. Correct answers are awarded points, allowing the quiz to be used as an assessment tool.

IV. Experimental Setup

The 3D Maze game for entrepreneurship education described above was used in a practical experiment together with two included mini-games – a quiz (3D Quiz – Fig. 3) and an arrangement of scattered context-related images (3D Zoom – Fig. 4). The selection of the educational content and mini-games was coordinated with leading experts in the area of entrepreneurial education, who possess many years of experience in entrepreneurship education. The essence of the effectuation theory, its principles, their comparison with those of causal relationships, as well as the process and entrepreneurial method identified by Sarasvathy [39], were described on boards and interactive whiteboards in six rooms of the maze. The test questions in the 3D Quiz referred to both the educational content presented in the halls and to basic concepts in entrepreneurship, such as what a feasibility study is, net profit, initial and overhead costs, etc.

23 students from the master's program in Technological Entrepreneurship at the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski" participated in the experiment. All participants had excellent computer skills and basic knowledge of entrepreneurship.

The practical experiment consisted of a presentation and demonstration of the 3D maze video game "STRAMAG", followed by individual game sessions, each lasting approximately 10 minutes. After the sessions, participants completed the GEQ [34], which assesses the level of engagement and experience during gameplay.

V. Results

A. Player profiles

23 students enrolled in the master's program in "Technology Entrepreneurship and Innovations in IT" at the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski" participated in this study. The age of the participants ranged from 20 to 35 years. The mean age was approximately 26.3 years, with a standard deviation of approximately 3.8 years. The gender distribution of the participants was 15 males and 8 females. Before participation, all students self-assessed their computer skills on a 5-point scale (1 = lowest, 5 = highest). The distribution of self-assessments was as follows: 0 participants rated their skills as a 1, 2 as a 2, 2 as a 3, 15 as a 4, and 4 as a 5. This distribution indicates a generally high level of self-perceived computer proficiency within the participant group.

B. GEQ results

Figure 5 and Table I illustrate the responses to the GEQ questions. Overall, students reported a positive experience with the STRAMAG game. High mean scores were observed across several dimensions: Immersion, Positive Affect, Competence, and Flow. These results suggest that students found the game engaging, enjoyable, and provided a strong sense of mastery.

TABLE I.
MEAN VALUES OF GEQ DIMENSIONS AND ASSOCIATED QUESTIONS FOR DIFFERENT MINI-GAMES (3D MAZE, 3D QUIZ, 3D ZOOM)

Dimension	Questions	3D Maze	3D Quiz	3D Zoom
Competence	2, 10, 15, 17, 21	3.91	3.74	4.09
Immersion (sensory & imaginative)	3, 12, 18, 19, 27, 30	4.35	4.13	4.17
Flow	5, 13, 25, 28, 31	3.70	3.96	3.87
Tension/ Annoyance	22, 24, 29	2.32	2.04	2.17
Challenge	11, 23, 26, 32, 33	2.09	2.17	2.26
Negative Affect	7, 8, 9, 16	2.14	2.06	2.36
Positive Affect	1, 4, 6, 14, 20	3.96	3.74	4.09

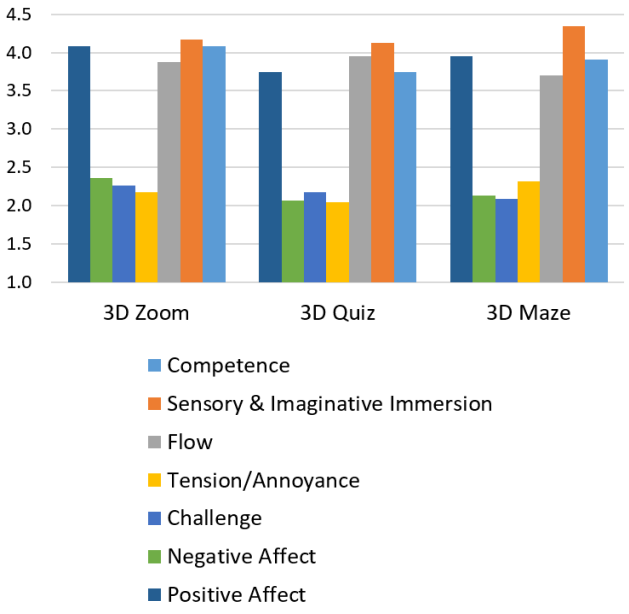


Fig. 5. Responses to the questions from the GEQ

The Negative Affect dimension showed consistently low scores across all game versions (3D Maze: 2.14, 3D Quiz: 2.06, 3D Zoom: 2.36), indicating minimal levels of frustration, boredom, or anxiety. Similarly, the Tension/Annoyance dimension yielded low scores (3D Maze: 2.32, 3D Quiz: 2.04, 3D Zoom: 2.17), suggesting that students experienced little discomfort. The Challenge dimension also showed moderate scores (3D Maze: 2.09, 3D Quiz: 2.17, 3D Zoom: 2.26), reflecting an optimal level of difficulty for maintaining engagement.

VI. Discussion

The findings of this study provide strong evidence for the usability and effectiveness of the STRAMAG game as an educational tool for entrepreneurship and strategic management. The game's 3D environment, combined with embedded mini-games, offers a dynamic and interactive learning experience that traditional methods often lack. Moreover, the novelty of the STRAMAG game consists in introducing an adaptation of learning content based on the learning style of the individual player. The results presented in Fig. 5 show that different video mini-games that adapt educational content to learning styles affect student engagement in a slightly different way.

The positive responses observed in the GEQ indicate that the game effectively engaged the students. The high scores for Immersion, positive effect, Flow, and Competence can be attributed to the design of the 3D maze environment and the integration of relevant entrepreneurial concepts within the maze structure and the embedded mini-games.

The Negative Affect dimension showed consistently low scores across all mini games (3D Maze: 2.14, 3D Quiz: 2.06, 3D Zoom: 2.36), indicating a minimal level of frustration, boredom, or anxiety. Similarly, the Tension/Annoyance dimension yielded low scores (3D Maze: 2.32, 3D Quiz: 2.04, 3D Zoom: 2.17), suggesting that students experienced little discomfort. The Challenge dimension also showed moderate scores (3D Maze: 2.09, 3D Quiz: 2.17, 3D Zoom: 2.26), reflecting an optimal level of difficulty for maintaining engagement.

It is important to acknowledge the limitations of this study. The sample size of 23 master's students, while sufficient for preliminary analysis, may limit the generalizability of the findings. Future research should aim to include larger and more diverse student populations. Additionally, longitudinal studies could investigate the long-term impact of the STRAMAG game on students' entrepreneurial skills and career trajectories. Further refinement of the game's adaptive mechanisms and exploration of additional game-based learning strategies are also warranted.

VII. Conclusions

The use of educational video games in higher education has the potential to significantly enhance the outcomes of traditional e-learning by making it more engaging and effective. Video games offer innovative ways to acquire knowledge and practical skills through informal and often implicit problem-solving.

This research was motivated by the existing lack of customizable and free video games for educating young people in entrepreneurship and strategic management [40]. The developed software system for declarative description, personalization, and generation of 3D video mazes, or such that offer a free solution for mass game-based learning, not only in entrepreneurship. It allows teachers to construct mazes presenting content from a given subject area in a personalized and interactive way, where the learner chooses how to continue navigating the maze. In addition to becoming acquainted with new ideas, concepts, and theories in a 3D maze, learners can play entertaining embedded mini-games with didactic value that develop motor skills, visual and spatial thinking, and context-based reasoning. The positive results from the Game Experience Questionnaire indicate that students found the STRAMAG game to be highly immersive and perceived the learning content as useful, suggesting the potential of this game for engaging and effective entrepreneurship education.

Future research and development should focus on several key areas. Firstly, expanding the learning content within STRAMAG to encompass a broader range of entrepreneurship topics would increase its applicability. Secondly, implementing more nuanced adaptive learning mechanisms that dynamically adjust difficulty and content based on individual student progress and learning styles could further enhance its effectiveness and foster a greater sense of competence. To further capture and maintain student attention, future iterations could explore incorporating more dynamic visual elements and unexpected challenges. Additionally, enhancing the perceived relevance of the content through personalized pathways and real-world scenarios, and integrating a more robust reward system to increase learner satisfaction, are important avenues for future development. The authors plan to compare the knowledge acquisition of students who played the game with those who did not by conducting a controlled experiment with a larger sample size, dividing participants into experimental and control groups. Finally, exploring the integration of collaborative gameplay features could foster social learning and engagement among students.

Acknowledgment

The research leading to these results has received funding from the MOTIVATE project, funded by the Bulgarian National Science Fund, Grant Agreement No. КП-06-Н85/2 dated 05/12/2024, and from the EU ESI Funds by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No. BG-RRP-2.004-0008.

References

[1] T. H. Apperley, “Genre and game studies: Toward a critical approach to video game genres,” *Simulation & Gaming*, 2006, 37(1), pp. 6-23.

[2] T. Paris, “The structure and dynamics of the video games industry,” in Legal challenges and market dynamics in the video games sector, 2024, IRIS, European Audiovisual Observatory, pp. 9-19.

[3] R. Zhang, "Research on the Progress of VR in Game," Highlights in Science, Engineering and Technology, 39, 2023, pp. 103-110.

[4] M. F. Dallaqua, B. Nunes and M. M. Carvalho, "Serious games research streams for social change: Critical review and framing," British Journal of Educational Technology, 55(2), 2024, pp. 460-483.

[5] M. Cremaschi, G. R. Avis, A. Q. Zhao, E. Guarnieri, A. Panzeri, and A. Spoto, "PENguIN: A mental health application employing gamification and token economy to boost therapeutic adherence in young users," Computers in Human Behavior Reports, vol. 17, 100586, 2025.

[6] S. Saeedi, M. Ghazisaeedi and S. Rezayi, "Applying Game-Based Approaches for Physical Rehabilitation of Poststroke Patients: A Systematic Review," J Healthc Eng., 2021, Art. ID 9928509.

[7] Piech, Joanna, and Krzysztof Czernicki. 2021. "Virtual Reality Rehabilitation and Exergames—Physical and Psychological Impact on Fall Prevention among the Elderly—A Literature Review" Applied Sciences 11 (9), 4098. https://doi.org/10.3390/app11094098..

[8] Š. Krištofík, M. Kasáš, Š. Neupauer and P. Malík, "StarCraft agent strategic training on a large human versus human game replay dataset," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, vol. 21, *Annals of Computer Science and Information Systems*, IEEE, 2020, pp. 391-399, doi: 10.15439/2020KM178

[9] N. Stein and T. Tamir, "Stackelberg Strategies for Weighted Load Balancing Games," in Proc. of the 17th Conference on Computer Science and Intelligence Systems, 2022, pp. 373–382.

[10] M. Macák, S. Bojnak and B. Buhnova, "Identification of Unintentional Perpetrator Attack Vectors using Simulation Game: A Case Study," in Proc. of the 16th Conference on Computer Science and Intelligence Systems, 2021, pp. 349–356.

[11] H. M. Neck and P. G. Greene, "Entrepreneurship Education: Known Worlds and New Frontiers," *Journal of Small Business Management*, 2011, pp. 55-70.

[12] T. Ahmadov, A. Karimov, S. Durst, M. Saarela, W. Gerstlberger, M. F. Wahl and T. Karkkainen, "A two-phase systematic literature review on the use of serious games for sustainable environmental education," Interactive Learning Environments, 33(3), 2024..

[13] N. M. Nasir, H. Ithnin, Z. A. Nasir, A. Zia and A. R. Ridzuan, "Effectiveness of Digital Game Based Learning (DGBL) Tools Usage in Macroeconomics among Tertiary Education Students: ARCS Motivation Model," International Journal of Research and Innovation in Social Science, Vol. VIII, 2024, pp. 5219-5230.

[14] A. Daniel, Y. Negre, J. Casaca, R. Patrício and R. Tsvetcoff, "The effect of game-based learning on the development of entrepreneurial competence among higher education students", ICERI2024 Proceedings, 2024, pp. 3294-3300.

[15] B. P. Bontchev, V. Terzieva and E. Paunova-Hubenova, "Personalization of serious games for learning," Interactive Technology and Smart Education, 18(1), 2021, pp. 50-68.

[16] R. E. Mayer, L. Fiorella and A. Stull, "Five ways to increase the effectiveness of instructional video," *Educational Technology Research and Development*, 68(2), 2020, pp. 837–852.

[17] H. M. Neck, C. G. Brush and P. G. Greene, (Eds.) *Teaching entrepreneurship, volume two: A practice-based approach*. Edward Elgar Publishing, 2021.

[18] S. Rahimi and V. J. Shute, *The effects of video games on creativity: A systematic review*. Handbook of lifespan development of creativity, 37, 2021.

[19] D. Ifenthaler and J. Y. K. Yau, "Utilising learning analytics to support study success in higher education: a systematic review," *Educational Technology Research and Development*, 68(4), 2020, pp. 1961-1990.

[20] E. Piette, C. J. F. S. Schiffel and C. J. T. W. Browne, "The Ludii Game Description Language is Universal," arXiv preprint arXiv:2205.00451, 2022.

[21] S. J. Metcalf, A. M. Kamarainen, E. Torres, T. A. Grotzer and C. Dede, "EcoMUVE: A Case Study on the Affordances of MUVEs in Ecosystem Science Education," in Y. Qian (Ed.), *Integrating Multi-User Virtual Environments in Modern Classrooms*, 2018, IGI Global Scientific Publishing, pp. 1-25. https://doi.org/10.4018/978-1-5225-3719-9.ch001

[22] S. A. Chelloug, H. Ashfaq, S. A. Alsuhibany, M. Shorfuzzaman, A. Alsufyani and A. Jalal, "Real Objects Understanding Using 3D Haptic Virtual Reality for E-Learning Education," *Computers, Materials & Continua*, 75(1), 2023.

[23] D. X. H. Mai, G. Loup and J.-Y. Didier, "Usability of An Immersive Authoring Tool: An Experimental Study for the Scenarization of Interactive Panoramic Videos," in *Proc. of 16th International Conference on Advances in Computer-Human Interactions (ACHI 2023),* Venice, Italy, 2023, April, pp. 174 - 180.

[24] W. Stewart and K. Baek, "Analyzing computational thinking studies in Scratch programming: A review of elementary education literature," *International Journal of Computer Science Education in Schools*, 6(1), 2023, pp. 35-58.

[25] Al Amri, A. Y., Osman, M. E. and Al Musawi, A. S. "The Effectiveness of a 3D-Virtual Reality Learning Environment (3D-VRLE) on the Omani Eighth Grade Students' Achievement and Motivation towards Physics Learning," *International Journal of Emerging Technologies in Learning (iJET)*, 15(05), 2020, pp. 4–16, https://doi.org/10.3991/ijet.v15i05.11890

[26] T. Beranič and M. Heričko, "The Impact of Serious Games in Economic and Business Education: A Case of ERP Business Simulation," Sustainability, 14(2), 2022, Art. ID 683.

[27] I. Martins, J. P. P. Perez, D. Osorio and J. Mesa, "Serious Games in En trepreneurship Education: A Learner Satisfaction and Theory of Planned Behaviour Approaches," The Journal of Entrepreneurship, 32(1), 2023, pp. 157-181.

[28] R. Sørheim, T. Aadland and D. Håkon, "19. Venture creation programs: what kinds of ventures do students create?" *Innovation in global entrepreneurship education: Teaching entrepreneurship in practice*, 274, 2021.

[29] Wilson Center, *Serious Games Initiative*. 2025, Retrieved from https://www.wilsoncenter.org/program/serious-games-initiative

[30] F. Bellotti, R. Berta, A. De Gloria, E. Lavagnino, A. Antonaci, F. Dagnino, M. Ott, M. Romero, M. Usart and I. S. Mayer, "Serious games and the development of an entrepreneurial mindset in higher education engineering students," Entertainment Computing, Volume 5, Issue 4, 2014, pp. 357-366, ISSN 1875-9521, https://doi.org/10.1016/j.entcom.2014.07.003.

[31] N. P. Bastos, B. M. Sánchez-Cantón and I. H. Costas, "ENTRExplorer: A Serious Game for Immersive Entrepreneurs," presented at the 1.º EJML - Encontro sobre Jogos e Mobile Learning, Coimbra (Faculdade de Psicologia e de Ciências da Educação da Universidade de Coimbra), 2012, October, https://doi.org/10.13140/2.1.4219.8724

[32] A. Mitchell and C. Savill-Smith, "The use of computer and video games for learning: A review of the literature," 2004, http://www.lsda.org.uk/files/pdf/1529.pdf

[33] E. Paunova-Hubenova, E. "Are the school teachers ready to start using smart adaptive video games for education?", in *Proc. of the 13th International Technology, Education and Development Conference (INTED2019)*, Valencia, Spain, ISBN: 978-84-09-08619-1, ISSN: 2340-1079, 2019, March, doi:10.21125/inted.2019.1294, pp. 5191-5199.

[34] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart and J. N. Pidruzny, "The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing," *Journal of Experimental Social Psychology*, 45(4), 2009, pp. 624-634.

[35] A. B. Zanoni, *Strategic Analysis: Processes and Tools,* Routledge, 2012, ISBN 9781138203099.

[36] E. Elgersma, *The Strategic Analysis Cycle Hand Book: How Advanced Data Collection and Analysis Underpins Winning Strategies*, LID Publishing, 2017, ISBN-10: 1911498363, ISBN-13: 978-1911498360.

[37] P. Honey and A. Mumford, *The learning styles helper's guide*, 2000.

[38] S. D. Sarasvathy and S. Venkataraman, *Acting on, reacting to, and interacting with markets as artifacts*, The Oxford Handbook of Organizational Change and Innovation, 131, 2021.

[39] S. D. Sarasvathy and M. A. Perez, "Stopping the Slide and Rising Above the Tide: Entrepreneurial Education to Move Out of Necessity into Opportunity," in *Necessity Entrepreneurship: Getting Beyond the Binary,* Emerald Publishing Limited, 2025, pp. 39-64.

[40] HBR, *HBR Guide to Managing Strategic Initiatives*, Harvard Business Review, Wiley, 2020.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 499–506
DOI: 10.15439/2025F3908
ISSN 2300-5963 ACSIS, Vol. 43

ELEVATE-AI: Evaluation of Learning Environments Via Assessment Tools Enhanced by AI

Linda Burchiellaro
0009-0008-1098-1398
University of Modena and Reggio Emilia
Department of Physics, Informatics and Mathematics
Via Campi 213, 41125 Modena, Italy
Email: 273342@studenti.unimore.it

Francesco Faenza, Claudia Canali
0000-0002-7258-7192, 0000-0001-8448-7693
University of Modena and Reggio Emilia,
Department of Engineering "Enzo Ferrari",
Via P. Vivarelli 10, 41125 Modena, Italy
Email: {francesco.faenza, claudia.canali}@unimore.it

***Abstract*—The number of STEM camps and extracurricular initiatives has risen considerably in recent years, driven by the increasing emphasis on the workforce shortage in STEM fields. However, despite their growth, these programs often suffer from a lack of structured evaluation practices, a well-known issue that hinders a comprehensive understanding of their effectiveness. This work focuses specifically on outreach initiatives and proposes ELEVATE-AI, a standardized evaluation platform that includes data-cleaning procedures, Exploratory Factor Analysis (EFA), and regression analysis to measure impacts effectively. Furthermore, we discuss the potential integration of AI-based tools to support non-experts in interpreting the results of the proposed analysis flow. The platform aims to lower technical barriers, promote systematic assessment, and encourage the widespread adoption of data-driven practices in evaluating CS and STEM outreach activities. To facilitate adoption and reproducibility, the platform will be made available as an open-source tool.**

I. Introduction

IN RECENT years, particularly since the mid-2010s, there has been growing attention at all levels of society toward increasing the number of graduates in STEM fields, particularly in computer science (CS). Public institutions such as the European Union have highlighted the urgent need for more professionals in digital and technological sectors, citing projected shortages of ICT specialists and the importance of closing the gender gap in these domains [1]. Similarly, the OECD and national education authorities have acknowledged the underrepresentation of women in STEM careers and the need for inclusive educational strategies [2].

Among the most prominent approaches to address this challenge are extracurricular and outreach initiatives, activities not formally associated with standard curricula but aimed at introducing students to STEM topics through informal, hands-on learning. Particularly in the case of girls, experiential and mentorship-based learning environments have shown effectiveness in fostering self-efficacy and interest in STEM subjects [3]. Broader studies confirm that these benefits extend to students of all genders when appropriate pedagogical strategies are used [4].

Despite their increasing prevalence, many outreach initiatives lack systematic evaluation, limiting their scalability and long-term impact [4], [5]. Programs often shift focus to trending topics, such as artificial intelligence, without assessing the impact of each iteration, resulting in design decisions based more on intuition than evidence [6]. One of the reasons for this absence of evaluation is the informal nature of these activities: unlike curricular settings, where learning outcomes can be tested directly, outreach initiatives tend to rely on self-perception measures rooted in Bandura's theory of self-efficacy [7].

Nevertheless, recent research demonstrates that robust evaluation is feasible even in informal contexts [5]. In this paper, we propose an open-source, self-hostable platform that provides a standardized, literature-informed workflow for evaluating extracurricular STEM initiatives. The platform, intended to be accessible to both researchers and educators, automates the main key phases of the evaluation process: survey design and deployment, data collection and validation, Exploratory Factor Analysis (EFA), and regression-based impact evaluation. The platform is designed to promote broader adoption of data-driven evaluation practices and to support community-driven contributions to its development. A valuable additional contribution of this paper is that the platform will be publicly released as an open-source resource, to maximize its impact and facilitate rapid adoption by researchers and practitioners.

The remaining part of the paper is structured as follows. Sec. II discusses the motivational background of the proposal. Sec. III describes the methodology including the three main phases of survey design, data collection and validation, and data analysis and implementation. Sec. IV presents the platform architecture and Sec. V illustrates the user interaction workflow. Finally, Sec. VI provides some concluding remarks and depicts the future research directions.

II. Motivation and Background

The evaluation of STEM and CS outreach activities, particularly in extracurricular settings, presents methodological challenges not typically encountered in formal education. In such informal contexts, learners often engage voluntarily, and activities are less structured, with outcomes being more affective or motivational than knowledge-based.

In contrast, the curricular context benefits from the increasing automation of evaluation processes, which often extends to automatic analysis and feedback generation. Several examples

illustrate this trend, particularly in STEM education. Cipriano et al. [8] present Drop Project, a platform designed to automate test result analysis and feedback for programming assignments. Similarly, Web-CAT[9] offers an automated grading system that compiles, tests, and evaluates student code submissions, providing immediate feedback. Majerník [10] introduces an e-Assessment Management System aimed at comprehensively evaluating medical students' knowledge, while, more generally, Stanescu et al. [11] propose a solution for automatic assessment of narrative answers, further demonstrating how structured digital platforms can support rigorous assessment practices even in highly specialized domains. These platforms are just some examples of how it is not only feasible but also widely applied to automate evaluation systems when grading and formal tests are involved.

Another contribution to formal testing automation in STEM is *LASSO* [12], a tool designed to support instructors in the assessments of their courses. LASSO's primary purpose is to promote the implementation of research-based teaching practices; in fact, it guides the teacher through a rigorous process of pre-test and post-test analysis. The platform consists of a generalized version of the previously cited Web-CAT and Drop Project, providing a platform to facilitate rigorous formal evaluation in all STEM fields.

More broadly, in formal education, structured assessments such as knowledge tests allow for systematic and replicable evaluation. Notably, Hattie's Visible Learning synthesis [13] aggregated thousands of studies to quantify the effect size of nearly every measurable educational intervention. This kind of quantification is feasible mainly because of the controlled nature of formal instruction and the availability of consistent outcome measures.

Nevertheless, the principle behind Visible Learning, making instructional choices based on solid evidence rather than trends or intuition, is just as applicable in informal learning. Even in the absence of formal testing, it is possible to build robust evaluation models using survey-based self-perception data as long as those instruments are rigorously validated and analyzed. By supporting outreach practitioners in collecting and interpreting valid data, we argue that evidence-informed design is not only possible but necessary in extracurricular education.

While standardized questionnaires and evaluation frameworks developed for formal environments cannot be directly transplanted into outreach contexts without adaptation, several initiatives have attempted to improve evaluation in informal contexts, such as the CISE REU toolkit [5] and the survey repository curated by Decker and McGill [14]. Although these resources provide a helpful starting point, they require contextual refinement to be used effectively in pre-college or extracurricular settings.

Knekta et al. [15] indeed underline the issue of contextualization by arguing that validity is not a property of an instrument itself but of the inferences drawn from its use in a specific context. Their study, titled "One Size Doesn't Fit All," emphasizes the importance of context-specific validation using techniques such as Exploratory Factor Analysis (EFA). In their view, even well-designed instruments must be empirically validated with data from the actual population and learning setting in which they are used. This is particularly relevant for outreach initiatives targeting diverse, often underrepresented groups in non-standard learning environments.

In line with the need for contextualization, the *Towards s'more Connected Coding Camps* study [16] outlines the vision of the European OSCAR initiative, which seeks to integrate informal STEM learning experiences, such as coding camps, into students' broader educational and professional trajectories. The initiative focuses on the current fragmentation of the panorama of extracurricular activities. The proposal is to give continuity and align these activities with formal education, developing a coherent learning pathway in which informal learning is no longer treated as an isolated experience but contributes meaningfully to long-term educational outcomes.

TheFragebogen [17], a web-based, open-source questionnaire framework initially developed for Quality of Experience (QoE) research, provides a compelling model for digital instrument delivery. Its architecture is guided by three design principles: responsiveness to research needs (e.g., multimedia support, scalability), extensibility and flexibility (including serverless deployment), and a focus on robustness and reproducibility (e.g., long-term data archiving). While it originates outside the domain of education, TheFragebogen exemplifies how modular, browser-based systems can meet the needs of research-grade data collection. These principles are particularly relevant to the design of evaluation tools for outreach education, where technological robustness must be paired with usability and adaptability.

The evaluation platform proposed in this paper results from several years of iterative refinement grounded in field experience. Specifically, its development began in 2016 within the context of *Digital Girls* [18], [19], a long-running Italian initiative launched in 2013 to engage female students aged 16–18 and introduce them to computer science (CS). The project is recognized in the Observatory for Public Sector Innovation Case Study Library [1] and was cited in the European Commission's 2021 *She Figures* report[2].

Digital Girls evolved from a single-university summer camp into a regionally coordinated program supported by local universities, schools, and stakeholders. It combines site visits, guest lectures, and project-based learning in a female-only setting, addressing confidence gaps and gender stereotypes before university entry. The program spans two weeks and offers approximately 50 hours of immersive activities in university settings.

The iterative enhancement of this program and its evaluation methodology followed a design-based research model [20]. The research adopted an iterative and literature-informed de-

[1] https://oecd-opsi.org/innovations/digital-girls-emilia-romagna/

[2] https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/she-figures-2021_en

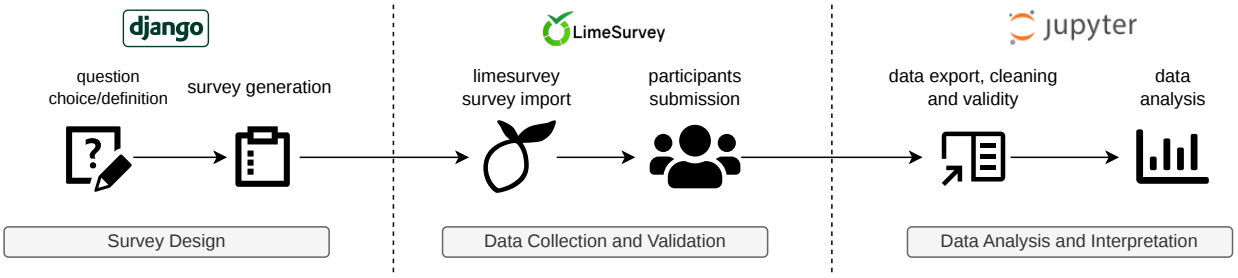


Fig. 1. Overall Platform Flow and Underlying Technologies

sign approach [21] involving cycles of instrument refinement, deployment, and validation. Survey instruments were informed by existing literature references but were adapted through feedback gathered via focus groups with past participants [22].

The resulting methodology includes two core survey phases, pre-camp, and post-camp, designed to capture changes in participants' perceptions, intentions, and confidence regarding CS. Several questions are repeated across both surveys to enable a longitudinal comparison of key indicators. Additional dimensions include background data (school type, video game experience, coding exposure), social influence (parental background), and affective constructs (motivation, stereotype perception), which align with Bandura's theory of self-efficacy [7] and Ajzen's theory of planned behavior [23].

Given the informal learning setting commonly associated with extracurricular initiatives, the evaluation strategy commonly deliberately excludes formal knowledge testing. Instead, it relies on participants' self-assessment to understand whether and how their attitudes have shifted as a result of the experience. A typical example is the measurement of self-efficacy in relation to the main topic of the camp, such as programming or algorithmic thinking. These constructs are typically evaluated through Likert-scale items, which provide a straightforward basis for statistical analysis. In some cases, free-text questions are added as they can provide meaningful additions or subtractions to the Likert items [24].

One of the main goals of STEM extracurricular initiatives is to engage participants and encourage them to consider future academic or professional paths in STEM fields. These programs are designed with the hope that, after the experience, participants will be more inclined to include STEM-related options among their future choices. In this context, Ajzen's theory of planned behavior [23] is particularly relevant, as it provides a framework for analyzing changes in intention. For instance, one of the main objectives of the *Digital Girls* initiative is to investigate whether the camp experience influences participants to reconsider their university plans, specifically, whether they become more likely to pursue a degree in computer science. To this end, the pre- and post-camp surveys are structured to detect shifts in intention, which can be analyzed as dependent variables in regression models to assess the impact of specific features of the camp.

III. Methodology

The multi-year experience derived from the organization of Digital Girls and other initiatives related to STEM disciplines led to the formalization of the platform presented in this paper. The platform is open-source, self-hostable, and modular, being written in Django, a Python framework for web platforms. It is structured, as shown in Figure 1, to assist both researchers and educators in evaluating CS outreach activities through three main phases: Survey Design, Data Collection and Validation, and Data Analysis and Interpretation.

A. Survey Design

One of the most critical components of any evaluation system is the design of the survey instrument, which essentially involves selecting appropriate questions for the context being evaluated.

Designing a proper survey is a complex activity, as it involves several steps that may be difficult to carry out without specific expertise. Some of the initial phases in survey development, such as literature reviews, interviews, or focus groups, require time and methodological competence [25].

The proposed platform supports this process by offering a curated set of questions, collected over time from relevant literature and organized according to the underlying constructs they aim to measure. Nonetheless, users can add new questions, provided with literature references, allowing more experienced users to expand and tailor the instrument to their specific needs.

Questions are grouped by thematic categories, more precisely referred to as *constructs*, defined in educational research as latent variables that represent abstract concepts measured through multiple related items [25]. Examples of such constructs include self-efficacy, motivation, or future academic intentions. Each construct can be measured at different levels of detail. For instance, satisfaction may be assessed through a single global item or through more granular questions addressing specific aspects such as satisfaction with the activity, teamwork, or instructor interaction. Similarly, future career intentions in STEM fields may be captured through a general yes/no item or a more detailed question listing potential academic or career paths.

While the platform provides a validated and structured base, it allows users to control the level of depth and complexity of their evaluation. Users can select only a minimal set of items for a lightweight assessment or opt for a more in-depth investigation depending on their goals and available resources.

This modular structure also supports the automatic generation of tailored analysis notebooks. When a construct is assessed through multiple items, the analysis notebook includes pre-configured aggregation options, such as mean scores, sum scores, or principal component analysis, based on the selected configuration.

B. Data Collection and Validation

The data collection phase occurs outside the platform itself. The platform provides an export functionality that generates a survey in a format compatible with LimeSurvey, an open-source and self-hostable tool for questionnaire administration. LimeSurvey enables the collection of responses in a way that allows pre- and post-survey data to be linked while still preserving respondent anonymity through the use of pseudonymized identifiers.

A crucial aspect of this phase, as emphasized by Knekta et al. [15], is the importance of validating the survey instrument in the specific context in which it is used. Even well-designed instruments must be empirically tested to ensure they measure the intended constructs within the target population. The platform facilitates this validation process by providing pre-configured scripts for performing Exploratory Factor Analysis (EFA) and reliability checks, such as Cronbach's alpha, based on the collected data.

However, the implementation of the validation process, including when and how it is conducted, remains the responsibility of the organizers. Depending on available resources, this may involve administering the generated surveys to a control group before the camp, or performing post-hoc validation using the responses from a single cohort at the end of the activity.

Furthermore, generating the questionnaire in LimeSurvey through the platform enables the seamless export of data that can be directly analyzed using the pre-configured notebooks. By assigning matching labels to the survey questions during the generation process, the platform ensures that the collected data can be automatically linked with the corresponding analysis.

C. Data Analysis and Interpretation

Data analysis and interpretation represent the most critical phase of the evaluation process, enabling organizers and stakeholders to derive meaningful insights about the impact of the activities undertaken. However, data analysis is rarely a linear or fully generalizable process. As emphasized in the literature on educational measurement and social science research (e.g., Maxwell's work on qualitative inquiry [26]), effective analysis depends heavily on the context, goals, and structure of the collected data. For non-expert users, this complexity can pose significant challenges.

To address these difficulties, the proposed platform supports a dual approach. On one hand, it provides a set of ready-to-use Jupyter notebooks tailored to the survey configuration chosen by the user. These notebooks guide the analyst through a basic data analysis pipeline, including data import, cleaning, outlier detection, and the application of standard techniques such as regression and multivariate analysis. While the notebooks are not meant to replace the role of a trained data scientist, they offer a transparent and extensible framework that reduces the technical burden on users.

On the other hand, to further support interpretation, the platform includes an optional notebook equipped with prompts and suggestions designed to work in conjunction with large language models (LLMs), aligning with a growing trend in educational research to adopt LLMs for supporting instructional tasks [27], [28]. This tool assists users, particularly those without a strong background in statistics, in understanding the meaning of key outputs and how they might inform future program design.

IV. Platform Architecture

The platform is built to be released free and open-source. Surveys in fields involving underage students often require strict privacy measures; as a consequence, the ability to self-host the platform allows for more control over the data being handled. The choice of technologies reflects these principles; in particular, the platform is divided into a central application and several integrated external tools, all open-source and self-hostable.

As shown in Figure 2, the central platform is responsible for managing the survey lifecycle, from creation to deployment, as well as integrating with external tools for data collection and analysis. The technology chosen for this part is Django [3], a Python framework widely used in open-source projects. Django was selected for its scalability, security, and extensibility. Its internal architecture is well-suited to handle complex workflows like survey management while also providing an admin interface that simplifies user interaction.

Once the user completes the survey definition, a Python package called Citric[4] is used to interact with the LimeSurvey platform[5], automating the transfer of the survey to LimeSurvey for deployment. LimeSurvey is a well-known, open-source, self-hosted tool for survey creation and data collection, which aligns with the platform's goals of maintaining control over survey data. While many survey platforms are not specifically designed for scientific research [17], LimeSurvey is a widely used and well-maintained solution that reduces both development time and maintenance burden.

The central platform also generates Jupyter Notebooks [6] for data analysis. The choice of Jupyter Notebooks is based on their widespread use in the research community and their ability to integrate code and text in a single, interactive

[3]https://www.djangoproject.com/
[4]https://github.com/edgarrmondragon/citric
[5]https://www.limesurvey.org/
[6]https://jupyter.org/

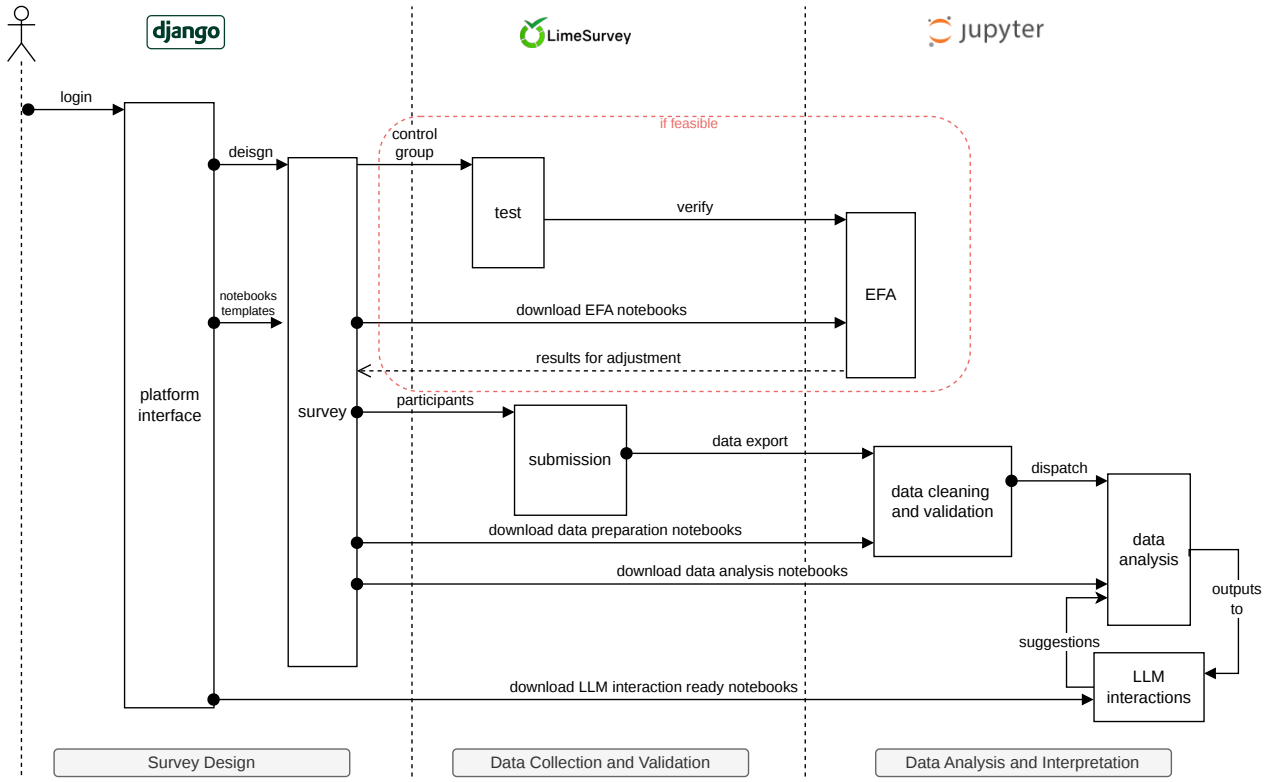


Fig. 2. User interaction flow

environment. This supports reproducibility in research [29], allowing users to document their analysis while performing it. Additionally, the Django template system enabled the customization of notebooks for each survey, allowing the creation of notebooks that fit research questions chosen in the design phase.

The generated notebooks integrate both explanatory text and executable code, allowing users to understand each step of the analysis and modify it as needed. The notebooks are divided into preparation and analysis notebooks. The formers are designed to verify the validity of the survey through Exploratory Factor Analysis (EFA), clean the data by removing any outliers, and assess consistency using Cronbach's alpha. Each step is accompanied by visualizations of the results, such as scree plots and factor loadings, to help users better understand the findings. The analysis notebooks assist users in performing further analysis; in addition to providing descriptive representations of the results, they offer templates for executing regression analysis on key variables and provide appropriate verifications and suggestions for improving the model fit.

Furthermore, the platform offers additional notebooks with the integration of Large Language Models (LLMs) to support users in interpreting their analysis. The main purpose is to provide contextualised suggestions and interpretations for users who do not have a strong background in data analysis. While it is well-known that LLMs cannot substitute for statisticians and that improper use of LLMs can even lead to mistakes that invalidate results, recent research by Zhu et al. [30] demonstrated that certain LLMs can perform reasonably well for relatively simple tasks, such as providing interpretations and suggestions for basic statistical analysis. This research helped guide the selection of an LLM that balances performance with the specific needs of our platform.

The final step in ensuring that the platform is truly self-hostable was the creation of a standardized deployment process that includes all the necessary components. Docker, and specifically Docker Compose, was selected for this task. A Docker Compose configuration[7] is provided, which encapsulates all the services required for the platform to function correctly. This configuration allows end users to launch the entire platform on their premises with minimal setup, requiring only a machine with Docker pre-installed. This approach greatly simplifies deployment, ensures consistency across different environments, and reduces the setup effort for users, allowing them to take full control of their data and platform infrastructure. All source code, configuration files, and usage instructions are openly available in the project repository at

[7]https://www.docker.com/

https://gitlab.com/frfaenza/elevate-ai.

V. User Interaction Workflow

The platform's design prioritizes usability, allowing non-expert users, such as researchers and educators, to effortlessly deploy and analyze surveys for outreach programs. The interaction process is intuitive and modular, guiding users through the entire survey lifecycle, from design to data analysis. The proposed workflow reflects our own experience with outreach initiatives [31].

Following the flow presented in Figure 2, users begin by creating a new evaluation process. They define a title, a description, and optional metadata to help document the assessment process. Additionally, they specify a start and end date, which are used by the platform to automate the creation of surveys in LimeSurvey. The overall sequence of steps and system interactions is also graphically illustrated in Figure 3.

The next step is to select questions to include in the evaluation. To encourage adoption, a set of questions frequently used in the literature is already embedded in the Django migration system, ensuring the database is pre-populated at deployment. This does not prevent users from creating new, custom questions tailored to their specific needs. Each question is linked to a literature reference, making its origin traceable. Questions are also grouped under constructs (e.g., self-efficacy, motivation), a feature implemented in the model schema to facilitate notebook generation for subsequent analysis.

By default, the evaluation is structured as a pre-post comparison. During survey composition, the user can specify whether each question should appear in the pre-survey, the post-survey, or both. It is also possible to mark questions as mandatory or conditionally displayed based on prior responses. While these are just a subset of LimeSurvey's full capabilities, they cover the most commonly used options in educational research.

Once the questions, order, and pre/post assignment have been finalized, the user can invoke the function that uses Citric to generate and upload the surveys to the LimeSurvey instance.

An essential aspect of the process is survey validation via Exploratory Factor Analysis (EFA). As highlighted in Figure 2, this step is optional and marked in red, as its feasibility depends on the specific constraints of the organizers. A robust validation process ideally requires a control group to complete the survey, followed by EFA to verify that the constructs align with the collected data (see Knekta et al. [15]). The platform provides a notebook to perform EFA, allowing practitioners to make informed decisions about modifying or retaining the current questionnaire structure.

If validation with a control group is not feasible, the user can proceed directly to deployment using LimeSurvey's native interface and then validate with participants' data. The platform generates both a public link and a QR code for distribution, though direct interaction with LimeSurvey remains possible for users with advanced needs.

One factor that determines whether further interaction with the LimeSurvey platform is needed is the method used to connect pre- and post-surveys. The default approach allows for multiple options. The first and easiest way is to use participants' email addresses. In this case, the provided notebook will assign a pseudonym and remove the email for privacy reasons. Alternatives could be assigning random codes or letting users generate their own codes through a specific question.

Once data has been collected, the analysis phase begins. The deployment includes a Jupyter server, allowing users to launch pre-generated analysis notebooks directly. Based on the selected survey configuration, the following notebooks are provided:

- **data cleaning notebook**: Processes exported responses, cleans the dataset, and connects pre- and post-survey entries
- **outliers and validity notebook**: Identifies outliers (e.g., through response time [32] _) and performs validity checks using EFA and Cronbach's alpha
- **descriptive analysis notebook**: Provides summary statistics and visualizations based on survey constructs
- **regression notebook**: Offers pre-configured regression models with customization instructions

While the analysis notebooks include embedded explanations, statistical knowledge is still required. To support non-expert users, the platform offers an additional notebook that integrates with a Large Language Model (LLM). The user will be able to carry on the analysis while interacting with LLM for assistance.

Following Zhu et al. [30], who compared LLM performance in statistical reasoning, the platform incorporates LLM-generated prompts to help users interpret outputs and navigate the analysis steps. To preserve the open-source and self-hostable nature of the platform, we support integration with local models via the Hugging Face framework[8]. Although these models can run on modern laptops with longer response times, they may be too resource-intensive for some users. Therefore, we also provide a notebook that connects to the ChatGPT-4 API[9], accompanied by prompt templates similar to those used locally. In both cases, however, users must register on a platform, either Hugging Face or OpenAI (for ChatGPT), to obtain an API key, which must be inserted into the notebook for it to function correctly.

VI. Conclusions and Future Development

In this paper, we presented an open-source, self-hostable platform designed to support the evaluation of STEM and CS outreach initiatives. The platform aims to address the lack of appropriate assessment during extracurricular activities. By automating survey generation, facilitating data validation, and guiding users through standard analysis workflows, including support from integrated Large Language Models (LLMs), our platform lowers the technical barrier to performing robust evaluations in informal settings.

The platform's design is grounded in educational research and informed by practical field experience. It aligns with

[8]https://huggingface.co
[9]https://chatgpt.com/

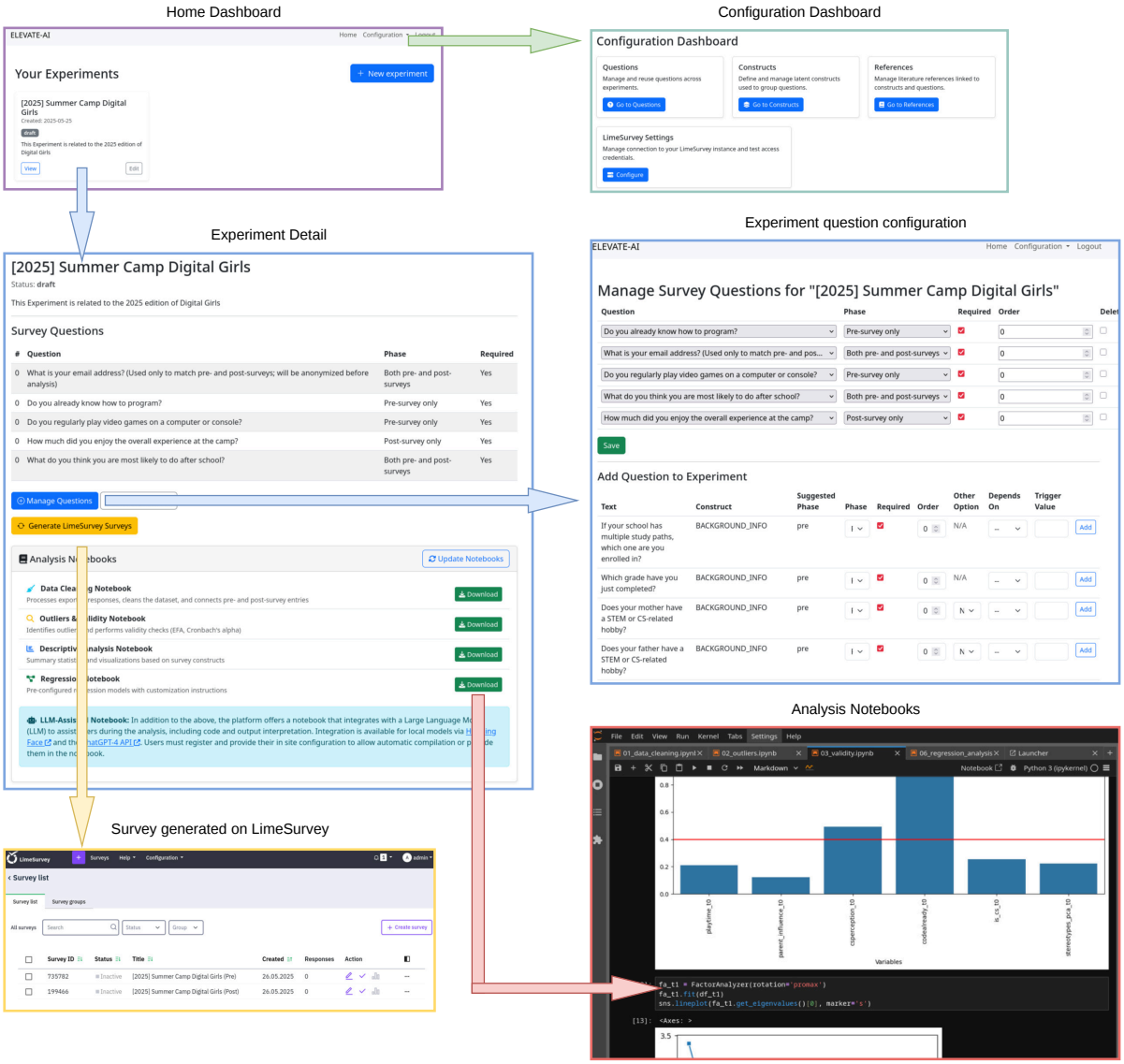


Fig. 3. Use case flow

widely adopted practices in survey validation and quantitative analysis. The tool builds upon several years of implementation in gender-focused STEM outreach, notably the Digital Girls initiative, and has been refined to meet the needs of organizers working in contexts where both methodological rigor and usability are essential.

The project's open-source nature is intended to foster a community of contributors who can collaboratively expand the platform by adding validated questions with corresponding literature references, sharing analysis templates, and suggesting feature improvements. All source code, deployment files, and usage documentation are freely available at https://gitlab.com/frfaenza/elevate-ai.

Looking ahead, several lines of development are planned. First, we aim to expand the LLM-assisted functionalities to support users not only during data analysis but also in survey design and validation. Furthermore, inspired by recent findings from Zhu et al. [30], we are exploring the possibility of integrating a fine-tuned model optimized explicitly for the types of evaluation workflows supported by the platform.

Finally, we plan to further simplify the user experience by streamlining access to the analysis tools. A step in this direction is the integration of Jupyter notebooks directly within the Django interface, removing the need for separate notebook deployment or configuration. Similar approaches have been adopted in prior research focused on improving the accessi-

bility of computational notebooks for non-technical users [33].

REFERENCES

[1] European Commission, "STEM Education Strategic Plan," https://education.ec.europa.eu/sites/default/files/2025-03/STEM_Education_Strategic_Plan_COM_2025_89_1_EN_0.pdf, Mar. 2025, accessed: 2025-07-15.

[2] OECD, "Gender, Education and Skills," https://www.oecd.org/en/publications/gender-education-and-skills_34680dd5-en.html, 2023, accessed: 2025-07-15.

[3] M. M. Msambwa, K. Daniel, C. Lianyu, and F. Antony, "A systematic review using feminist perspectives on the factors affecting girls' participation in stem subjects," *Science & Education*, pp. 1–32, 2024. doi: 10.1007/s11191-024-00524-0

[4] F. Beroíza-Valenzuela and N. Salas-Guzmán, "Stem and gender gap: A systematic review in wos, scopus, and eric databases (2012–2022)," in *Frontiers in Education*, vol. 9. Frontiers Media SA, 2024. doi: 10.3389/feduc.2024.1378640 p. 1378640.

[5] A. S. Rorrer, "An evaluation capacity building toolkit for principal investigators of undergraduate research experiences: A demonstration of transforming theory into practice," *Evaluation and program planning*, vol. 55, pp. 103–111, 2016. doi: 10.1016/j.evalprogplan.2015.12.006

[6] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two decades of artificial intelligence in education," *Educational Technology & Society*, vol. 25, no. 1, pp. 28–47, 2022.

[7] E. A. Locke, "Self-efficacy: The exercise of control," *Personnel psychology*, vol. 50, no. 3, p. 801, 1997.

[8] B. P. Cipriano, N. Fachada, and P. Alves, "Drop project: An automatic assessment tool for programming assignments," *SoftwareX 18 (2022) 101079*, vol. 18, 2022. doi: 10.1016/j.softx.2022.101079

[9] S. H. Edwards and M. A. Perez-Quinones, "Web-cat: automatically grading programming assignments," in *Proceedings of the 13th annual conference on Innovation and technology in computer science education*, 2008. doi: 10.1145/1384271.1384371 pp. 328–328.

[10] J. Majerník, "E-assessment management system for comprehensive assessment of medical students knowledge," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2018. doi: 10.15439/2018F138 pp. 795–799.

[11] L. Stanescu and B. Savu, "Automatic assessment of narrative answers using information retrieval techniques," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019. doi: 10.15439/2019F96 pp. 355–358.

[12] B. Van Dusen, "Lasso: A new tool to support instructors and researchers," *arXiv preprint arXiv:1812.02299*, 2018. doi: 10.48550/arXiv.1812.02299

[13] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008.

[14] A. Decker and M. M. McGill, "A topical review of evaluation instruments for computing education," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019. doi: 10.1145/3287324.3287393 pp. 558–564.

[15] E. Knekta, C. Runyon, and S. Eddy, "One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research," *CBE—Life Sciences Education*, vol. 18, no. 1, p. rm1, 2019. doi: 10.1187/cbe.18-04-0064

[16] I. Fronza, P. Ihantola, O.-P. Riikola, G. Iaccarino, T. Mikkonen, L. García Rytman, V. Lappalainen, C. Rebollo Santamaría, I. Remolar Quintana, and V. Rossano, "Towards s'more connected coding camps," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 2025. doi: 10.1145/3641554.3701849 pp. 353–359.

[17] D. Guse, H. R. Orefice, G. Reimers, and O. Hohlfeld, "Thefragebogen: A web browser-based questionnaire framework for scientific research," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019. doi: 10.1109/QoMEX.2019.8743231 pp. 1–3.

[18] F. Faenza, C. Canali, M. Colajanni, and A. Carbonaro, "The digital girls response to pandemic: Impacts of in presence and online extracurricular activities on girls future academic choices," *Education Sciences*, vol. 11, no. 11, p. 715, 2021. doi: 10.3390/educsci11110715

[19] F. Faenza, C. Canali, and A. Carbonaro, "Ict extra-curricular activities: The "digital girls" case study for the development of human capital," in *The International Research & Innovation Forum*. Springer, 2021. doi: 10.1007/978-3-030-84311-3_18 pp. 193–205.

[20] P. Cobb, K. Jackson, and C. Dunlap, "Design research: An analysis and critique," in *Handbook of international research in mathematics education*. Routledge, 2015, pp. 481–503.

[21] L. Markauskaite, P. Freebody, and J. Irwin, *Methodological choice and design: Scholarship, policy and practice in social and educational research*. Springer Science & Business Media, 2010, vol. 9.

[22] C. Canali and F. Faenza, "An evaluation tool for extracurricular activities to reduce the gender gap in computer science," in *ICGR 2023 6th International Conference on Gender Research*. Academic Conferences and publishing limited, 2023. doi: 10.34190/icgr.6.1.1026

[23] I. Ajzen, "Nature and operation of attitudes," *Annual review of psychology*, vol. 52, no. 1, pp. 27–58, 2001. doi: 10.1146/annurev.psych.52.1.27

[24] J. F. McKenzie, M. L. Wood, J. E. Kotecki, J. K. Clark, and R. A. Brey, "Establishing content validity: Using qualitative and quantitative steps." *American Journal of Health Behavior*, vol. 23, no. 4, 1999.

[25] A. R. Artino Jr, J. S. La Rochelle, K. J. Dezee, and H. Gehlbach, "Developing questionnaires for educational research: Amee guide no. 87," *Medical teacher*, vol. 36, no. 6, pp. 463–474, 2014. doi: 10.3109/0142159X.2014.889814

[26] J. A. Maxwell, *Qualitative research design: An interactive approach: An interactive approach*. sage, 2013.

[27] E. Rudolph, H. Seer, C. Mothes, and J. Albrecht, "Automated feedback generation in an intelligent tutoring system for counselor education," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024. doi: 10.15439/2024F1649 pp. 501–512.

[28] T. C. Freitas, M. J. V. Pereira, A. C. Neto, and P. R. Henriques, "Goliath, a programming exercises generator supported by ai," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024. doi: 10.15439/2024F8479 pp. 331–342.

[29] F. Perez and B. E. Granger, "Project jupyter: Computational narratives as the engine of collaborative data science," *Retrieved September*, vol. 11, no. 207, p. 108, 2015.

[30] Y. Zhu, S. Du, B. Li, Y. Luo, and N. Tang, "Are large language models good statisticians?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 697–62 731, 2024.

[31] D. Maniglia, F. Faenza, and C. Canali, "Empowering girls in cs: The impact of digital girls outreach camp," in *Proceedings of The 7th International Conference on Gender Research*. Academic Conferences and publishing limited, 2025. doi: 10.34190/icgr.8.1.3528

[32] G. Szyjewski and L. Fabisiak, "Survey as a source of low quality research data," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2017. doi: 10.15439/2017F266 pp. 939–943.

[33] S. Pigozzi, F. Faenza, and C. Canali, "Sophon: An extensible platform for collaborative research," in *Practice and Experience in Advanced Research Computing 2022: Revolutionary: Computing, Connections, You*, 2022. doi: 10.1145/3491418.3535163 pp. 1–4.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 507–518
DOI: 10.15439/2025F8393
ISSN 2300-5963 ACSIS, Vol. 43

Bridging Global Language Models and Local Spatial Data: The JackDaw Approach to Context-Aware Agriculture and Rural Planning

Karel Charvát [1], (Stein) Runar Berhgeim [2], Raul Palma [3], Adam Aron Rynkiewicz [3], Matúš Botek [5], Alexander Kovalenko [4], Pavel Kordík [4], Antonín Kubíček [5], Šárka Horáková [1], Markéta Kollerová [5]

[1] Czech Center for Science and Society, Radlická 663/28, Smíchov, 150 00 Praha, Czech Republic
Email: charvat@ccss; horakova@ccss.cz
[2] Asplan Viak Internet Ltd, Havnegården (Kystveien 14), Postboks 701 Stoa
Arendal, Aust-Agder 4808, Norway
Email: rb@avinet.no
[3] Poznan Supercomputing and Networking Center, ul. Jana Pawła II 10, 61-139 Poznan, Poland
Email: rpalma@man.poznan.pl; adam.rynkiewicz@doctorate.put.poznan.pl
[4] Faculty of Information Technology, Czech Technical University in Prague
Thákurova 9, 160 00 Praha 6, Czech Republic
Email: alexander.kovalenko@fit.cvut.cz; Pavel.Kordik@fit.cvut.cz
[5] Plan4all z.s., K Rybníčku 557, 33012 Horní Bříza, Czech Republic
Email: matus.botek@gmail.com, tony.kubicek@gmail.com; marketa.kollerova@plan4all.eu

***Abstract*—Large Language Models (LLMs) excel at synthesising globally documented knowledge but lack the fine-grained, real-time awareness required for field-level agricultural and rural-planning decisions. This paper introduces JackDaw, a spatially enabled chat-agent architecture that couples foundation-model reasoning with multi-modal geospatial data streams and a retrieval-augmented generation (RAG) pipeline. JackDaw implements a tool-prefiltering mechanism that selects only those data connectors whose topical, temporal and spatial metadata match the current query, thereby mitigating the diminishing returns observed when LLMs are exposed to large, flat toolsets. Through LangChain-based orchestration the platform dynamically assembles workflows that range from lightweight natural-language processing models to domain-specific analytic kernels, while a value-engineering strategy allocates computationally intensive models (e.g., GPT-4-class) only to tasks that require broad contextual reasoning. Benchmark experiments on forestry-asset discovery and vineyard-site assessment demonstrate that JackDaw delivers location-specific, traceable answers that outperform a standalone proprietary LLM, which provides only generic or spatially misattributed responses. The results confirm that bridging global language models with local spatial intelligence markedly reduces hallucination rates and enhances the operational readiness of AI for sustainable agriculture and rural development.**

***Index Terms*—Large language models; geospatial AI; retrieval-augmented generation; context-aware agriculture; spatial decision support; tool prefiltering; JackDaw system; rural planning.**

I. Introduction

ARTIFICIAL intelligence (AI) has transitioned from a niche computational technique to a pervasive epistemic lens through which agriculture and rural development are increasingly interpreted. In agronomic contexts, AI is not merely a collection of algorithms but a socio-technical apparatus that converts raw environmental signals into actionable knowledge, thereby mediating the long-standing tension between global food-system efficiency and local stewardship of land, labour and biodiversity. By formalising inductive reasoning at unprecedented spatial and temporal scales, AI promises to reconcile high-frequency decision support for producers with broader rural-development directives, resilience, equity, sustainability, provided its deployment remains transparent, uncertainty-aware and ethically aligned [1], [2].

AI pipelines now ingest heterogeneous data streams from in-situ sensors (soil-moisture probes, automated weather stations), passive and active satellite constellations (multispectral Sentinel-2 MSI, C-band Sentinel-1 SAR, DESIS hyperspectral, LiDAR), low-altitude unmanned-aerial systems [3], and physics-based or statistical climate re-analyses (GFS, ERA-5-Land). Early studies relied on single-sensor inputs and shallow classifiers; contemporary frameworks fuse multi-temporal, multi-modal observations via machine-and deep-learning architectures optimized for high-dimensional spatiotemporal covariance structures. Sentinel-2 reflectance cubes harmonised with

Sentinel-1 back-scatter underpin continental land-use-intensification inventories [4] and sub-field crop-type delineation [5]. Binary classifiers trained on multi-season Sentinel mosaics further attain F1 $\geq$ 0.90 for key commodity classes [6]. Ensemble decision-tree learners, notably Random Forest, remain baseline regressors for temperate-cereal yield [7] and national wheat statistics [8], while deep neural networks gain traction where label density is sufficient: One study [9] stacked convolutional, gated-recurrent, and transformer encoders on fused Sentinel-1/2 sequences for pixel-level classification; The method proposed in [10] introduced the 3-D $\rightarrow$ 2-D HypsLiDNet that couples DESIS voxels with LiDAR-derived canopy structure; multi-site neural networks now invert hyperspectral reflectance directly to nitrogen and leaf-area traits with ≈10 % relative RMSE [11]. Research in [12] demonstrated multimodal convolutional fusion of Sentinel-2, edaphic layers and mesoscale meteorology for field-scale yield forecasting. Object-based learning exploiting explicit geostatistical segmentation persists for landscape stratification, findings from [13] combined Sentinel-2A, ALOS DSM, and PALSAR layers with Random Forest to derive agro-ecological strata lightweight transformer networks distilled from deep vision models extend the approach to UAV weed-mapping with reduced compute overhead [14], and temporal self-attention mechanisms have been applied to integrated crop-livestock delineation in the subtropics [15]. Wave-equation-informed convolutions also elevate spatio-temporal forest-monitoring accuracy, pointing to transferable architectures for land-cover dynamics [16].

Operational deployment is constrained by observation gaps (cloud cover, revisit frequency) and by aleatoric uncertainty in multi-sensor fusion. Research described in [17] mitigated these issues by embedding Sentinel-1 SAR into the UnCRtainTS cloud-removal network, lowering spectral MAE to ≈ 0.025 reflectance units, and enabling continuous NDVI trajectories across Central-European croplands. On the prognostic front, high-resolution climatic forcing has been advanced by up-scaling 0.25° GFS products to ERA-5-Land resolution via a U-Net super-resolution model, merging the output with HadISD station records, and applying CatBoost regression to reduce 24-h temperature MAE to ~1 °C [18]. Nevertheless, most agronomic AI studies still rely on internal cross-validation; spatially independent or inter-annual robustness tests remain rare; examples include paddock-out validation [19] and temporally stratified ten-fold splits [20]. Consequently, the methodological frontier is defined by scalable multi-sensor fusion, uncertainty-aware ensembles or transformers, and rigorous cross-region benchmarking to support operational use in dynamic agroecosystems.

Large-language-model (LLM) research entered agricultural informatics only recently, yet adoption is rapid. A systematic scan identified 26 geospatial-LLM papers up to early 2024, almost all post-November 2022, describing the field as “embryonic but rapidly accelerating” [21]; a parallel survey of multimodal ag-LLMs reached similar conclusions and noted rudimentary benchmarks [22]. Developments have progressed from demonstration chatbots to end-to-end systems: ChatGeoAI converts plain-language queries into PyQGIS code [23]; OmniGeo demonstrates zero-shot reasoning over satellite imagery, vector polygons, and free text using an instruction-tuned vision-language model [24]. Text-focused agronomic advisers include a retrieval-augmented fruit-and-vegetable system coupling Baichuan-2 with RAG to reduce hallucination rates by 10–40 % [25] and a cotton-soil engine where a fine-tuned GPT-2 outperforms LLaMA-2 for nutrient recommendations [26]. Synthetic data generation is exemplified by an LLM-driven agentic workflow that fabricates realistic fault patterns for smart-tractor telemetry, enabling predictive maintenance testing). Document intelligence pipelines empower domain-agnostic LLMs to extract structured pest information from agronomic literature in zero-shot mode [27]. Complementary graph-generation pipelines capture pest–symptom ontologies for tea production, easing structured decision-support [28].

Across these studies, shared hurdles are evident. High-quality labeled agri-text corpora and paired image–text datasets remain scarce, forcing reliance on generic foundation models or few-shot prompting [25] [22]. Factual reliability is fragile, motivating retrieval-augmented generation [21]. Multimodal grounding across satellite, drone, sensor, and text streams is largely heuristic, as shown in [22] [24], while the compute and energy costs of fine-tuning multi-billion-parameter models raise sustainability concerns. Explainability, data governance, and ethical compliance are critical yet under-explored prerequisites for commercial farm deployment [21].

Within two harvest seasons, the community has advanced from isolated demonstrations to retrieval-augmented, multimodal assistants addressing yield prediction, soil management, pest intelligence, and equipment maintenance. Consolidating open data resources, curbing hallucinations, mastering efficient geo-text fusion, and delivering transparent, farmer-oriented explanations now constitute the central agenda for AI-driven rural development.

II. MOTIVATION

Large Language Models (LLMs) have significantly impacted various domains by leveraging vast datasets to provide comprehensive responses on diverse topics. Their capability to process global and well-documented information allows them to efficiently recognize general trends, interpret historical data patterns, and detail well-known geographical areas. However, despite these substantial strengths, LLMs frequently exhibit limitations

concerning the provision of precise, real-time insights required for location-specific and dynamic contexts, such as agriculture and rural development.

In agricultural practices, where decisions rely heavily on real-time data about local environmental conditions, precise field-level analytics, and seasonal variability, the limitations of standalone LLMs become particularly pronounced. For instance, while an LLM may comprehensively discuss global agricultural trends or extensively documented practices, it may lack critical, timely insights into crop conditions in a less-documented rural area or the subtle seasonal variations affecting local agricultural productivity.

Integrating LLMs with geospatial data offers a robust solution to bridge these gaps. Real-time geospatial data, encompassing remote sensing imagery from satellites, drones, weather stations, and in-situ sensors, can complement the generalized insights of LLMs with detailed, actionable information at specific locations and moments in time. By coupling spatial analysis technologies (already established through remote sensing platforms such as Sentinel-1 and Sentinel-2, as well as advanced AI-driven methodologies) with the broad contextual knowledge offered by LLMs, systems like JackDaw have demonstrated the potential to significantly enhance rural-urban planning and agricultural decision-making.

The JackDaw prototype has underscored the value of this integrative approach by successfully combining the power of LLM reasoning with spatial analysis to provide context-aware insights critical for agriculture and rural management. Its approach exemplifies how synthesizing multi-modal data streams with deep-learning architectures can mitigate the limitations stemming from gaps in sensor observations or uncertainties in data fusion. Furthermore, JackDaw exemplifies the potential of multi-modal integration to improve practical agricultural outcomes by offering more precise yield forecasts, targeted resource management recommendations, and improved environmental monitoring.

Thus, developing an integrative platform that combines the strengths of LLMs with standard agricultural AI tools and GIS solutions represents a strategic opportunity. Such integration promises substantial advancements in rural agricultural practices, enabling more robust, evidence-based decision-making, enhancing sustainability, and ultimately improving rural livelihoods and economic development.

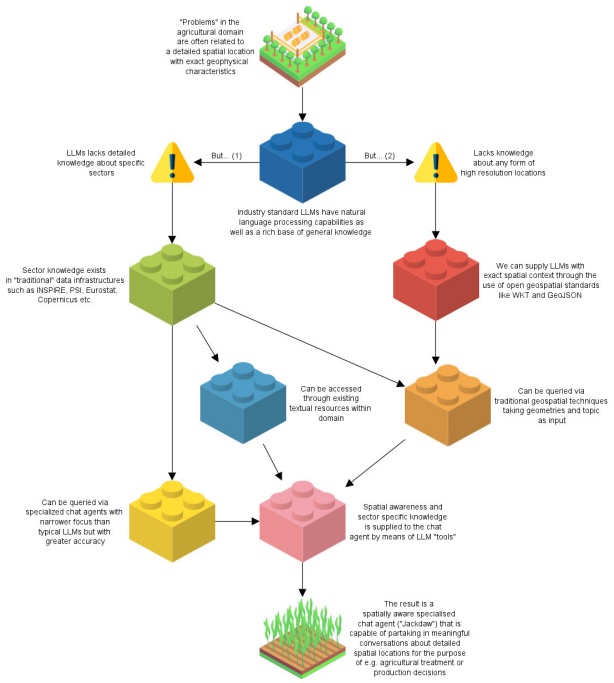


Fig 1. The motivation for developing the JackDaw chat agent

III. SOLUTION

The JackDaw solution is designed to enhance the functionality of LLMs by addressing their inherent limitations in accessing and interpreting localized, real-time data. This enhancement is significant in agricultural and rural planning professions, where precise localized knowledge is crucial. The gap between the global knowledge embedded in LLMs and the need for detailed local contextual information is significant. JackDaw attempts to bridge this gap by integrating advanced geospatial data analysis and specialised AI tools into conversational AI interfaces. This integration enables more accurate and contextually aware responses from LLMs than what is possible with the foundation models.

The conceptual solution underlying JackDaw is facilitating meaningful conversations about geographical locations through LLMs. To achieve this, it is essential for LLMs to be not only aware of the location being discussed but also to possess an understanding of what specific data is relevant to that place. Furthermore, to provide useful responses, JackDaw requires the model to discern which data sources should be utilised based on the topic, purpose, and spatial location of the conversation. These LLMs are expected to engage iteratively with various data sources to foster an incremental understanding of the context-specific to each inquiry.

Conceptual workflow of a conversation with the JackDaw chat agent

JackDaw operates using a systematic workflow that begins when a user initiates a query related to a specific geographic location through a user message to the agent.

The system interprets this query and tries to match it against a large number of available “tools” that connect the LLM to a wide range of relevant data sources. These sources include spatial datasets, such as those defined by the Infrastructure for Spatial Information in Europe (INSPIRE), public sector information (PSI), and specialised AI models optimised for agricultural analytics.

Upon identifying which “tools” may be relevant to answering the question, JackDaw proceeds to pass them to the language model, which collects both spatial and tabular data sets pertinent to the user's query. This includes sensor-derived field data, satellite imagery, current weather forecasts, soil properties, and agronomic statistics. Specialised AI models process these datasets to generate analytical insights, which are then contextualised and embedded into the chat interface.

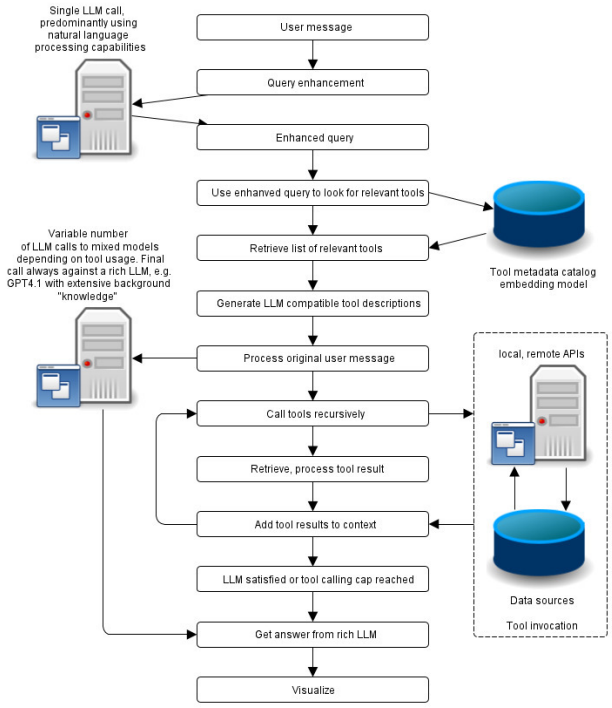


Fig 2. Conceptual flow of a conversation with the JackDaw spatial-enabled chat agent

Visualisations derived from the processed datasets are integrated into the user interface, providing graphical context to the inquiries. These visual representations facilitate clearer insights and more actionable decision-making. By embedding enriched data insights into the LLM's reasoning process, JackDaw effectively bridges the gap between general global knowledge and the need for localized, actionable information. This approach offers robust and informed answers to user inquiries, promoting enhanced decision-making processes tailored specifically to rural and agricultural contexts.

Through the innovative integration of geographic information systems (GIS) and advanced AI-driven data analytics, JackDaw capitalises on existing expertise within these domains. This harnessing of advanced technologies promotes a nuanced and effective decision-making process, significantly enhancing LLM capabilities in understanding and responding to location-specific questions. Consequently, JackDaw transforms the application of LLMs in agricultural planning and rural development, making them more responsive and accurate in addressing context-specific inquiries.

Enhancing Chat Conversations with Spatial Context

Integrating an area of interest into a chat query can significantly augment user interaction by providing spatial context to the conversation. This enhancement involves adding spatial features that encapsulate the geographical areas users wish to inquire about. The spatial data is represented using Well-Known Text (WKT) geometry features, a format allowing a detailed description of geometric shapes in a human-readable text string. By incorporating WKT geometry, users can define specific areas of interest directly within the chat interface.

Ways of Incorporating Spatial Features

Users can add spatial features to a chat query in two primary ways: through a client interface or via natural language processing (NLP). When utilising a client interface, users can directly draw the areas of interest onto a map. This method relies on graphical tools that allow precise selection and adjustment of shapes to represent the desired geographical space accurately. These shapes are then converted into WKT format, maintaining their geometric characteristics while enabling text-based communication in the chat. This method provides users with a tangible and intuitive way to specify areas of interest.

Alternatively, spatial features can be derived from NLP queries. In this approach, a tool or algorithm processes the user's natural language inputs to extract relevant geographical information and convert it into WKT geometry. This method leverages advances in NLP to interpret user intent and spatial references embedded within the chat. Such tools can identify place names, landmarks, or specific directions mentioned by users and translate them into spatial features. This capability allows users to describe their areas of interest using everyday language, simplifying the process and making it accessible to a broader range of users.

Facilitating Function and Tool Integration

Once the spatial features are integrated into the chat, they are encapsulated within code blocks. These code blocks are clearly delineated to distinguish them from other elements of the chat. This demarcation ensures that

the data is accurately identified and processed by subsequent tools or functions that the chat agent may call upon. Code blocks serve not only to organise information but also to maintain data integrity when transferring inputs across different operational environments.

Incorporating spatial features as part of chat queries empowers chat agents to utilise this data in subsequent interactions. The defined area of interest can be used as a parameter for various function or tool calls. For instance, a chat agent might access weather data, perform geographical analysis, or retrieve demographic information specific to the identified area. This capacity expands the functionality of chat agents, enabling more sophisticated queries and responses tailored to the user's spatial context.

Adding areas of interest to chat queries through WKT geometry represents a meaningful advancement in interactive technology. By allowing users to define spatial parameters using both graphical interfaces and natural language processing, the system becomes more flexible and user-centric. Furthermore, the encapsulation of data within distinct code blocks facilitates seamless integration with other tools and functions. Such enhancements lead to more effective and targeted interactions, bridging the gap between spatial awareness and conversational interfaces. This integration serves as a transformative enhancement in fields where geographical context is critical, such as urban planning, logistics, and environmental monitoring, thereby expanding the utility and applicability of chat-based systems in professional and research environments.

Connecting data sources to LLMs

Implementing knowledge acquisition in large language models (LLMs) involves equipping these models with an understanding of various "things," or categories of information, which can be used to improve the applicability and accuracy of chat agents. This methodology necessitates the development of versatile tools that connect LLMs with relevant data sources. LLM tools serve a crucial function in facilitating this connection, providing the necessary scaffolding to bridge the gap between an LLM's inherent language skills and its access to a structured knowledge base.

Versatility is a fundamental requirement when constructing these tools. They must be able to interlink effectively with multiple data sources, thereby expanding the functional capabilities of chat agents. To achieve this, tool prompt engineering is an essential process. This involves designing specific prompts that guide the retrieval and processing of data in a manner that aligns with the operational goals of the LLM. Making each tool available through a web service enhances accessibility and usability, allowing these tools to be integrated seamlessly into web-based applications.

Developing numerous specialised tools can significantly enhance the LLM's ability to handle complex inquiries. In the realm of geophysical information, for instance, tools can analyse terrain by examining landforms, elevation, aspect, slope, and surface roughness. This data provides essential context for applications ranging from environmental monitoring to urban planning. Similarly, climatic tools that gather information about temperature, wind patterns, cloud cover, precipitation, and solar radiation contribute to a comprehensive understanding of environmental conditions across different regions and times, enriching predictive models and research studies.

Demographic information represents another vital category, enabling insights into population characteristics and trends. Tools that analyse demographic data must consider population size, growth projections, historical trends, and the composition of populations based on gender and age. This data supports social dynamics, policy making, and economic forecast analyses. Economic tools can assess business demography and market sizes, offering valuable input for economic modeling, investment analysis, and strategic planning.

Temporal tools are indispensable for temporal analysis, which is critical in assessing real-time changes and zeitgeist trends. When coupled with historical data, a snapshot of current weather conditions enables detailed climatic assessments. Moreover, zeitgeist tools, such as those utilising the Global Database of Events, Language, and Tone (GDELT), are instrumental in capturing contemporary socio-political dynamics, offering insights into prevailing public sentiment and events affecting international relations and market behaviours.

Incorporating statistical data with georeferenced identifiers is instrumental for deriving meaningful insights based on location-specific features. By identifying Nomenclature of Territorial Units for Statistics (NUTS) identifiers for particular locations, it is possible to gather relevant demographic and business demography statistics. This georeferenced approach enables tailored analyses considering local variations, which is essential for targeted policymaking and regional development strategies.

The utilisation of raster and grid data complements this analysis by providing detailed spatial information. Sources such as the European Space Agency (ESA), the Group on Earth Observations (GEO), and Sentinel data offer high-resolution satellite imagery and derived raster datasets. These datasets are optimised for large-scale multidimensional queries, allowing users to extract information that spans multiple dimensions and scales. This capability is particularly useful in applications like precision agriculture, land use planning, and disaster management.

Vector data, on the other hand, involves using line and polygon data formats typical of traditional Geographic Information System (GIS) infrastructures. Adhering to standards such as those established by the Infrastructure for Spatial Information in the European Community (INSPIRE), vector data provides a detailed locational

context crucial for detailed mapping and spatial analysis. This infrastructure supports a wide array of applications, from infrastructure development to environmental conservation efforts.

Document data serves an integral role in the holistic use of LLM tools, particularly in contexts that require the retrieval of legal documents, plans, reports, and other contextual information. Access to such documents allows LLMs to provide informed responses that are grounded in the regulatory and historical framework of the area in question. For instance, in urban development scenarios, understanding zoning laws and historical city planning documents is essential to delivering relevant insights and recommendations.

Limiting the number of tools made available to the LLM for improved accuracy and efficiency

A significant aspect of working with these models involves the invocation of external tools to enhance their problem-solving capabilities. However, as the number of tools accessible to an LLM increases, the efficiency and effectiveness of these tool invocations tend to diminish. To address this issue, it is critical to develop a robust method for prefiltering tools based on the specific requirements of a query or question.

LLMs are inherently equipped with flat tool invocation capabilities. This means that they can be supplied with a limited set of tools to assist in fulfilling a task. However, as the array of available tools expands, the efficiency of the model's capacity to appropriately call upon these tools declines. This is mainly due to the overwhelming number of potential options, which can obscure the model's decision-making process and reduce its ability to select the most pertinent tools for a given task. Therefore, a prefiltering step becomes necessary to manage tool selection effectively.

Prefiltering tools is an essential process that relies on structural metadata to determine which tools should be considered by the model in relation to the question at hand. This step ensures that only relevant tools are available for use, thereby optimising the model's performance and reducing computational overhead. The prefiltering step can be executed by employing several criteria, each aimed at ensuring the model's proficiency in addressing the specific nuances of the question.

The initial criterion for filtering tools involves sorting by topic relevance. This means identifying tools that align with the subject matter of the inquiry. Questions often pertain to specific topics, and tools designed with an understanding of these subjects can offer more accurate and precise insights. For instance, if a question relates to environmental science, the prefiltering system would bypass tools irrelevant to this field and focus on those offering insights or functionality that complement environmental studies.

Another critical filtering criterion is temporal resolution. Questions may address events or data related to particular time periods, and not all tools possess the capability to handle temporal queries comprehensively. The prefiltering system must, therefore, identify and prioritise tools that are equipped to deal with the specified temporal context. For example, if a query involves historical climate data from the 1980s, the system should filter out tools that only provide contemporary data, ensuring the most temporally relevant tools are selected.

In addition to temporal relevance, the selection of tools may also hinge on determining the best source of information, especially in scenarios with overlapping data coverage. This means the system should favour tools that offer the most precise or contextually appropriate data. For example, if a model for Norway offers superior data specificity compared to a broader European model for a particular question, the prefiltering mechanism should select the Norwegian model. This ensures that the data used is both relevant and of the highest quality possible for the query.

Spatial resolution is another determinant in the prefiltering process. Many questions necessitate specific spatial data resolutions, and tools may vary greatly in the granularity of their data. It is essential to filter tools based on the spatial requirements specified in the question. If a question demands high-resolution spatial data, such as a 1-meter resolution for detailed topographic analysis, the prefiltering mechanism should prioritise those tools capable of fulfilling such needs. Conversely, for less detailed queries, tools offering broader, 100-meter resolution data may suffice.

The implementation of a structured prefiltering step enhances the ability of LLMs to select and utilise tools effectively, leading to more accurate and efficient outcomes. By categorising tools based on topic, temporal and spatial capabilities, and data quality, the system reduces the cognitive load on the model and ensures that only the most relevant tools are engaged. This not only improves the model's response time and accuracy but also conserves computational resources.

Agent implementation using Langchain for Python

The integration and orchestration of calls to various machine learning models constitute a key component in the efficient deployment and utilisation of these models in real-world applications. Among the technologies facilitating this process, Langchain emerges as a pivotal tool. It offers an interface designed specifically for dynamic interaction with language models, significantly enhancing applied solutions' flexibility and maintainability.

Langchain serves as an abstraction layer that simplifies the setup of client calls to large language models (LLMs). Through its design, Langchain provides a mechanism by

which these models can be interacted with seamlessly, ensuring that the underlying complexities of each model are effectively managed. The abstraction layer implemented by Langchain allows developers to perform model orchestration in a manner that separates the business logic from model-specific code. This, in turn, supports the ease of integration and modular development, facilitating a more structured approach to software development that incorporates these advanced models.

Further, the abstraction layer not only supports seamless communication with a single LLM but also allows for the straightforward replacement of one model with another. This feature is essential in today's rapidly evolving machine-learning landscape, where the rapid development of new models can often leave practitioners with outdated tools. By using Langchain, developers can efficiently swap out LLMs at any workflow stage to achieve improved performance, accuracy, or other desirable traits inherent in more advanced models.

The utility of this abstraction layer extends beyond the wholesale replacement of models. It also permits granular control over individual steps within the model workflow. This means that each component of the process can be optimised by integrating the most suitable model for that specific task. Such granular control is invaluable, as it provides the opportunity to tailor each step of the interaction with the LLM to align with specific requirements or constraints. Consequently, applications deploying these models become more robust and targeted in their outcomes.

Langchain's design enhances the portability and scalability of machine learning systems, allowing for simple expansion in scope and scale. Abstracting model interactions enables systems to cater to increased demand or incorporate complex functionalities without significantly reengineering existing systems. The ability to efficiently manage and organise various LLMs through Langchain allows developers to create more sophisticated applications that can handle a wider range of tasks and challenges.

Moreover, Langchain plays a significant role in encouraging best practices in software engineering and model deployment. Enforcing separation of concerns ensures that the model logic does not hinder the development of other software components. Developers are encouraged to engage in clean code practices and modular development, which enhances the maintainability and longevity of software systems. Using an abstraction layer thus aids in creating a more streamlined development process, where different teams can work on separate components without interference, leading to a more collaborative and efficient workflow.

In addition to improving development processes, Langchain may also contribute to the reduction of deployment risks. It allows testing new models within existing infrastructures without extensive overhauls, facilitating smoother transitions and updates. This capacity to swap models quickly and without significant downtime is especially critical in environments where continuous integration and deployment are essential requirements.

Fostering an ecosystem where evolving models can be integrated without friction is fundamental to maintaining a competitive edge in fields relying heavily on machine learning. Langchain thus acts as a linchpin in the operational management of these models, aligning technological capabilities with business objectives. Organisations are thereby empowered to harness the full potential of LLM advancements with minimal disruption to their existing systems.

Value engineering in the back-end solution

Back-end value engineering is a crucial aspect in the development of computational tools, particularly tools like JackDaw that involve complex model interactions. This discipline focuses on the optimal allocation of computational resources to enhance performance and reduce operational costs. Within the context of back-end value engineering, a nuanced approach to model selection and implementation can lead to substantial improvements in efficiency and effectiveness.

At its core, the process of back-end value engineering involves the strategic use of different models for varying steps within a computational sequence. By tailoring the choice of model to the specific requirements of each task, developers can achieve better resource allocation, ultimately enhancing the tool's performance.

For initial tasks requiring basic language processing capabilities, models possessing natural language processing (NLP) capacity are sufficient. Such tasks often involve identifying appropriate arguments to be passed to a function. NLP models excel in understanding linguistic input, making them ideal for these preliminary steps. Their use ensures that computational resources are not expended unnecessarily on more complex models when simpler ones suffice.

In contrast, certain steps necessitate the employment of a base model with a comprehensive understanding of diverse knowledge areas. Models like GPT-4po serve this purpose effectively. These models are designed to possess a wide-ranging repository of generic knowledge, thereby enabling them to tackle tasks that demand a broad understanding. Their utility lies in their ability to provide insights and execute functions that require a generalist perspective.

Moreover, some stages in a computational process necessitate the application of bespoke models with domain-specific expertise. An example might be a model engineered to handle species tolerance parameters. These custom models are tailored to address specialised tasks, offering insights and solutions that generic models might overlook. Their usage is particularly advantageous in niche

applications where precision and specialised knowledge are critical.

By implementing an optimised model selection strategy, the overall cost associated with complex tool-calling chains can be significantly mitigated. This approach entails identifying the most suitable model for any given task and deploying it accordingly. Such strategic allocation not only reduces computational expenses but also enhances the tool's efficacy in performing its designated functions.

The process of determining the optimal model for each step involves several considerations. These include the computational requirements of the task, the desired level of accuracy, and the potential impact on overall performance. By weighing these factors, developers can make informed choices that align with both technical objectives and budgetary constraints.

There is a crucial balance to be struck between leveraging the simplicity and efficiency of NLP models and harnessing the expansive capabilities of base models like GPT-4po. Additionally, the precision offered by customised models with specialised knowledge must be integrated judiciously to maximise effectiveness. Such decisions require careful assessment and strategic planning to ensure that the right model is deployed at the right stage.

In scenarios where tasks are relatively straightforward, deploying simpler models can free up computational resources, thereby allowing more complex models to be reserved for tasks that truly necessitate their capabilities. Conversely, when tasks demand a more in-depth understanding or specialised insight, more sophisticated models are utilised to achieve the desired outcomes.

Furthermore, this tiered approach to model selection also facilitates scalability. As computational demands evolve, new models can be integrated, or existing models can be updated to reflect the latest advancements and requirements. This adaptability ensures the system remains robust and responsive to changing needs without requiring complete overhauls or excessive expenditure.

Integrating diverse models in back-end processes also fosters innovation, as developers are encouraged to explore novel combinations and applications to enhance performance. By considering how different models can interact and complement one another, new solutions can be devised that push the boundaries of current capabilities, all while keeping resource use and costs in check.

Through careful analysis and planning, back-end value engineering plays an instrumental role in the development and deployment of computational tools. By strategically selecting and implementing models tailored to specific tasks, developers can transform their systems into more efficient and cost-effective solutions. This calculated approach not only optimises performance but also paves the way for ongoing innovation and refinement.

While of lesser scientific value, value engineering is critical to ensuring adoption and a realistic path to a real and tangible market.

RAG module

One of the key components that enables JackDaw to deliver more credible and less hallucination-prone responses is the Retrieval Augmented Generation (RAG) module, illustrated in Figure 3. This module enhances JackDaw's responses by grounding them in external references and supporting the ingestion of new data sources. The RAG module consists of three main components:

1. **Embedding Model** - Transforms input text into high-dimensional vector embeddings, enabling semantic search within a vector database.
2. **PDF-to-Markdown Conversion Tool** - Parses user-submitted PDF files and extracts both content and structure. It converts the documents into clean, structured Markdown, preserving key elements such as headings, paragraphs, tables, and lists to maintain readability and information integrity.
3. **RAG Application** - The core service that handles user queries, retrieves relevant content from the vector database, and compiles the results for response generation.

The module uses a microservices architecture to support scalability and simplify maintenance. At its core is the RAG Application, which orchestrates the query processing and retrieval workflow.

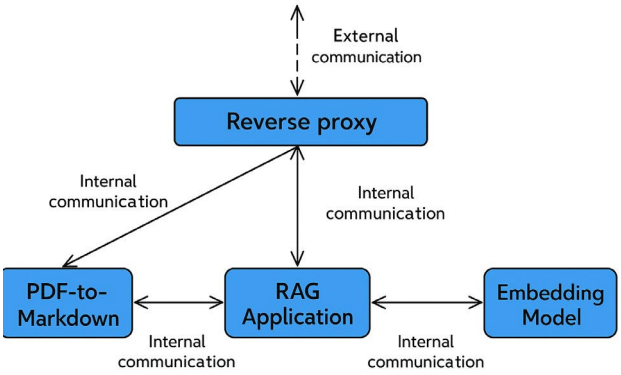


Fig 3. Overview of the RAG module architecture

The RAG module interfaces with the JackDaw system through a reverse proxy, facilitating secure external communication. Internally, the reverse proxy routes requests to two main components: the RAG Application and the PDF-to-Markdown Conversion tool. The RAG Application acts as the central component, orchestrating interactions with both the PDF-to-Markdown Conversion tool and the Embedding Model. All communication between these components occurs internally, ensuring efficient data flow within the system.

Before deployment, the system was seeded with trusted content - policy documents and scientific articles. These documents were processed using the same PDF-to-Markdown tool described above. This tool utilized Docling Library [29] to perform the parsing. The resulting Markdown files were segmented into smaller units of approximately 500 tokens, referred to as chunks.

Each chunk was embedded using the **Alibaba-NLP/gte-Qwen2-1.5B-instruct Model** [30], a high-performing multilingual language model. The resulting embeddings, along with their corresponding text chunks, were stored in **ChromaDB**, an open-source vector database optimized for similarity search.

At runtime, when a user submits a query, the RAG Application first generates an embedding of the query. This embedding is then used to retrieve semantically similar chunks from the vector database. These passages are returned to JackDaw along with metadata that includes source information, enabling more accurate responses and full traceability. This approach ensures that users can access the original documents from which answers are derived, promoting transparency and trust.

Currently, the PDF-to-Markdown tool is limited to processing individual PDF files and returning the parsed content directly to JackDaw. This setup is designed for scenarios where users want assistance understanding or extracting information from specific documents.

In the future, the tool will be expanded to support the creation of personalized knowledge bases. Users will be able to upload multiple PDF files, effectively building a private vector database similar to the existing global one. This personal database can then be used to enrich the responses generated by the language model, resulting in more context-aware, tailored, and useful interactions. Moreover, this setup will support the continuous growth of the data source, enabling users to expand and evolve their assistant's knowledge over time incrementally.

IV. Results

We tested the performance of JackDaw on various scenarios related to geospatial data tasks in agriculture and forestry and compared its solutions with standard proprietary LLMs to determine its benefits and limitations. For each user's query, our system invokes tools with information about the geographical area of interest and reasons over the retrieved content to provide a final answer. To make the comparison more balanced, we include the geographical area in the form of bounding box coordinates in the input query so that the standard LLM has some information about the location. Without this information, the standard LLM would have no way of providing relevant information. The following scenarios showcase that JackDaw is able to provide location-specific relevant information, whereas the standard LLM can only give general advice, sometimes not even related to the area of interest. We used GPT-4o-mini as the LLM.

Scenario 1

In this scenario, we asked about a partially forested area; the user query was: "Is this area forested, and are there any specific utilities regarding forestry?". Figure 4 shows responses from JackDaw and a standard LLM. It also contains the selected area of interest. GPT-4o-mini managed to offer general advice and derived the location from the coordinates as close to Pilsen, which is true, but it was unable to determine the real location. JackDaw retrieved data by utilizing tools and was able to list existing facilities related to forestry and also determine the distribution of land in the area.

Scenario 2

In this scenario, we asked whether an area in the Czech Republic is suitable for vineyards. The user query was: "Determine if this area is suitable for a vineyard and provide a concise answer." Figure 5 shows both responses together with the area of interest. This time, JackDaw provided a concise answer consisting of climate, land cover, and elevation analysis. The standard LLM wrongly determined the area to be in Austria, with a general recommendation for the area for vineyards without any backing data.

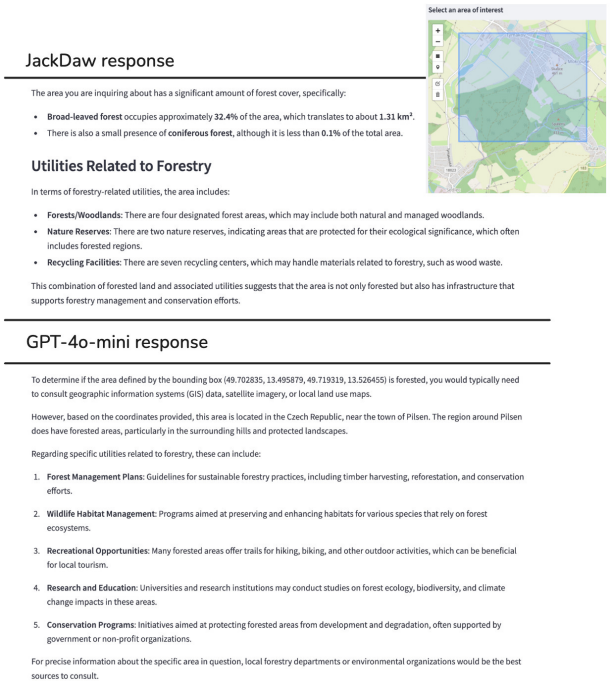


Fig 4. Comparison of responses regarding a query about forestry

Context and Visualisation Imperatives

In the years leading up to generative AI becoming "public property", the volume of text emerging from argumentative sources such as consultancy reports was already expanding noticeably. The increase was often

characterised by esoteric industry language and overcomplicated sentence structures, far beyond what was necessary for effective communication. Reports usually offered vague reasoning and inconclusive advice. Consequently, readers, typically non-scientific, technically oriented professionals had to develop new approaches to extract key ideas or recommendations from these wordy passages, which again led to altered reading habits and gave root to the scientifically unjustified and yet not entirely unfounded maxim that “nobody reads”.

Following the advent of the Language Model (LLM), the creation of well-structured text is now within reach of literally anyone, leading to a further accelerated growth in well-padded textual content, resulting in a situation where now not only “nobody reads” but neither does anyone write. While this is anecdotal and should not be considered a scientific finding, it does describe a situation where visualisation remains as important as ever to reveal and convey trends and exceptions in data in a manner that makes it more easily accessible to readers who may find a good-looking block of text deceptively attractive without taking into consideration what it actually says.

Effective visualisation in this context means applying the building blocks derived from the more traditional analytical tool suites, such as geoportals (maps, 2D, 3D) and business intelligence (charts, tables), as well as more advanced forms of visualisation.

LLMs generate text and often use the "markdown" language to include formatting. This includes the capability of outputting code blocks that can contain structured data intermixed with the text. A code block may contain any text-based or text-representable format, such as GeoJSON or GeoCSV, making rendering LLM output as any of the above visual indicators easy. JackDaw utilises custom markdown renderers to accomplish render tables, charts, and maps.

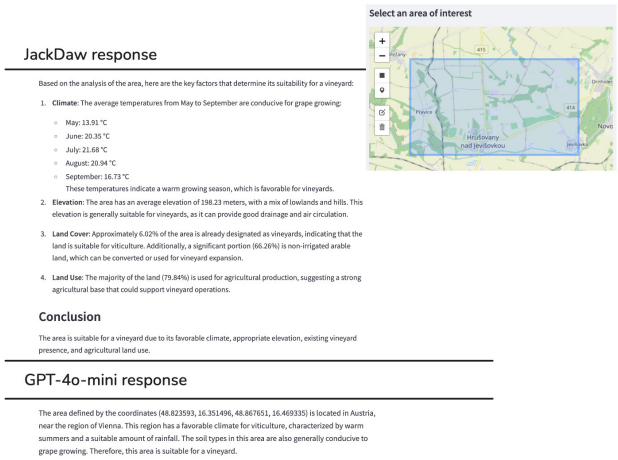


Fig 5. Comparison of responses to an area analysis query related to vineyards

V. Conclusion

This paper has demonstrated that a language-model agent can be made spatially competent by constraining its reasoning to a query-specific subset of geospatial tools and by grounding each answer in an external retrieval layer. The proposed JackDaw architecture contributes three principal advances:

1. Structured tool pre-filtering. Tools are surfaced to the LLM only when their topical, temporal and spatial metadata intersect the user’s query, avoiding the accuracy and latency penalties that arise from flat tool lists.

2. Value-engineered model orchestration. A LangChain abstraction allocates lightweight NLP models to syntactic steps, reserves foundation models for broad-context reasoning, and inserts domain-specific kernels where specialised knowledge is required.

3. Retrieval-augmented generation with transparent provenance. A micro-service RAG layer embeds trusted documents in a vector store and returns source passages with every response, enabling fact-checking and reducing hallucinations.

Benchmark experiments on forestry-asset discovery and vineyard-site assessment show that JackDaw delivers concise, location-specific answers, whereas a baseline proprietary LLM produces only generic or spatially misattributed guidance. These findings confirm that coupling global language understanding with local spatial intelligence materially improves decision support in agriculture and rural planning.

Limitations. The evaluation covered two use-cases within Central Europe and a single growing season; cross-regional transferability, additional thematic domains and longer temporal windows remain untested. The current RAG pipeline ingests individual PDF documents but does not yet support user-specific knowledge bases or large-scale cloud retrieval, and cost modelling excluded vector-query overheads.

Future work. Planned extensions include (i) adaptive ontology-driven tool selection to further shrink the action space, (ii) ingestion of multimodal sensor streams (e.g., hyperspectral cubes, IoT telemetry), (iii) private vector databases for institutional users, and (iv) benchmarking against emerging geospatial-LLM test suites. Governance requirements—explainability, bias auditing and energy accounting—will be addressed to ensure regulatory compliance and sustainable operation.

In summary, the results substantiate that an architecture which bridges foundation-scale language models with dynamically selected, high-resolution spatial data streams is a viable path toward reliable, context-aware AI services for agriculture and rural development.

Acknowledgment

Work is also supported by the EU project PoliRuralPlus – Fostering Sustainable, Balanced, Equitable, Place-based and Inclusive Development of Rural-Urban Communities'

Using Specific Spatial Enhanced Attractiveness Mapping ToolBox (Grant agreement ID: 101136910).

References

[1] Ryan, M., Isakhanyan, G., & Tekinerdogan, B. (2023). An interdisciplinary approach to artificial intelligence in agriculture. NJAS: Impact in Agricultural and Life Sciences, 95(1), 2168568, https://doi.org/10.1080/27685241.2023.2168568

[2] Ruiz-Real, J. L., Uribe-Toril, J., Torres Arriaza, J. A., & de Pablo Valenciano, J. (2020). A look at the past, present, and future research trends of artificial intelligence in agriculture. Agronomy, 10(11), 1839, https://doi.org/10.3390/agronomy10111839

[3] G. Cisotto, D. D. Tegegn, A. Zancanaro, I. Reguzzoni, E. Lotti, S. L. Manzoni, and I. F. Zoppis, "An AI-Empowered Energy-Efficient Portable NIRS Solution for Precision Agriculture: A Pilot Study on a Citrus Fruit," Proc. 19th Conf. Computer Science and Intelligence Systems, ACSIS, vol. 39, pp. 313–318, 2024, DOI: 10.15439/2024F5111

[4] Bolfe, É. L., Parreiras, T. C., Sano, E. E., Bettiol, G. M., & Vicente, L. E. (2023). Mapping agricultural intensification in the Brazilian Savanna: A machine-learning approach using harmonised Landsat–Sentinel-2 data. ISPRS International Journal of Geo-Information, 12(7), 263, https://doi.org/10.3390/ijgi12070263

[5] Dakir, A., Barramou, F., & Bachir Alami, O. (2023). Towards a Machine Learning-based Model for Automated Crop Type Mapping. International Journal of Advanced Computer Science and Applications (IJACSA), 14(1). https://doi.org/10.14569/IJACSA.2023.0140185

[6] P. Bertellini, G. D'Addese, G. Franchini, S. Parisi, C. Scribano, D. Zanirato, and M. Bertogna, "Binary Classification of Agricultural Crops Using Sentinel Satellite Data and Machine Learning Techniques," Proc. 18th Conf. Computer Science and Intelligence Systems, ACSIS, vol. 35, pp. 859–864, 2023, DOI: 10.15439/2023F1703

[7] Dhillon, M. S., Dahms, T. C., Kübert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I., & Ullmann, T. (2023). Integrating random forest and crop modelling improves the yield prediction of winter wheat and oil-seed rape. Frontiers in Remote Sensing, 3, 1010978. https://doi.org/10.3389/frsen.2022.1010978

[8] Gumma, M. K., Nukala, R. M., Panjala, P., Bellam, P. K., Gajjala, S., Dubey, S. K., & Deevi, K. C. (2024). Optimizing Crop Yield Estimation through Geospatial Technology: A Comparative Analysis of a Semi-Physical Model, Crop Simulation, and Machine Learning Algorithms. *AgriEngineering*, 6(1), 786–802, https://doi.org/10.3390/agriengineering6010045

[9] Qi, H., Wang, L., Zhu, H., Gani, A., & Gong, C. (2023). The barren plateaus of quantum neural networks: review, taxonomy and trends. *Quantum Information Processing*, 22(12), 435. https://doi.org/10.1007/s11128-023-04188-7

[10] Farmonov, N., Esmaeili, M., Abbasi-Moghadam, D., Sharifi, A., Amankulova, K., & Mucsi, L. (2024). HypsLiDNet: 3D-2D CNN model and spatial-spectral morphological attention for crop classification with DESIS and LiDAR data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. https://ieeexplore.ieee.org/document/10571369, DOI: 10.1109/JSTARS.2024.3418854

[11] L. Parigi, G. Candiani, I. Gallo, P. Toscano, and M. Boschetti, "Towards Crop Traits Estimation from Hyperspectral Data: Evaluation of Neural Network Models Trained with Real Multi-Site Data or Synthetic RTM Simulations," Proc. 19th Conf. Computer Science and Intelligence Systems, ACSIS, vol. 39, pp. 475–484, 2024, DOI: 10.15439/2024F4108

[12] Miranda, M., Pathak, D., Nuske, M., & Dengel, A. (2024, July). Multi-modal fusion methods with local neighborhood information for crop yield prediction at field and subfield levels. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium* (pp. 4307–4311). IEEE. https://doi.org/10.1109/IGARSS53475.2024.10640993

[13] Pavlova M.A., Sidorchuk D.S., Bocharov D.A. Classification of crops by NDVI time series of reduced dimensionality // Sensornye sistemy. - 2023. - Vol. 37. - N. 2. - P. 306-312. DOI: 10.31857/S023500922302004X

[14] G. Castellano, P. De Marinis, and G. Vessio, "Applying Knowledge Distillation to Improve Weed Mapping With Drones," Proc. 18th Conf. Computer Science and Intelligence Systems, ACSIS, vol. 35, pp. 393–400, 2023, DOI: 10.15439/2023F960.

[15] Toro, A. P., Bueno, I. T., Werner, J. P., Antunes, J. F., Lamparelli, R. A., Coutinho, A. C., ... & Figueiredo, G. K. (2023). SAR and optical data applied to the early-season mapping of integrated crop-livestock systems using deep and machine learning algorithms. Remote Sensing, 15(4), 1130. DOI: https://doi.org/10.3390/rs15041130

[16] R. Maskeliunas and R. Damaševičius, "d'Alembert Convolution for Enhanced Spatio-Temporal Analysis of Forest Ecosystems," Proc. 19th Conf. Computer Science and Intelligence Systems, ACSIS, vol. 39, pp. 141–148, 2024, DOI: 10.15439/2024F4228.

[17] Pihrt, J., Šimánek, P., Kovalenko, A., Kvapil, J., & Charvát, K. (2024). AI-Based Spatiotemporal Crop Monitoring by Cloud Removal in Satellite Images. In Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Annals of Computer Science and Information Systems, Volume 39, pp. 485–492, DOI: 10.15439/2024F5446

[18] F. Koutenský, J. Pihrt, M. Čepek, V. Rybář, P. Šimánek, M. Kepka, K. Jedlička, and K. Charvát, "Combining Local and Global Weather Data to Improve Forecast Accuracy for Agriculture," in *Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024, pp. 77–82. DOI:10.15439/2024F5990

[19] Filippi, P., Jones, E., Bishop, T., Acharige, N., Dewage, S., Johnson, L., Ugbaje, S., Jephcott, T., Paterson, S., & Whelan, B. (2017, October). A Big Data Approach to Predicting Crop Yield. In *Proceedings of the 7th Asian-Australasian Conference on Precision Agriculture* (pp. 16–18). Hamilton, New Zealand. Zenodo. https://doi.org/10.5281/zenodo.893668

[20] M. Hosseini, I. Becker-Reshef, R. Sahajpal, L. Fontana, P. Lafluf, G. Leale, and C. Justice, "Crop Yield Prediction Using Integration of Polarimetric Synthetic Aperture Radar and Optical Data," in *Proceedings of the 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, 2020, pp. 17–20. DOI: 10.1109/InGARSS48198.2020.9358978

[21] Wang, S., Hu, T., Huang, X., Li, Y., Zhang, C., Ning, H., Zhu, R., Li, Z., & Ye, X. (2024). GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review. *International Journal of Digital Earth*, 17(1), Article 2353122.LinkedIn+2University of Bristol+2University of Bristol+2, DOI: 10.1080/17538947.2024.2353122

[22] Sapkota, R., Qureshi, R., Hassan, S. Z., Shutske, J., Shoman, M., & Sajjad, M. (2024). *Multi-Modal LLMs in Agriculture: A Comprehensive Review*. TechRxiv Preprint. DOI: 10.36227/techrxiv.172651082.24507804/v1

[23] Mansourian, A., & Oucheikh, R. (2024). *ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models*. ISPRS International Journal of Geo-Information, 13(10), 348.Lund University, DOI: 10.3390/ijgi13100348

[24] L. Yuan, F. Mo, K. Huang, W. Wang, W. Zhai, X. Zhu, Y. Li, J. Xu, and J.-Y. Nie, "OmniGeo: Towards a Multimodal Large Language Models for Geospatial Artificial Intelligence," *arXiv preprint arXiv:2503.16326*, Mar. 2025. https://arxiv.org/html/2503.16326v1

[25] L. Wang, J. Chou, A. Tien, X. Zhou, and D. M. Baumgartner, "AviationGPT: A Large Language Model for the Aviation Domain," in *Proceedings of the AIAA AVIATION Forum and ASCEND 2024*, 2024, p. 4250. DOI: 10.2514/6.2024-4250

[26] Ali. S. H, Shahid. M. F, Tanveer. M. H, Rauf. A, "Integrating LLM for Cotton SoilAnalysis in Smart Agriculture System", IJIST, Special Issue pp 283-294, (2024) *(PDF) Integrating LLM for Cotton Soil Analysis in Smart Agriculture System*. Available from: https://www.researchgate.net/publication/389436000_Integrating_LLM_for_Cotton_Soil_Analysis_in_Smart_Agriculture_System [accessed May 23 2025]

[27] Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*. https://doi.org/10.48550/arXiv.2304.03277

[28] Q. Huang, Y. Tao, S. Ding, Y. Liu, and F. Marinello, "Automatic Construction of Knowledge Graph of Tea Diseases and Pests," Communication Papers of the 18th Conf. Computer Science and

Intelligence Systems, ACSIS, vol. 37, pp. 141–146, 2023, DOI: 10.15439/2023F6100

[29] Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Berrospi Ramis, C., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., Teixeira de Lima, R., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., & Staar, P. W. J. (2024). *Docling Technical Report* (Version 1.0.0). arXiv preprint arXiv:2408.09869. https://doi.org/10.48550/arXiv.2408.09869

[30] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*. https://doi.org/10.48550/arXiv.2308.03281

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 519–530
DOI: 10.15439/2025F8874
ISSN 2300-5963 ACSIS, Vol. 43

While-guard Synthesis by Abstract Static Analysis and CHC Solving

Aleksandar S. Dimovski
0000-0002-3601-2631
Mother Teresa University
Skopje, North Macedonia
aleksandar.dimovski@unt.edu.mk

***Abstract*—This paper introduces a novel technique for automatically synthesizing assertion-safe `while`-guards in imperative programs. Given a partial program (sketch) with missing `while`-guards, the proposed algorithm synthesizes complete Boolean expressions for the missing ones, such that the obtained complete program satisfies the given assertions. To solve this problem, our technique uses forward and backward abstract static analyses of programs to generate (logical) constraints with unknown predicates that are subsequently solved by employing the logical Constraint Horn Clause (CHC) solvers.**

We have implemented our synthesis algorithm in a proof-of-concept tool, and evaluated it on a set of C programs with missing `while`-guards. By experiments we prove the effectiveness of the proposed technique for synthesizing arbitrary Boolean expressions defined over program variables for some interesting program sketches written in C.

***Index Terms*—Program Synthesis, Abstract Static Analysis, Constraint Horn Clause (CHC) solving.**

I. Introduction

PROGRAM synthesis [1] is a process of automatically constructing a program from a given specification. Program sketches [2], [3] are partial programs with missing (arithmetic and Boolean) expressions called *holes* (`??`). They simplify the synthesis problem by limiting the search to programs of particular form. The earliest approaches to program sketching [2], [3], [4] can automatically synthesize only integer "constants" for the holes. In particular, given a sketch and an assertion, the synthesizer (e.g., `Sketcher` [2], [3]) automatically finds integer constants for the holes such that the synthesized complete program satisfies the assertion. Subsequently, the generalized program sketcher [5] is introduced, which uses abstract interpretation [6], [7], [8] and logical abduction [9] to infer arbitrary expressions, rather than only integer constants, for the holes in program sketches. The missing holes occur either as guards (Boolean expressions) in `if`-s and `while`-s or as right-hand sides (arithmetic expressions) in assignments. However, in the case of missing `while`-guards, this approach [5] is cumbersome due to the difficulty to infer precise inductive invariants for `while`-s, and so it often fails to generate satisfactory solutions.

In this paper, we explore the problem of inferring missing `while`-guards in program sketches by introducing a novel algorithm based on abstract static analysis of programs [6], [7], [8] and logical Constraint Horn Clause (CHC) solving [10], [11]. In particular, our synthesis algorithm consists of two phases: *constraint generation* and *constraint solving*. In *the first phase*, we employ forward and backward abstract static analyses of programs, whose aim is to determine dynamic program properties in a fully automatic manner. The forward abstract analysis [6], [7], which is similar to strongest postcondition inference [12], works by propagating the initial facts towards the final locations, thus computing the invariant postconditions before the holes. The backward abstract analysis [6], [7], which is similar to weakest precondition inference [12], works by propagating the final (assertion) facts towards the initial locations, thus computing the sufficient preconditions that ensure assertion validity after the holes. The numerical abstract domains, such as Intervals [6], Octagons [13], Polyhedra [14], are widely used in practice for static analyses to infer information about the possible values of program variables together with the possible relations between them. Thus, they are employed for computing invariant postconditions and sufficient preconditions in the form of linear integer constraints over program variables. Subsequently, in *the second phase*, we employ the logical constraints generated in the first phase by abstract static analyses as well as the logical CHC solving technique to synthesize the missing Boolean expressions. That is, the `while`-guard synthesis problem is encoded as a set of CHCs, which represent a class of first-order logical formulas modulo some background theories, like linear integer arithmetic (LIA). More specifically, a CHC is a logical implication formula containing unknown predicates. This way, all holes in the sketch as well as inductive invariants for the `while`-s are represented by unknown predicates in the CHCs. A solution to a set of CHCs assigns a logical formula to each unknown predicate in such a way that all CHCs become valid formulas, and moreover their validity guarantees the correctness of the resulting complete program. By solving the obtained CHCs with off-the-shelf CHC solvers, such as `Eldarica` [11], `Spacer` [10], `Duality` [15], and `FreqHorn` [16], we can synthesize complete Boolean expressions for the missing holes `??` in the program sketch, which will guarantee the assertion validity of the resulting complete program.

We have implemented the proposed algorithm in a proof-of-concept tool, called $\texttt{WhileSketching}_{\texttt{CHC}}$. For the constraint generation phase, we use the numerical abstract domains

```
void main() {
① int x = 10, y = 0;
② while ⓗ (??) do {
③   x = x-1;
④   y = y+1;}
⑤ assert (y==10);
//assert (y>x);
//assert (y==x);
//assert (x=10); }
//assert (x>15); }
```

Fig. 1. `while.c`.

(e.g., Polyhedra [14]) from the APRON library [17], while for the constraint solving phase, we use the `Eldarica` [11] and `Z3` [18] tools for CHC- and SMT-solving. We perform an experimental evaluation on a selected set of benchmarks written in C, which are taken from the SV-COMP (https://sv-comp.sosy-lab.org/) and the literature [19], [5]. The experiments demonstrate the practicality of our technique to successfully synthesize assertion-safe `while`-guards in some interesting C sketches.

This work brings the following contributions:

(1) We present the `while`-guard synthesis problem for automatically synthesizing arbitrary complete Boolean expressions for missing holes in sketches;
(2) We introduce a novel algorithm for solving this problem by interaction between the abstract static analysis and the logical CHC solving;
(3) We implement our algorithm and evaluate it on a set of interesting program sketches written in C.

This work is organized in the following way. Section II presents a small example illustrating how our algorithm infers missing `while`-guards. Section III defines a small imperative language that is used for the formal development as well as its concrete and abstract semantics. Moreover, the CHC problem is also defined here. Section IV presents the main `while`-guard synthesis algorithm. Section V briefly describes another existing approach for resolving program sketches based on logical abduction. Section VI presents an experimental evaluation of the introduced approach, comparing our tool with a program sketching tool based on logical abduction. Finally, Section VII and VIII discuss related work, summarize the introduced approach, and highlight future research direction.

II. Illustration

We now illustrate our approach through one motivating example taken from [5]. The program sketch given in Fig. 1, called `while.c`, contains one unknown hole `??` in the `while`-guard. Our aim is to replace the unknown hole at loc. ⓗ with a Boolean expression, such that the assertion at loc ⑤ is valid.

Our approach consists of two phases: constraint generation and constraint solving. In the first phase, our technique starts by performing forward and backward abstract static analyses based on the Polyhedra abstract domain [14]. The over-approximating forward static analysis starts with the initial fact $\top$ (i.e., true) at loc. ①, and by propagating it forwards it infers the invariants $(\texttt{x=10} \wedge \texttt{y=0})$ at loc. ② before the `while` as well as $(\texttt{x} \leq 10 \wedge \texttt{x+y=10})$ at loc. ⓗ. The under-approximating backward static analysis starts with the assertion fact $(\texttt{y=10})$ at loc. ⑤, and by propagating it backwards it infers the preconditions $(\texttt{y=10})$ at loc. ⑤ after the `while` as well as $\bot$ (i.e., false) at loc. ⓗ. The loc. ⓗ is reached upon entering the `while`-loop and after each `while`-iteration, so the facts inferred by both analyses at loc. ⓗ correspond to the *inductive `while`-invariant*, which expresses a relationship between program variables that is preserved by any execution of the `while`'s body [12]. However, since both analyses treat the unknown hole `??` as non-deterministic choice $\{\text{true}, \text{false}\}$, the forward analysis infers the over-approximation of the inductive `while`-invariant, whereas the backward analysis infers the under-approximation of the inductive `while`-invariant. Thus, we use the inferred facts by forward and backward static analyses to generate the following implication constraints, called Constrained Horn Clauses (CHCs):

$$(\texttt{x=10} \wedge \texttt{y=0}) \implies Inv(\texttt{x},\texttt{y}) \quad (1)$$

$$Inv(\texttt{x},\texttt{y}) \wedge R(\texttt{x},\texttt{y}) \wedge \texttt{x'=x-1} \wedge \texttt{y'=y+1} \implies Inv(\texttt{x'},\texttt{y'}) \quad (2)$$

$$Inv(\texttt{x},\texttt{y}) \wedge \neg R(\texttt{x},\texttt{y}) \implies (\texttt{y=10}) \quad (3)$$

$$Inv(\texttt{x},\texttt{y}) \implies (\texttt{x} \leq 10 \wedge \texttt{x+y=10}) \quad (4)$$

$$\text{false} \implies Inv(\texttt{x},\texttt{y}) \quad (5)$$

$Inv(\texttt{x},\texttt{y})$ encodes an unknown predicate for the inductive `while`-invariant, and $R(\texttt{x},\texttt{y})$ encodes an unknown predicate for the missing `while`-guard `??`. Both unknown predicates $Inv(\texttt{x},\texttt{y})$ and $R(\texttt{x},\texttt{y})$ are defined over program variables `x` and `y` that are in the scope of the `while`. The implications (1) and (2) state that the `while`-invariant $Inv(\texttt{x},\texttt{y})$ is inductive. That is, (1) states that $Inv(\texttt{x},\texttt{y})$ is implied by the `while`'s precondition at loc. ②, whereas (2) states that $Inv(\texttt{x},\texttt{y})$ is preserved in each iteration of the `while`'s body, i.e. if $Inv(\texttt{x},\texttt{y})$ and the `while`-guard $R(\texttt{x},\texttt{y})$ are true and we execute the `while`'s body then $Inv(\texttt{x'},\texttt{y'})$ will be true after it, where `x',y'` are updated values of `x,y` after executing the `while`'s body. The implication (3) stipulates that when the `while` terminates, and so $Inv(\texttt{x},\texttt{y})$ and $\neg R(\texttt{x},\texttt{y})$ are true, then they are strong enough to imply the `while`'s postcondition at loc. ⑤ ensuring that the assertion must be valid. The implication (4) (resp., (5)) states that the analysis fact inferred at loc. ⓗ by forward (resp., backward) analysis is over- (resp., under-) approximation of $Inv(\texttt{x},\texttt{y})$.

In the second phase, our aim is to solve the generated CHCs by inferring interpretations (logical formulas) for $Inv(\texttt{x},\texttt{y})$ and $R(\texttt{x},\texttt{y})$ under which all CHCs (1) – (5) are valid. The CHC solving algorithm [11] reports the following solutions for the unknown predicates: $Inv(\texttt{x},\texttt{y}) \equiv (\texttt{x} \leq 10 \wedge \texttt{x+y=10})$ and $R(\texttt{x},\texttt{y}) \equiv (\texttt{x} \leq 10 \wedge \texttt{x+y=10} \wedge \texttt{x} \neq 0)$. Then, we call a SMT solver to simplify the formula $R(\texttt{x},\texttt{y}) \wedge \neg Inv(\texttt{x},\texttt{y})$, thus excluding sub-formulas of $R(\texttt{x},\texttt{y})$ that are also present in the invariant $Inv(\texttt{x},\texttt{y})$. Thus, the synthesized `while`-guard

will be only the strengthening of $R(\mathtt{x},\mathtt{y})$ with respect to $Inv(\mathtt{x},\mathtt{y})$ that is not part of the invariant $Inv(\mathtt{x},\mathtt{y})$. Hence, we fill the hole `??` at loc. ⓗ with the solution $(\mathtt{x} \neq 0)$.

In a similar vein, we can synthesize concrete `while`-guards in `while.c` (Fig. 1) with respect to different assertions. Assume that we consider $\mathtt{assert}(be)$. The CHCs (1), (2), (4), (5) stay the same as above, and we only change the right-hand side of CHC (3) that becomes: $Inv(\mathtt{x},\mathtt{y}) \wedge \neg R(\mathtt{x},\mathtt{y}) \implies (be)$. For example, for assertions $(\mathtt{y} > \mathtt{x})$, $(\mathtt{y}{=}\mathtt{x})$, $(\mathtt{x}{=}10)$, and $(\mathtt{x} > 15)$, the synthesized `while`-guards are: $(\mathtt{x} \geq 5)$, $(\mathtt{x} \neq 5)$, (false), and the problem is unrealizable (i.e., there is no solution that can be derived), respectively. Note that for $\mathtt{assert}(\mathtt{x} > 15)$, we obtain the same solutions for $Inv(\mathtt{x},\mathtt{y})$ and $R(\mathtt{x},\mathtt{y})$, that is $Inv(\mathtt{x},\mathtt{y}) \equiv R(\mathtt{x},\mathtt{y}) \equiv (\mathtt{x} \leq 10 \wedge \mathtt{x}{+}\mathtt{y}{=}10)$. For example, the implication (3): $Inv(\mathtt{x},\mathtt{y}) \wedge \neg R(\mathtt{x},\mathtt{y}) \implies (\mathtt{x} > 15)$ becomes vacuously valid in this case as the left-hand side of the above implication is (false). However, $(R(\mathtt{x},\mathtt{y}) \wedge \neg Inv(\mathtt{x},\mathtt{y})) \equiv$ false, so we conclude that there is no (non-vacuous) solution for this assertion (i.e., the problem is unrealizable). On the other hand, for $\mathtt{assert}(\mathtt{x}{=}10)$, we obtain the CHC solution $Inv(\mathtt{x},\mathtt{y}) \equiv (\mathtt{x} \leq 10 \wedge \mathtt{x}{+}\mathtt{y}{=}10)$ and $R(\mathtt{x},\mathtt{y}) \equiv$ (false), so we infer the solution (false) for `??` in this case.

III. Background

This section introduces the language we consider as well as its concrete and abstract semantics. Furthermore, we also formally define the CHC problem.

A. Language

We use a simple C-like imperative language for writing general-purpose programs. The program variables *Var* are statically allocated and the only data type is the set $\mathbb{Z}$ of mathematical integers. We use the hole constructs $\mathtt{??}_i$ to encode unknown `while`-guards. The holes $\mathtt{??}_i$ are placeholders that the synthesizer must replace with suitable Boolean expressions, such that the resulting complete program will satisfy all assertions. The syntax of the language is:

$$
\begin{array}{l}
s\,(s \in \mathit{Stm}) ::= \mathtt{skip} \mid \mathtt{x}{=}ae \mid s; s \mid \mathtt{assume}(be) \\
\qquad\qquad \mid \mathtt{if}\,(be')\ s\ \mathtt{else}\ s \mid \mathtt{while}\ l_h : (be)\ \mathtt{do}\ s \\
ae\,(ae \in \mathit{AExp}) ::= n \mid [n, n'] \mid \mathtt{x} \mid ae \oplus ae \\
be\,(be \in \mathit{BExp}) ::= \mathtt{??}_i \mid be' \\
be' ::= ae \bowtie ae \mid \neg be' \mid be' \wedge be' \mid be' \vee be'
\end{array}
$$

where n ranges over integers $\mathbb{Z}$, $\mathtt{x}$ ranges over program variables *Var*, $[n, n']$ over integer intervals, and $\oplus \in \{+, -, *, /\}$, $\bowtie \in \{<, \leq, =, \neq\}$. Integer intervals $[n, n']$ denote a random choice of an integer in the interval. We assume that statements are tagged with unique syntactic labels $l \in \mathbb{L}$. Note that `while` has an extra syntactic label $l_h \in \mathbb{L}$ at the `while`-head, i.e. between the `while` keyword and the guard. We assume, without loss of generality, that a program p is a sequence of statements followed by a single assertion "$l_i : s; l_f : \mathtt{assert}(be^f)$". We say that the program p satisfies its assertion (or the assertion is valid) if Boolean expression be^f evaluates to (true) whenever the control reaches the label l_f. Otherwise, the assertion is violated. We say that a program (statement) is *complete* if it contains no holes; otherwise, we say that it is *partial* (or *program sketch*).

$$
\begin{array}{l}
\overrightarrow{[\![\mathtt{skip}]\!]} S = S \\
\overrightarrow{[\![\mathtt{x} = ae]\!]} S = \{\sigma[\mathtt{x} \mapsto n] \mid \sigma \in S, n \in [\![ae]\!]\sigma\} \\
\overrightarrow{[\![s_1\ ;\ s_2]\!]} S = \overrightarrow{[\![s_2]\!]}(\overrightarrow{[\![s_1]\!]} S) \\
\overrightarrow{[\![\mathtt{assume}(be)]\!]} S = \{\sigma \in S \mid \mathsf{true} \in [\![be]\!]\sigma\} \\
\overrightarrow{[\![\mathtt{if}\ (be)\ \ s_1\ \mathtt{else}\ s_2]\!]} S = \overrightarrow{[\![s_1]\!]}\{\sigma \in S \mid \mathsf{true} \in [\![be]\!]\sigma\} \cup \\
\qquad\qquad\qquad\qquad\qquad\qquad \overrightarrow{[\![s_2]\!]}\{\sigma \in S \mid \mathsf{false} \in [\![be]\!]\sigma\} \\
\overrightarrow{[\![\mathtt{while}\ (be)\ \mathtt{do}\ s]\!]} S = \{\sigma \in \mathtt{lfp}\ \overrightarrow{\phi_S} \mid \mathsf{false} \in [\![be]\!]\sigma\} \\
\overrightarrow{\phi_S}(X) = S \cup \overrightarrow{[\![s]\!]}\{\sigma \in X \mid \mathsf{true} \in [\![be]\!]\sigma\}
\end{array}
$$

Fig. 2. Definitions of $\overrightarrow{[\![s]\!]} : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$.

Remark. Note that in programs p we allow assumptions ($\mathtt{assume}(be)$) to freely occur in a sequence of statements s, while the assertion ($\mathtt{assert}(be)$) occurs only at the final location l_f. An assertion $\mathtt{assert}(be)$ specifies that every execution reaching it is expected to evaluate be to true, in which case we say that the assertion is valid. Otherwise, if the assertion is not valid, the program terminates early with an error. On the other hand, an assumption $\mathtt{assume}(be)$ specifies that every execution reaching it with be evaluated to false is terminated. However, unlike before, this termination is not erroneous, but it means that the programmer does not want to consider the rest of this execution when verifying the correctness of the program.

Let H be a set of uniquely labelled holes $\mathtt{??}_i$ in program p. A *control function* ϕ is a mapping from the set of holes H to complete Boolean expressions be'. We say that ϕ is *complete* if $dom(\phi) = H$, i.e. ϕ is defined for each hole in the program. Otherwise, if $dom(\phi) \subset H$, i.e. ϕ is $\perp$ (undefined) for some holes, we say that ϕ is a *partial* control function. We write $p[\phi]$ to denote the program obtained by substituting each $\mathtt{??}_i$ with $\phi(\mathtt{??}_i)$, if $\phi(\mathtt{??}_i)$ is defined.

Definition 3.1: A complete control function ϕ is a *solution* to the `while`-guard synthesis problem defined by program sketch p if $p[\phi]$ is a complete program that satisfies its assertion.

B. Concrete Semantics and Analyses

The concrete semantics describes the properties of the possible executions of a program. We now define the concrete semantics of our language, and use it to construct *reachability* or *invariance* (forward) as well as *sufficient condition* (backward) *concrete analyses*.

A memory *store*, denoted $\sigma \in \Sigma$, is a function mapping each program variable to an integer value: $\Sigma = \mathit{Var} \to \mathbb{Z}$. The concrete domain is the complete lattice of the powerset of stores [1] $\langle \mathcal{P}(\Sigma), \subseteq, \cup, \cap, \emptyset, \Sigma \rangle$. The semantics of arithmetic expressions $[\![ae]\!] : \Sigma \to \mathcal{P}(\mathbb{Z})$ and Boolean expressions $[\![be]\!] : \Sigma \to \mathcal{P}(\{\mathsf{true}, \mathsf{false}\})$ are the sets of possible (numerical

[1] The powerset of a set S, denoted $\mathcal{P}(S)$, is the set of all subsets of S.

$$
\begin{array}{l}
\overleftarrow{[\![\texttt{skip}]\!]}S = S \\
\overleftarrow{[\![\texttt{x = } ae]\!]}S = \{\sigma \mid \forall n \in [\![ae]\!]\sigma, \sigma[\texttt{x} \mapsto n] \in S\} \\
\overleftarrow{[\![s_1 \texttt{ ; } s_2]\!]}S = \overleftarrow{[\![s_1]\!]}(\overleftarrow{[\![s_2]\!]}S) \\
\overleftarrow{[\![\texttt{assume}(be)]\!]}S = S \cup \{\sigma \mid [\![be]\!]\sigma = \{\text{false}\}\} \\
\overleftarrow{[\![\texttt{if } (be) \texttt{ } s_1 \texttt{ else } s_2]\!]}S = \big(\overleftarrow{[\![s_1]\!]}S \cup \{\sigma \mid [\![be]\!]\sigma = \{\text{false}\}\}\big) \cap \\
\qquad\qquad \big(\overleftarrow{[\![s_2]\!]}S \cup \{\sigma \mid [\![\neg be]\!]\sigma = \{\text{false}\}\}\big) \\
\overleftarrow{[\![\texttt{while } (be) \texttt{ do } s]\!]}S = \texttt{gfp}\ \overleftarrow{\phi}_S \\
\overleftarrow{\phi}_S(X) = \big(S \cup \{\sigma \mid [\![\neg be]\!]\sigma = \{\text{false}\}\}\big) \cap \big(\overleftarrow{[\![s]\!]}X \cup \{\sigma \mid [\![be]\!]\sigma = \{\text{false}\}\}\big)
\end{array}
$$

Fig. 3. Definitions of $\overleftarrow{[\![s]\!]} : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$.

and Boolean) values for expressions ae and be in a given store σ. For example, $[\![n]\!]\sigma = n$ and $[\![[n, n']]\!]\sigma = \{n, \ldots, n'\}$ for arithmetic expressions n and $[n, n']$, whereas $[\![\texttt{??}_i]\!]\sigma = \{\text{true}, \text{false}\}$ for Boolean expression $\texttt{??}_i$. A semantics definition associates with each statement (program) a concrete semantics property from $\mathcal{P}(\Sigma)$ representing some characteristics of its possible executions. We consider two semantics of statements (programs): a *reachability or invariance* (forward) semantics $\overrightarrow{[\![s]\!]} : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$ that infers a set of reachable stores (invariants) from a given set of initial stores; and a *sufficient condition* (backward) semantics $\overleftarrow{[\![s]\!]} : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$ that infers a set of stores (sufficient condition) from which only stores satisfying a given postcondition are reached. The definitions of $\overrightarrow{[\![s]\!]}$ and $\overleftarrow{[\![s]\!]}$ are given in Fig. 2 and Fig. 3, respectively. The reachability semantics [6] is built forward, so each function $\overrightarrow{[\![s]\!]}$ takes as input a set of stores S before statement s and returns a set of possible stores reached after executing s from S. The sufficient condition semantics [7] is built backward, so each function $\overleftarrow{[\![s]\!]}$ takes as input a set of stores S after statement s and returns a set of possible stores before s from which only stores from S are reached after executing s. For example, given a set of stores $S \in \mathcal{P}(\Sigma)$, we have that $\overrightarrow{[\![\texttt{x = } ae]\!]}S$ [2] is the set of final stores reachable after executing $\texttt{x = } ae$, whereas $\overleftarrow{[\![\texttt{x = } ae]\!]}S$ is the set of initial stores such that any execution of $\texttt{x = } ae$ from them leads to a store in S. The semantics of a `while` statement is given in a standard fixed-point formulation [6], [7] using the least and greatest fix-point operators `lfp` and `gfp`, where the fixed-point functionals $\overrightarrow{\phi}_S, \overleftarrow{\phi}_S : \mathcal{P}(\Sigma) \to \mathcal{P}(\Sigma)$, parameterized with the input set of stores S, accumulate possible stores after another `while` iteration from a set of stores X going in a forward and backward direction, respectively.

C. Abstract Semantics

The language we consider is Turing complete, thus making the concrete semantics and analyses uncomputable. We now define computable abstract analyses [6], [7], [8] that are approximations of the concrete semantics and analyses. The first basic choice in abstract interpretation is to design an abstract semantics domain $\mathbb{D}$, which is an approximate version of the concrete semantics domain $\mathcal{P}(\Sigma)$. More specifically, we replace the computation in the concrete domain $\mathcal{P}(\Sigma)$ with a computation in some numerical abstract domain $\mathbb{D}$ that reasons on the numerical properties of variables, such that there exists a concretization-based abstraction $\langle \mathcal{P}(\Sigma), \subseteq \rangle \xleftarrow{\gamma_{\mathbb{D}}} \langle \mathbb{D}, \sqsubseteq_{\mathbb{D}} \rangle$. [3] This abstraction works only with a concretization function $\gamma_{\mathbb{D}} : \mathbb{D} \to \mathcal{P}(\Sigma)$ expressing the meaning of abstract elements from $\mathbb{D}$ in terms of concrete elements from $\mathcal{P}(\Sigma)$. The abstract domain $\mathbb{D}$ is a set of computer-representable properties, called *abstract elements*, together with effective algorithms to implement sound abstract operators for forward and backward analyses. In particular, they have abstract operators for ordering $\sqsubseteq_{\mathbb{D}}$, least upper bound (join) $\sqcup_{\mathbb{D}}$, greatest lower bound (meet) $\sqcap_{\mathbb{D}}$, bottom $\bot_{\mathbb{D}}$, top $\top_{\mathbb{D}}$, widening $\triangledown_{\mathbb{D}}$, and narrowing $\triangle_{\mathbb{D}}$. There are forward transfer functions for assignments $\texttt{ASSIGN}_{\mathbb{D}} : \textit{Stm} \times \mathbb{D} \to \mathbb{D}$ and tests $\texttt{FILTER}_{\mathbb{D}} : \textit{BExp} \times \mathbb{D} \to \mathbb{D}$, which are sound over-approximations of the corresponding concrete functions. We let $\texttt{lfp}^{\#}$ denote an abstract fix-point operator, derived using widening $\triangledown_{\mathbb{D}}$ and narrowing $\triangle_{\mathbb{D}}$, that over-approximates the concrete `lfp`. There are also backward transfer functions for assignments $\texttt{B-ASSIGN}^u_{\mathbb{D}} : \textit{Stm} \times \mathbb{D} \to \mathbb{D}$, tests $\texttt{B-FILTER}^u_{\mathbb{D}} : \textit{BExp} \times \mathbb{D} \to \mathbb{D}$, and a lower widening $\underline{\triangledown}_{\mathbb{D}}$ [7], which are sound under-approximations of the corresponding concrete functions. We let $\texttt{gfp}^{\#}$ denote an abstract fixpoint operator, derived using lower widening $\underline{\triangledown}_{\mathbb{D}}$, that under-approximates the concrete `gfp`.

The abstract domain $\mathbb{D}$ can be employed to derive abstract reachability (resp., abstract sufficient condition) analysis that represents over-approximation (resp., under-approximation) of the corresponding concrete analysis. For each statement s, we define the abstract reachability semantics $\overrightarrow{[\![s]\!]}^{\sharp}$ and the abstract sufficient condition semantics $\overleftarrow{[\![s]\!]}^{\sharp}$ in Fig. 4 and Fig. 5, respectively. For a `while` loop, $\texttt{lfp}^{\sharp}\ \overrightarrow{\phi}^{\sharp}_d$ is the limit of the following increasing chain: $y_0 = d$, $y_{n+1} = y_n \triangledown_{\mathbb{D}} \overrightarrow{\phi}^{\sharp}_d(y_n)$ for forward analysis; whereas $\texttt{gfp}^{\sharp}\ \overleftarrow{\phi}^{\sharp}_d$ is the limit of the following decreasing chain: $y_0 = \texttt{B-FILTER}^u_{\mathbb{D}}(\neg be, d)$, $y_{n+1} = y_n \underline{\triangledown}_{\mathbb{D}} \overleftarrow{\phi}^{\sharp}_d(y_n)$ for backward analysis. Since $\texttt{FILTER}_{\mathbb{D}}(be, d) \sqsubseteq_{\mathbb{D}} d$ and $\texttt{B-FILTER}^u_{\mathbb{D}}(be, d) \sqsupseteq_{\mathbb{D}} d$, we use $\texttt{FILTER}_{\mathbb{D}}(\texttt{??}_i, d) = d$ and $\texttt{B-FILTER}^u_{\mathbb{D}}(\texttt{??}_i, d) = d$ to handle the Boolean expression hole $\texttt{??}_i$.

By the soundness of the operators of abstract domain $\mathbb{D}$ [6], [7], we prove the soundness of abstract semantics with respect to concrete semantics. That is, $\overrightarrow{[\![s]\!]}^{\sharp}$ over-approximates the set of reachable stores obtained by concrete semantics $\overrightarrow{[\![s]\!]}$, whereas $\overleftarrow{[\![s]\!]}^{\sharp}$ under-approximates the set of sufficient condition stores obtained by concrete semantics $\overleftarrow{[\![s]\!]}$.

Proposition 3.2 ([6], [7]): We have: $\overrightarrow{[\![s]\!]}\gamma_{\mathbb{D}}(d) \subseteq \overrightarrow{[\![s]\!]}^{\sharp}d$ and $\overleftarrow{[\![s]\!]}\gamma_{\mathbb{D}}(d) \supseteq \overleftarrow{[\![s]\!]}^{\sharp}d$.

[2] $\sigma[\texttt{x} \mapsto n]$ denotes the store that updates σ at variable $\texttt{x}$ to be equal to n.

[3] Concretization-based abstraction, which uses only a concretization function $\gamma_{\mathbb{D}}$ (e.g. Polyhedra domain), represents a relaxation of the well-known Galois connection that uses both abstraction and concretization functions.

$$
\begin{array}{l}
\overrightarrow{[\![\texttt{skip}]\!]^\sharp} d = d \\
\overrightarrow{[\![\texttt{x = } ae]\!]^\sharp} d = \texttt{ASSIGN}_{\mathbb{D}}(\texttt{x = } ae, d) \\
\overrightarrow{[\![s_1 \texttt{ ; } s_2]\!]^\sharp} d = \overrightarrow{[\![s_2]\!]^\sharp}(\overrightarrow{[\![s_1]\!]^\sharp} d) \\
\overrightarrow{[\![\texttt{assume}(be)]\!]^\sharp} d = \texttt{FILTER}_{\mathbb{D}}(be, d) \\
\overrightarrow{[\![\texttt{if } be \texttt{ } s_1 \texttt{ else } s_2]\!]^\sharp} d = \overrightarrow{[\![s_1]\!]^\sharp}(\texttt{FILTER}_{\mathbb{D}}(be, d)) \sqcup_{\mathbb{D}} \\
\qquad\qquad \overrightarrow{[\![s_2]\!]^\sharp}(\texttt{FILTER}_{\mathbb{D}}(\neg be, d)) \\
\overrightarrow{[\![\texttt{while } be \texttt{ do } s]\!]^\sharp} d = \texttt{FILTER}_{\mathbb{D}}(\neg be, \texttt{lfp}^\sharp\ \overrightarrow{\phi}^\sharp_d) \\
\overrightarrow{\phi}^\sharp_d(x) = d \sqcup_{\mathbb{D}} \overrightarrow{[\![s]\!]^\sharp}(\texttt{FILTER}_{\mathbb{D}}(be, x))
\end{array}
$$

Fig. 4. Definitions of $\overrightarrow{[\![s]\!]^\sharp} : \mathbb{D} \to \mathbb{D}$.

$$
\begin{array}{l}
\overleftarrow{[\![\texttt{skip}]\!]^\sharp} d = d \\
\overleftarrow{[\![\texttt{x = } ae]\!]^\sharp} d = \texttt{B-ASSIGN}^u_{\mathbb{D}}(\texttt{x = } ae, d) \\
\overleftarrow{[\![s_1 \texttt{ ; } s_2]\!]^\sharp} d = \overleftarrow{[\![s_1]\!]^\sharp}(\overleftarrow{[\![s_2]\!]^\sharp} d) \\
\overleftarrow{[\![\texttt{assume}(be)]\!]^\sharp} d = \texttt{B-FILTER}^u_{\mathbb{D}}(be, d) \\
\overleftarrow{[\![\texttt{if } be \texttt{ } s_1 \texttt{ else } s_2]\!]^\sharp} d = \texttt{B-FILTER}^u_{\mathbb{D}}(be, \overleftarrow{[\![s_1]\!]^\sharp} d) \sqcap_{\mathbb{D}} \\
\qquad\qquad \texttt{B-FILTER}^u_{\mathbb{D}}(\neg be, \overleftarrow{[\![s_2]\!]^\sharp} d) \\
\overleftarrow{[\![\texttt{while } be \texttt{ do } s]\!]^\sharp} d = \texttt{gfp}^\sharp\ \overleftarrow{\phi}^\sharp_d \\
\overleftarrow{\phi}^\sharp_d(x) = \texttt{B-FILTER}^u_{\mathbb{D}}(\neg be, d) \sqcap_{\mathbb{D}} \texttt{B-FILTER}^u_{\mathbb{D}}(be, \overleftarrow{[\![s]\!]^\sharp} x)
\end{array}
$$

Fig. 5. Definitions of $\overleftarrow{[\![s]\!]^\sharp} : \mathbb{D} \to \mathbb{D}$.

The abstract domain $\mathbb{D}$ can be instantiated with different numerical abstract domains including Intervals [6], Octagons [7], and Polyhedra [14]. They differ in precision and computational complexity. In this work, we will mainly use the Polyhedra domain due to its precision. The *Polyhedra domain* [14] is a fully relational numerical abstract domain, which allows manipulating conjunctions of linear inequalities (constraints) of the form $\alpha_1 \texttt{x}_1 + \ldots + \alpha_n \texttt{x}_n \geq \beta$, where $\texttt{x}_1$, ..., $\texttt{x}_n$ are program variables and $\alpha_i, \beta \in \mathbb{Q}$ (rationals). The abstract operations, transfer functions, and the soundness of the Polyhedra domain are defined in [14].

Example 3.3: Consider the program sketch `while.c` in Fig. 1. Assume that the analysis fact $\top$ (true) holds at loc. ①. and we want to perform forward reachability analysis based on the Polyhedra domain. In order to enforce the convergence of the analysis, we apply the widening operator at loc. ⓗ. The analysis results inferred by our forward analysis at locs. from ① to ⑤ are shown in Fig. 6 (left).

Similarly, we assume that the analysis fact $(\texttt{y=10})$ holds at loc. ⑤ and we want to perform backward sufficient condition analysis based on the Polyhedra domain. To enforce the convergence, we apply the lower widening operator at loc. ⓗ. The analysis results inferred by our backward analysis at locs. from ⑤ to ① are shown in Fig. 6 (right).

D. Constrained Horn Clauses (CHCs)

Given a formula ϕ in a first-order theory $\mathcal{T}$, the SMT (Satisfiability Modulo Theory) solvers [18] decide the satisfiability of ϕ, i.e. they check if there is an assignment of values to variables in ϕ that makes ϕ true. To check the validity of ϕ, they check the unsatisfiability of $\neg\phi$. Examples of first-order theories are: linear integer arithmetic (LIA), arrays, etc.

① $\top$	⑤ $(\texttt{y=10})$
② $(\texttt{x=10} \wedge \texttt{y=0})$	ⓗ $\bot$
ⓗ $(\texttt{x} \leq 10 \wedge \texttt{x+y=10})$	④ $\bot$
③ $(\texttt{x} \leq 10 \wedge \texttt{x+y=10})$	③ $\bot$
④ $(\texttt{x} \leq 9 \wedge \texttt{x+y=9})$	② $\bot$
⑤ $(\texttt{x} \leq 10 \wedge \texttt{x+y=10})$	① $\bot$

Fig. 6. Analysis results of `while.c` inferred by forward (left) and backward (right) analysis.

Constrained Horn Clauses (CHCs) are a fragment of first-order theory $\mathcal{T}$ defined over a finite set of unknown predicates $\mathcal{R}$. A CHC is a formula that has one of the following forms:

$$\phi(\overrightarrow{\mathbf{x}_0}) \implies R_0(\overrightarrow{\mathbf{x}_0}) \quad (6)$$

$$\wedge_{0 \leq i \leq n} R_i(\overrightarrow{\mathbf{x}_i}) \wedge \phi(\overrightarrow{\mathbf{x}_0}, \ldots, \overrightarrow{\mathbf{x}_{n+1}}) \implies R_{n+1}(\overrightarrow{\mathbf{x}_{n+1}}) \quad (7)$$

$$\wedge_{0 \leq i \leq n} R_i(\overrightarrow{\mathbf{x}_i}) \implies \phi(\overrightarrow{\mathbf{x}_0}, \ldots, \overrightarrow{\mathbf{x}_n}) \quad (8)$$

where $R_i(\overrightarrow{\mathbf{x}_i}) \in \mathcal{R}$ $(0 \leq i \leq n+1)$ is an unknown predicate defined over a vector of variables $\overrightarrow{\mathbf{x}_i}$, and $\phi(\overrightarrow{\mathbf{x}_0}, \ldots, \overrightarrow{\mathbf{x}_n})$ is a first-order formula that does not contain any unknown predicates. The *body* (resp., *head*) of a CHC represents the left (resp., right) side of the implication. A CHC of type (6) is called *fact*, of type (7) is called *inductive*, and of type (8) is called *query*. A CHC is called *linear* if the number of unknown predicates in its body is one, otherwise it is called *non-linear*.

A *solution* to a set of CHCs defined over unknown predicates $\mathcal{R}$ is a mapping $\mathcal{M}$ from $\mathcal{R}$ to first-order formulas, such that all implications in the set become valid. That is, for each $R(\overrightarrow{\mathbf{x}}) \in \mathcal{R}$, $\mathcal{M}(R)$ represents a first-order formula ϕ defined over the variables $\overrightarrow{\mathbf{x}}$. The CHC solvers, like `Eldarica` [11], `Spacer` [10], `Duality` [15], and `FreqHorn` [16], contain algorithms, denoted `CHC-solve`, for computing the correct solutions to sets of CHCs.

IV. Synthesis Algorithm

In this section, we present our algorithm for solving the `while`-guard synthesis problem. In particular, we employ the abstract static analyses, $\overrightarrow{[\![s]\!]^\sharp}$ and $\overleftarrow{[\![s]\!]^\sharp}$, as well as the CHC solving procedure, `CHC-solve`, to automatically find Boolean expressions for the missing holes in a program sketch, so that the resulting complete program satisfies its assertion. That is, we establish a link between checking an assertion in a program and solving CHCs via abstract static analyses, which provide a way to transition from programs to CHCs.

A. Description

The $\texttt{WhileSketching}_{\texttt{CHC}}(p, H)$ synthesis procedure is shown in Algorithm 1. The procedure takes as input two parameters: a program sketch p, and a set of holes H in p representing missing `while`-guards. It returns as output a solution (complete control function ϕ) or the empty mapping $\emptyset$ if no solution exists. For each hole $\texttt{??}_i$ in H, we first generate an initial solution, which is a complete Boolean expression be'_i, such that all other holes are treated as non-deterministic

Algorithm 1: WhileSketching$_{\texttt{CHC}}(p, H)$

Input: Program sketch p, and a set of holes H
Output: Complete control function ϕ or an empty mapping $\emptyset$

1 $\phi := \phi' := \emptyset$;
2 **for** *(* $??_i \in H$*)* **do**
3 $(l: \texttt{while}\, l_h : (??_i)\, \texttt{do}\, \{l_b : s_b\}\, l' :) := \texttt{Extract}(p, ??_i)$;
4 $\phi_i^{ndet} := [??_i \mapsto \bot] \uplus \{[??_j \mapsto \mathcal{P}(\{\text{true}, \text{false}\})] \mid j \neq i\}$;
5 $be_i' := \texttt{OneHoleSynth}(p[\phi_i^{ndet}], s_b, l, l_h, l')$;
6 **if** $(be_i' = \emptyset)$ **then return** $\emptyset$ **;**
7 $\phi := \phi \uplus [\mathbf{??}_i \mapsto be_i']$
8 **while** $(\phi \neq \phi')$ **do**
9 **if** $(\phi' \neq \emptyset)$ **then** $\phi = \phi'; \phi' = \emptyset$;
10 **for** *(* $??_i \in H$*)* **do**
11 $(l: \texttt{while}\, l_h : (??_i)\, \texttt{do}\, \{l_b : s_b\}\, l' :) := \texttt{Extract}(p, ??_i)$;
12 $\phi_i := \phi[??_i \mapsto \bot]$;
13 $be_i'' := \texttt{OneHoleSynth}(p[\phi_i], s_b, l, l_h, l')$;
14 **if** $(be_i'' = \emptyset)$ **then return** $\emptyset$ **;**
15 $\phi' := \phi' \uplus [\mathbf{??}_i \mapsto be_i'']$
16 **return** ϕ

choices over Booleans, $\mathcal{P}(\{\text{true}, \text{false}\})$ (lines 2–7). First, we identify the statement "$(l : \texttt{while}\ l_h : (??_i)\ \texttt{do}\ \{l_b : s_b\}\ l' :)$" in which $??_i$ occurs by using $\texttt{Extract}(p, ??_i)$ procedure (line 3). Second, we construct a partial control function ϕ_i^{ndet}, such that $\phi_i^{ndet}(??_i) = \bot$ (undefined) and $\phi_i^{ndet}(??_j) = \mathcal{P}(\{\text{true}, \text{false}\})$ for all other $??_j \in H$ (line 4), and we generate a program sketch $p[\phi_i^{ndet}]$ in which the hole $??_i$ is the only one. This has the effect of treating any hole other than $??_i$ in $p[\phi_i^{ndet}]$ as a non-deterministic choice. Finally, we call the one-hole synthesis procedure, called $\texttt{OneHoleSynth}(p[\phi_i^{ndet}], s_b, l, l_h, l')$ to generate the solution be_i' corresponding to hole $??_i$ (line 5). This way, we generate an *initial* complete control function $\phi : [??_i \mapsto be_i']$ by using the above initial solutions for all holes $??_i$ in H.

Subsequently, Algorithm 1 goes through an iterative while-loop (lines 8-15) performing the weakening of the current solution ϕ, thus generating more general solutions ϕ'. The algorithm terminates when the refinement is no longer enabled, i.e. $\phi = \phi'$ (line 8). In particular, in each while-iteration we weaken the current solution ϕ by iteratively weakening ϕ-solutions for all holes (lines 10–15). To weaken the ϕ-solution for each $??_i$, we fix the ϕ-solutions be_j' for all other $??_j$ in a partial control function ϕ_i, such that $\phi_i(??_i) = \bot$ and $\phi_i(??_j) = \phi(??_j)$ for all other $??_j \in H$ (line 12). Then, we generate a program sketch $p[\phi_i]$ in which the hole $??_i$ is the only one. Finally, we call $\texttt{OneHoleSynth}(p[\phi_i^{ndet}], s_b, l, l_h, l')$ to find a weaker Boolean expression be_i'' for $??_i$ (line 13), which is located at statement $(l : \texttt{while}\ l_h : (??_i)\ \texttt{do}\ \{l_b : s_b\}\ l' :)$. The generated be_i'' will be included in the updated solution ϕ' (line 15). In particular, we use the solutions given by the current control function for all other holes and infer more general Boolean expression for $??_i$ that implies the assertion validity.

Algorithm 2: OneHoleSynth(p, s_b, l, l_t, l')

Input: Program sketch $p \equiv l_i : s; l_f : \texttt{assert}(be^f)$, while-body s_b, and labels l, l_h, l'
Output: Complete Boolean expression be' or an empty mapping $\emptyset$

1 $\texttt{Reach}^\sharp := \overrightarrow{[\![s]\!]}^\sharp \top_{\mathbb{D}}$;
2 $\texttt{Cond}^\sharp := \overleftarrow{[\![s]\!]}^\sharp \texttt{FILTER}_{\mathbb{D}}(be^f, \top_{\mathbb{D}})$;
3 $C_1 := (\texttt{Reach}^\sharp(l) \implies Inv(\overrightarrow{\mathbf{x}}))$;
4 $C_2 := (Inv(\overrightarrow{\mathbf{x}}) \wedge R(\overrightarrow{\mathbf{x}}) \wedge (\overrightarrow{[\![\overrightarrow{\mathbf{x'}} = \overrightarrow{\mathbf{x}}; s_b[\overrightarrow{\mathbf{x'}}/\overrightarrow{\mathbf{x}}]]\!]^\sharp} \top_{\mathbb{D}}) \implies Inv(\overrightarrow{\mathbf{x'}}))$;
5 $C_3 := (Inv(\overrightarrow{\mathbf{x}}) \wedge \neg R(\overrightarrow{\mathbf{x}}) \implies \texttt{Cond}^\sharp(l'))$;
6 $C_4 := (Inv(\overrightarrow{\mathbf{x}}) \implies \texttt{Reach}^\sharp(l_h))$;
7 $C_5 := (\texttt{Cond}^\sharp(l_h) \implies Inv(\overrightarrow{\mathbf{x}}))$;
8 $CHC := \{C_1, C_2, C_3, C_4, C_5\}$;
9 $\mathcal{M} := \texttt{CHC-solve}(CHC)$;
10 **if** $(\mathcal{M}(R(\overrightarrow{\mathbf{x}})) \Leftrightarrow \mathcal{M}(Inv(\overrightarrow{\mathbf{x}})))$ **then return** $\emptyset$ **;**
11 **return** $\texttt{SMTSimpl}((\mathcal{M}(R(\overrightarrow{\mathbf{x}})) \wedge \neg\mathcal{M}(Inv(\overrightarrow{\mathbf{x}}))))$

B. One-hole Synthesis Procedure

The $\texttt{OneHoleSynth}(p, s_b, l, l_t, l')$ procedure, shown in Algorithm 2, takes five parameters: a program sketch $p \equiv l_i : s; l_f : \texttt{assert}(be^f)$ with one hole in it, a body s_b of the while with missing guard in s, and three labels: l, l_h, and l', where the while-guard $??$ we want to handle is located (l and l' are locations before and after the while, and l_h is location at the while-head). We first call the forward abstract analysis $\overrightarrow{[\![s]\!]}^\sharp \top_{\mathbb{D}}$ (line 1) and the backward abstract analysis $\overleftarrow{[\![s]\!]}^\sharp d_{\mathcal{F}}$ (line 2), where $d_{\mathcal{F}} = \texttt{FILTER}_{\mathbb{D}}(be^f, \top_{\mathbb{D}})$ represents the assertion fact, that proceed by structural induction on the program syntax of s. This way, we compute the invariants $\texttt{Reach}^\sharp$ as well as the sufficient conditions $\texttt{Cond}^\sharp$ in all locations of p. Assume that $\texttt{Reach}^\sharp(l)$ is the inferred invariant at loc. l before the while, while $\texttt{Cond}^\sharp(l')$ is the inferred sufficient condition at loc. l' after the while that ensures assertion-safety of the code that comes after the hole $??$. Moreover, we assume that $\texttt{Reach}^\sharp(l_h)$ and $\texttt{Cond}^\sharp(l_h)$ are the inferred analysis facts at loc. l_h for the inductive while-invariant by the forward and backward analysis, respectively. By using the inferred $\texttt{Reach}^\sharp(l)$, $\texttt{Reach}^\sharp(l_h)$, $\texttt{Cond}^\sharp(l')$, and $\texttt{Cond}^\sharp(l_h)$, we then construct CHCs with two unknown predicates: $Inv(\overrightarrow{\mathbf{x}})$ and $R(\overrightarrow{\mathbf{x}})$ corresponding to the inductive while-invariant and the guard of the while with the hole $??$, where $\overrightarrow{\mathbf{x}}$ is a vector of all program variables defined in the scope of the given while. More specifically, we generate the following five CHC queries for all cases involving the given while. The inductive while-invariant $Inv(\overrightarrow{\mathbf{x}})$ expresses a relationship between the program variables that is implied by the while's precondition and is preserved by any execution of the while-body. For this aim, we use the CHCs C_1 and C_2 to ensure that the while-invariant $Inv(\overrightarrow{\mathbf{x}})$ is inductive. In particular, C_1 states that the invariant at the while's precondition $\texttt{Reach}^\sharp(l)$ implies $Inv(\overrightarrow{\mathbf{x}})$. To state

that $Inv(\overrightarrow{\mathbf{x}})$ is preserved in each iteration of the `while`'s body, we need a way of remembering the initial values of all variables $\overrightarrow{\mathbf{x}}$ in the scope of the given `while`. To do so, we analyze the program: $\overrightarrow{\mathbf{x'}} = \overrightarrow{\mathbf{x}}; s_b[\overrightarrow{\mathbf{x'}}/\overrightarrow{\mathbf{x}}]$, where $\overrightarrow{\mathbf{x'}} = \overrightarrow{\mathbf{x}}$ assign all program variables $\overrightarrow{\mathbf{x}}$ to their primed versions $\overrightarrow{\mathbf{x'}}$ and $s_b[\overrightarrow{\mathbf{x'}}/\overrightarrow{\mathbf{x}}]$ means that all variables $\overrightarrow{\mathbf{x}}$ in s_b are renamed to their primed versions $\overrightarrow{\mathbf{x'}}$. This way, we build the required CHC C_2 stating that if $Inv(\overrightarrow{\mathbf{x}})$ and $R(\overrightarrow{\mathbf{x}})$ hold and we analyze $\overrightarrow{\mathbf{x'}} = \overrightarrow{\mathbf{x}}; s_b[\overrightarrow{\mathbf{x'}}/\overrightarrow{\mathbf{x}}]$, then $Inv(\overrightarrow{\mathbf{x'}})$ will hold after that. That is, executing `while`'s body when $R(\overrightarrow{\mathbf{x}})$ holds re-establishes the `while`-invariant $Inv(\overrightarrow{\mathbf{x}})$. The CHC C_3 states that on `while` exit, the `while`-invariant $Inv(\overrightarrow{\mathbf{x}})$ strengthen with $\neg R(\overrightarrow{\mathbf{x}})$ are sufficiently strong to imply the `while`'s postcondition $\texttt{Cond}^\sharp(l')$ ensuring that the assertion is correct. The CHCs C_4 and C_5 state that the analysis facts $\texttt{Reach}^\sharp(l_h)$ and $\texttt{Cond}^\sharp(l_h)$ inferred at loc. l_h by forward and backward analysis are over- and under-approximations of $Inv(\overrightarrow{\mathbf{x}})$, respectively.

Finally, we call the `CHC-solve` procedure to find the solution (mapping) $\mathcal{M}$ of the constructed set of CHCs, such that $\mathcal{M}(Inv)$ and $\mathcal{M}(R)$ are first-order formulas defined over variables $\overrightarrow{\mathbf{x}}$. Since the solution $\mathcal{M}(R)$ represents a strengthen formula of $\mathcal{M}(Inv)$, we can use $\mathcal{M}(R) \wedge \neg\mathcal{M}(Inv)$ as the final solution be' to replace the given Boolean expression hole `??` (line 11). However, if $(\mathcal{M}(R) \Leftrightarrow \mathcal{M}(Inv))$, i.e. $\mathcal{M}(R)$ and $\mathcal{M}(Inv)$ are equivalent, then the given problem is not solvable and an empty mapping is returned (line 10).

C. Correctness

The following theorem states correctness of the $\texttt{WhileSketching}_{\texttt{CHC}}$ algorithm.

Theorem 1: $\texttt{WhileSketching}_{\texttt{CHC}}(p, H)$ is correct and terminates.

Proof: The procedure $\texttt{WhileSketching}_{\texttt{CHC}}(p, H)$ terminates since all steps in it are terminating. The `while`-loop (lines 8-15) terminates since in the worst case it will generate the weakest solutions true for all holes. The correctness of $\texttt{WhileSketching}_{\texttt{CHC}}(p, H)$ follows from the soundness of $\texttt{Reach}^\sharp$ and $\texttt{Cond}^\sharp$ (see Proposition 3.2) and the correctness of `CHC-solve` [11], [10], [15], [16]. ■

V. Abduction-based Synthesis

We now give an overview of the logical abduction-based approach for `while`-guard synthesis, called $\texttt{GenSketching}_{\texttt{Abd}}$, introduced in [5]. The logical abduction [9] is a technique for inferring a single unknown predicate $R(\overrightarrow{\mathbf{x}})$, which is defined over a vector of program variables $\overrightarrow{\mathbf{x}}$ from an implication formula:

$$R(\overrightarrow{\mathbf{x}}) \wedge \chi \implies C$$

where χ and C are first-order logical formulas. The solution to the above abduction problem is formula ϕ, which represents an interpretation of $R(\overrightarrow{\mathbf{x}})$, such that (1) $\phi \wedge \chi \not\models$ false; and (2) $\phi \wedge \chi \models C$. The procedure $\texttt{Abduce}(\chi, C, \overrightarrow{\mathbf{x}})$ that finds the logically weakest solution containing a fewest number of variables to the above abduction problem is implemented in the `EXPLAIN` tool [9].

We now show how to handle the case when there is a single hole in a sketch p, which occurs in the statement of the form: "l : (`while` l_h : (`??`) `do` $\{l_b : s_b\})\, l'$:". The case when there are more than one holes in a program sketch is handled similarly as in Algorithm 1. First, we perform forward and backward abstract static analyses of p to compute the invariants $\texttt{Reach}^\sharp$ and the sufficient conditions $\texttt{Cond}^\sharp$ in all locations of p. Then, we construct an abduction query where the premise is $\texttt{Reach}^\sharp(l_h)$, the desired conclusion is $\texttt{Cond}^\sharp(l')$, and the unknown predicate is $R_{false}(\overrightarrow{\mathbf{x}})$, that is: $\texttt{Reach}^\sharp(l_h) \wedge R_{false}(\overrightarrow{\mathbf{x}}) \implies \texttt{Cond}^\sharp(l')$. The meaning of this query is that when `while` terminates, the `while`-invariant $\texttt{Reach}^\sharp(l_h)$ and the negation of the missing `while`-guard, encoded via the unknown predicate $R_{false}(\overrightarrow{\mathbf{x}})$, should imply the sufficient condition $\texttt{Cond}^\sharp(l')$ ensuring that the code after `while` guarantees the assertion validity.

Next, we find $R_{true}(\overrightarrow{\mathbf{x}})$ that is the negation of the formula found for $R_{false}(\overrightarrow{\mathbf{x}})$, i.e. $R_{true}(\overrightarrow{\mathbf{x}}) \equiv (\neg R_{false}(\overrightarrow{\mathbf{x}}))$. Assume that $R_{true}(\overrightarrow{\mathbf{x}})$ is of the form: $\phi_1 \vee \ldots \vee \phi_n$, where ϕ_i is a conjunction of formulas. We perform n backward analyses defined as follows: $\overleftarrow{[\![p[\texttt{??} \mapsto (\phi_1 \wedge \texttt{??})]]\!]^\sharp}(\texttt{FILTER}_{\mathbb{D}}(be^f, \top_{\mathbb{D}})), \ldots, \texttt{Cond}^{\sharp,n} = \overleftarrow{[\![p[\texttt{??} \mapsto (\phi_n \wedge \texttt{??})]]\!]^\sharp}(\texttt{FILTER}_{\mathbb{D}}(be^f, \top_{\mathbb{D}}))$, where $p[\texttt{??} \mapsto (\phi_i \wedge \texttt{??})]$ is the sketch p in which the hole `??` is replaced with $(\phi_i \wedge \texttt{??})$. We create n abduction queries: $\texttt{Reach}^\sharp(l_h) \wedge R^1_{true}(\overrightarrow{\mathbf{x}}) \implies \texttt{Cond}^{\sharp,1}(l_b)$, ..., $\texttt{Reach}^\sharp(l_h) \wedge R^n_{true}(\overrightarrow{\mathbf{x}}) \implies \texttt{Cond}^{\sharp,n}(l_b)$. Finally, the solution for `??` is the Boolean expression: $(\phi_1 \wedge R^1_{true}(\overrightarrow{\mathbf{x}})) \vee \ldots \vee (\phi_n \wedge R^n_{true}(\overrightarrow{\mathbf{x}}))$. In the implementation, we can make some optimizations to reduce the number of calls to the abduction solver. For example, if $\texttt{Cond}^{\sharp,i}(l_b) =$ false then we can conclude that $R^i_{true}(\overrightarrow{\mathbf{x}}) \equiv$ (false) without calling the abduction solver.

Example 5.1: Consider the sketch `while.c` given in Fig. 1. The forward analysis infers the invariant $(\texttt{x} \leq 10) \wedge (\texttt{x+y=10})$ at loc. l_h. We construct the abduction query $(\texttt{x} \leq 10) \wedge (\texttt{x+y=10}) \wedge R_{false}(\texttt{x,y}) \implies (\texttt{y=10})$. The reported solution by EXPLAIN is $R_{false}(\texttt{x,y}) \equiv (\texttt{x=0})$. We compute $R_{true}(\texttt{x,y}) \equiv \neg R_{false}(\texttt{x,y}) \equiv (\texttt{x<0}) \vee (\texttt{x>0})$. Hence, we perform two backward analysis of `while.c` in which the `while`-guard is $(\texttt{x<0} \wedge \texttt{??}_1)$ and $(\texttt{x>0} \wedge \texttt{??}_2)$, respectively. They start with the fact $(\texttt{x=0} \wedge \texttt{y=10})$ at loc. ⑤. The first backward analysis for $(\texttt{x<0} \wedge \texttt{??}_1)$ infers the bottom ($\bot$) condition after the guard at loc. ③, whereas the second backward analysis for $(\texttt{x>0} \wedge \texttt{??}_2)$ infers $(1 \leq \texttt{x} \leq 11) \wedge (\texttt{x+y=10})$ after the guard at loc. ③. We construct two abduction queries: (1) $(\texttt{x} \leq 10 \wedge \texttt{x+y=10}) \wedge R^1_{true}(\texttt{x,y}) \implies$ false, and (2) $(\texttt{x} \leq 10 \wedge \texttt{x+y=10}) \wedge R^2_{true}(\texttt{x,y}) \implies (1 \leq \texttt{x} \leq 11) \wedge (\texttt{x+y=10})$. The obtained solutions are $R^1_{true}(\texttt{x,y}) \equiv$ false and $R^2_{true}(\texttt{x,y}) \equiv$ true. Hence, we fill the hole with $(\texttt{x>0})$.

VI. Evaluation

We now evaluate our approach based on abstract static analysis and CHC solving for synthesizing missing `while`-guards in numerical sketches written in C. We compare its performances against the approach that uses logical abduction for constraint solving.

```
void main(int n){
  assume(n ≥ 0)
  int i = 0;
  while (??) do{
      i = i+2;
  }
  assert(i==2*n);
 // assert(i==4*n);
 // assert(i==0);
 // assert(i<0);
}
```

Fig. 7. eq2.c.

```
void main(int x) {
  assume(x ≥ 0)
  int a = x;
  int y = 0;
  while(??) do{
    y = y+1;
    a = a-1;
  }
  assert(x==y);
 // assert(a==0);
 // assert(a+2==x);
 // assert(y<0);
}
```

Fig. 9. copy.c.

```
void main() {
  int i = 0;
  int j = 0;
  while(??) do{
    i = i+1;
    j = j+1;
  }
  assert(i+j==20);
 // assert(i+j>20);
 // assert(i+j≤15);
 // assert(i+j==15);
}
```

Fig. 8. loop2.c.

```
void main() {
  int x = 0;
  int y = 0;
  while(??) do{
    x = x+1;
    y = y+2;
  }
  assert(y==2*m);
 // assert(y==x);
 // assert(y==x && x>0);
 // assert(y==2*m+1);
}
```

Fig. 10. hola.c.

A. Implementation

We have implemented the proposed synthesis algorithm in a proof-of-concept tool, called $\texttt{WhileSketching}_{\texttt{CHC}}$. The abstract operations and transfer functions of the Polyhedra abstract domain [14] are provided by the APRON library [17]. The CHC and SMT queries are solved by the `Eldarica` [11] and the `Z3` [18] tools. Our proof-of-concept tool is written in OCAML and consists of around 7K lines of code. The tool accepts programs written in a subset of C. It currently provides only a limited support for arrays, pointers, `struct` and `union` types. The only basic data type is mathematical integers. The tool calls the APRON library for static analysis of the input program, then generates CHC constraints and calls the `Eldarica` tool to solve them. Finally, it handles the obtained results and reports the solutions.

B. Experiment setup and Benchmarks

All experiments are executed on a 64-bit Intel®Core^{TM} i5 CPU, VM Lubuntu 18.04, with 16 GB memory, and we use a timeout value of 60 sec. All times are reported as average over five independent executions. We compare two approaches:

(1) $\texttt{WhileSketching}_{\texttt{CHC}}$ based on abstract static analysis and CHC solving; and
(2) $\texttt{GenSketching}_{\texttt{Abd}}$ based on abstract static analysis and abduction solving.

We report TIME which is the total time to resolve a given problem, SOLVTIME which is the time taken by logical (CHC or abduction) solvers to solve the logical queries in the given synthesis task, and PREC which shows the precision of the reported solution. We use: ✓ when the tool finds correct solution for a given assertion; and × denotes the opposite, i.e. the tool cannot find a correct solution for the assertion.

The evaluation is performed on several C numerical sketches collected from SV-COMP, (https://sv-comp.sosy-lab.org/) and the literature [19], [5] In particular, we use the following benchmarks: `while.c` (Fig. 1), `eq2.c` (Fig. 7), `loop2.c` (Fig. 8), `copy.c` (Fig. 9), and `hola.c` (Fig. 10). In all benchmarks, we assume that `while`-guards are missing and we employ the aforementioned approaches to synthesize suitable Boolean expressions for them.

C. Performance Results

Table I shows the performance results of $\texttt{WhileSketching}_{\texttt{CHC}}$ and $\texttt{GenSketching}_{\texttt{Abd}}$.

We described in Section II how $\texttt{WhileSketching}_{\texttt{CHC}}$ successfully handles the sketch `while.c` given in Fig. 1. $\texttt{GenSketching}_{\texttt{Abd}}$ also finds the correct solutions for `while.c`, but for the first two assertions it needs two calls to the abduction solver, which results in slower running times.

The program sketch `eq2.c` in Fig. 7 has one local variable `i`, which is initialized to `0` and increased by `2` in every `while`-iteration. The forward static analysis infers the invariant $(\texttt{i=0} \wedge \texttt{n} \geq \texttt{0})$ before the `while` and $(\texttt{i} \geq \texttt{0} \wedge \texttt{n} \geq \texttt{0})$ at the `while`-head. For the specification `assert(i==2*n)`, $\texttt{WhileSketching}_{\texttt{CHC}}$ constructs the corresponding CHCs

TABLE I
PERFORMANCE RESULTS OF WHILESKETCHING$_{CHC}$ VS. GENSKETCHING$_{ABD}$. ALL TIMES IN SEC.

Bench.	assert	WhileSketching$_{CHC}$			GenSketching$_{Abd}$		
		TIME	SOLVTIME	Prec	TIME	SOLVTIME	Prec
while.c	(y=10)	0.920	0.904	✓	2.011	1.836	✓
while.c	(y>x)	0.924	0.901	✓	2.082	1.874	✓
while.c	(x=10)	0.817	0.796	✓	0.746	0.611	✓
while.c	(x>15)	0.921	0.900	✓	0.744	0.630	✓
eq2.c	(i=2⋆n)	0.939	0.889	✓	1.337	1.251	✓
eq2.c	(i=4⋆n)	0.901	0.871	✓	1.578	1.402	✓
eq2.c	(i=0)	0.813	0.789	✓	0.742	0.637	✓
eq2.c	(i<0)	0.903	0.872	✓	0.745	0.631	✓
loop2.c	(i+j=20)	0.945	0.911	✓	1.502	1.364	✓
loop2.c	(i+j>20)	0.936	0.902	✓	1.558	1.404	✓
loop2.c	(i+j<=20)	0.814	0.786	✓	1.543	1.376	✓
loop2.c	(i+j=15)	0.951	0.910	✓	0.751	0.665	✓
copy.c	(x=y)	0.979	0.937	✓	1.693	1.542	×
copy.c	(a=0)	0.954	0.921	✓	1.511	1.382	×
copy.c	(a+2=x)	0.950	0.918	✓	1.451	1.329	×
copy.c	(y<0)	0.947	0.916	✓	0.757	0.679	✓
hola.c	(y=2⋆m)	0.948	0.913	✓	1.511	1.364	×
hola.c	(y=x)	0.825	0.764	✓	1.542	1.350	✓
hola.c	(y=x∧x>0)	0.952	0.921	✓	0.756	0.667	✓
hola.c	(y=2⋆m+1)	0.926	0.902	✓	0.748	0.664	×

and generates the solution (i ≠ n) for the while-guard ??. For the assertions (i==4⋆n), (i==0), and (i<0), we synthesize the following while-guards: (i ≠ 2⋆n), false, and the problem is unrealizable, respectively. On the other hand, GenSketching$_{Abd}$ reports the correct solutions for eq2.c, but it shows slower performance times than WhileSketching$_{CHC}$ for the first two assertions.

Consider the program sketch loop2.c in Fig. 8. The forward static analysis infers the invariant (i=0∧j=0) before the while and (i=0 ∧ i=j) at the while-head. Hence, WhileSketching$_{CHC}$ infers the solutions (i ≠ 10), (i ≤ 10), and false for the assertions (i+j==20), (i+j>20), and (i+j ≤ 15), respectively. However, for the assertion (i+j==15), we obtain that the problem is unrealizable (i.e., it has no solution). GenSketching$_{Abd}$ finds the correct solutions as well, but is slower for three assertions.

The program sketch copy.c in Fig. 9 copies the input value of the argument x into the local variable y. This is specified by the final assertion (x==y). The forward static analysis infers the invariant (x ≥ 0 ∧ y=0 ∧ x=a) before the while and (x ≥ 0 ∧ y=x−a ∧ x−a ≥ 0) at the while-head. Hence, WhileSketching$_{CHC}$ infers the solution (x ≠ y) for the missing while-guard ??. For the assertions (a==0), (a+2==x), and (y<0), it generates the while-guards: (x−y ≥ 1), (y ≠ 2), and the problem is unrealizable, respectively. On the other hand, GenSketching$_{Abd}$ fails to find solutions for three assertions.

Consider the program sketch hola.c in Fig. 10 [19]. The forward static analysis infers the invariant (x=0 ∧ y=0) before the while and (x ≥ 0 ∧ y=2⋆x) at the while-head. This way, WhileSketching$_{CHC}$ finds the solution (m ≠ x) for the assertion (y==2⋆m). For assertion (y==x) the solution is (false), whereas for assertions (y==x)∧(x>0) and (y==2⋆m+1) we obtain that the problems are unrealizable. GenSketching$_{Abd}$ fails to produce correct solutions for two assertions in this case.

Discussion: In summary, as Table I shows, the proposed technique WhileSketching$_{CHC}$ is quite effective at successfully synthesizing the missing while-guards ?? in all considered benchmarks. Note that the tasks that synthesize the solution (false) are the fastest ones for WhileSketching$_{CHC}$ achieving around 10% speed-up with respect to the other tasks for the same benchmark. Moreover, the CHC and abduction solvers take more than 95% of the total synthesis time for both WhileSketching$_{CHC}$ and GenSketching$_{Abd}$. The synthesis time of GenSketching$_{Abd}$ depends on how many times the abduction solver is called. When the Abduction solver is called once, then GenSketching$_{Abd}$ and WhileSketching$_{CHC}$ achieve comparable synthesis times, since WhileSketching$_{CHC}$ always calls the CHC solver once. However, we can see that GenSketching$_{Abd}$ cannot find solutions for five assertions, thus returning empty sets. On the one hand, this is due to the limitation of the abduction solver and the difficulty of static analyzers to infer precise inductive while-invariants in the presence of missing while-guards, but on the other hand, this is also due to the suitability of CHC solvers for handling complex control flow.

The current tool supports an interesting subset of C, so we can handle many interesting programs. The selected benchmarks are chosen to show some distinctive properties of our approach and its ability to be applied in practice.

VII. RELATED WORK

We divide our discussion of related work into three categories: program sketching; abstract static analysis; and CHC solving.

A. Program sketching

The widely-known SKETCH tool [2], [3] represents one of the earliest attempts to resolve program sketches in which missing holes can be only integer constants. It uses SAT-based counterexample-guided inductive synthesis. In particular, SKETCH iteratively generates a finite set of inputs and performs SAT queries to identify values for the holes such that the obtained complete program satisfies all assertions for the considered inputs. Additional SAT queries are then checked to establish whether the obtained program is correct on all possible inputs. SKETCH handles `while`-s by unrolling them, which makes its performance very sensitive to the degree of unrolling. Since it is based on SAT solvers, SKETCH is especially suited for handling bit-manipulating programs.

FAMILYSKETCHER [4], [20], [21] is another tool for solving the sketching problem by using lifted (family-based) static analysis based on abstract interpretation [22], [23]. The key idea underlying this approach is that the set of all possible sketch realizations can be represented as a program family (Software Product Line) with numerical features [24], [25], [26], which is subsequently statically analyzed to establish assertion validity. This way, the effort of conducting an effective search of all possible hole realizations is delegated to an efficient lifted static analyzer for program families (SPLs), which uses a specifically designed decision tree abstract domain. Since FAMILYSKETCHER is based on numerical abstract domains, it can handle unbounded `while`-s via widening and narrowing, and moreover it is very suited for handling numerical programs with integer data types.

However, SKETCH and FAMILYSKETCHER can only resolve program sketches in which unknown holes can be replaced by one constant value from a finite set of integers. $\texttt{GenSketching}_{\texttt{Abd}}$ [5] considers so-called generalized sketching problem, where each hole can be replaced by an arbitrary expression. It uses abstract static analysis for constraint generation and abduction solvers for constraint solving. Similarly to FAMILYSKETCHER, $\texttt{GenSketching}_{\texttt{Abd}}$ can handle unbounded `while`-s and numerical programs. However, this approach is not very effective in handling missing `while`-guards, and it often fails to find any solution. In this paper, we pursue this line of work by using CHC solvers rather than abduction solvers, which are more suitable for automatically handling more complex `while`-loops. The abduction solvers have also been applied in program synthesis to infer missing `if`-guards from low-level C code such that all buffer accesses are memory safe [27], to infer missing `#if`-guards from numerical program families so that a given assertion is valid in all members of the family [28], to find the weakest specifications of unknown (library) functions that ensure the safety of the program calling those external functions [29], [30], etc.

B. Abstract static analysis

Abstract interpretation [6], [14], [7], [31], [32] is a general theory for approximating the semantics of programs. It provides sound and efficient static analyses for inferring run-time properties of programs. The most popular static analyzers employ the well-known numerical abstract domains of Intervals [6], Octagons [7], and Polyhedra [14], that can be used for inferring numerical properties of program variables. Mine [32] has introduced the under-approximating backward polyhedral analysis for inferring sufficient conditions of non-deterministic programs, which has been implemented as part of the APRON library in the BANAL tool [32].

Several successful static analyzers based on abstract interpretation have been developed recently for verifying real-world programs. For example, ASTREE [33] is a static analyzer for verifying avionics software; ULITIMATE TAIPAN [34] is a software model checker that uses abstract interpretation to derive invariants for the path program corresponding to a given spurious counterexample; PAGAI [35] combines SMT-solving with abstract interpretation, etc.

Several approaches use a combination of forward and backward static analyses to infer more precise interesting properties of programs. More specifically, Rival [36] uses this combination to detect a set of traces leading to a bug; Bourdoncle [31] to generate preconditions ensuring validity of invariant and intermittent assertions; Dimovski and Legay [37] to compute the probability that a given assertion is satisfied (or violated); and Urban and Miné [8] to find ranking functions for proving termination.

C. Constraint Horn Clauses

Recently, the problem of CHC solving has gained a lot of attention due to their wide applications in program verification and program synthesis. Various algorithms and tools for CHC solving have been developed. Some of the most prominent tools are: `Eldarica` [11] that combines predicate abstraction with counterexample-guided abstraction refinement; `Spacer` [10] that uses generalized property-driven reachability analysis; `Duality` [15] that uses predicate abstraction and Craig interpolation for abstraction refinement; `FreqHorn` [16] that uses Syntax-Guided Synthesis (SyGuS) [38]. Optimal CHC solving [39], which calculates maximal and minimal solutions of a given set of CHCs, uses techniques from program termination analysis to check the optimality of a solution of CHCs. Algorithms for solving a richer formalism of existentially quantified Horn clauses have also been studied [40], [41].

CHC solving has found a number of applications in program verification and program synthesis. SEAHORN [42] and JAYHORN [43] are tools for verification of safety properties of C and Java programs by combining abstract interpretation and CHC solving. They encode Hoare logic-style verification conditions as CHCs, which are subsequently solved via CHC solvers. CHC solving has also been used for determining refinement types of higher-order functional programs [44], for analysis of business processes expressed as Petri nets [45], etc. The existential Horn clauses have been used for the weakest strategy synthesis of safety games [46].

CHC solving has also been applied for synthesizing imperative programs from input-output examples as specifications in the semantics-guided synthesis (SemGuS) framework [47], [48]. The syntax, semantics, and specifications of imperative

programs are represented as CHCs, thus effectively reducing the program synthesis problem into a proof search over CHCs. This approach is specialized on proving unrealizable problems by showing that no imperative program from a given grammar exists that satisfies the specification. The SemGuS framework is also used for synthesizing regular expressions, Boolean formulas, and SyGuS problems. There are other approaches for imperative program synthesis that are specialized for solving realizable problems. SIMPL [49] combines the enumerative search with abstract static analysis to infer programs from input-output examples, while IMPSYNTH [50] combines enumerative search, abstract static analysis, and SMT mutations [51] to infer programs from logic specifications.

VIII. Conclusion

In this work, we introduce a synthesis algorithm for automatically inferring arbitrary Boolean expressions for unknown `while`-guards in program sketches. It is implemented by interaction between abstract static analyses and CHC solving. By means of experiments, we demonstrate the effectiveness of our algorithm on a set of interesting C benchmarks that cannot be handled by other existing tools.

In the future, we plan to extend our approach by considering program sketches in which apart from `while`-guards, unknown holes can also be `if`-guards, arithmetic expressions in assignments, and ultimately arbitrary statements. We envision achieving this aim by further interplay between abstract static analyses, CHC solving, and abduction solving.

References

[1] A. Pnueli and R. Rosner, "On the synthesis of a reactive module," in *Conference Record of the Sixteenth Annual ACM Symposium on Principles of Programming Languages*. ACM Press, 1989, pp. 179–190. [Online]. Available: https://doi.org/10.1145/75277.75293

[2] A. Solar-Lezama, "Program sketching," *STTT*, vol. 15, no. 5-6, pp. 475–495, 2013. [Online]. Available: https://doi.org/10.1007/s10009-012-0249-7

[3] A. Solar-Lezama, R. M. Rabbah, R. Bodík, and K. Ebcioglu, "Programming by sketching for bit-streaming programs," in *Proceedings of the ACM SIGPLAN 2005 Conference on Programming Language Design and Implementation*. ACM, 2005, pp. 281–294. [Online]. Available: https://doi.org/10.1145/1065010.1065045

[4] A. Dimovski, S. Apel, and A. Legay, "Program sketching using lifted analysis for numerical program families," in *NASA Formal Methods, NFM 2021*, ser. LNCS, vol. 12673. Springer, 2021, pp. 95–112. [Online]. Available: https://doi.org/10.1007/978-3-030-76384-8_7

[5] A. S. Dimovski, "Generalized program sketching by abstract interpretation and logical abduction," in *Static Analysis - 30th International Symposium, SAS 2023, Proceedings*, ser. LNCS, vol. 14284. Springer, 2023, pp. 212–230. [Online]. Available: https://doi.org/10.1007/978-3-031-44245-2_11

[6] P. Cousot and R. Cousot, "Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints," in *POPL'77*. ACM, 1977, pp. 238–252. [Online]. Available: http://doi.acm.org/10.1145/512950.512973

[7] A. Miné, "Tutorial on static inference of numeric invariants by abstract interpretation," *Foundations and Trends in Programming Languages*, vol. 4, no. 3-4, pp. 120–372, 2017. [Online]. Available: https://doi.org/10.1561/2500000034

[8] C. Urban and A. Miné, "A decision tree abstract domain for proving conditional termination," in *Static Analysis - 21st International Symposium, SAS 2014. Proceedings*, ser. LNCS, vol. 8723. Springer, 2014, pp. 302–318. [Online]. Available: https://doi.org/10.1007/978-3-319-10936-7_19

[9] I. Dillig and T. Dillig, "Explain: A tool for performing abductive inference," in *Computer Aided Verification - 25th International Conference, CAV 2013. Proceedings*, ser. LNCS, vol. 8044. Springer, 2013, pp. 684–689. [Online]. Available: https://doi.org/10.1007/978-3-642-39799-8_46

[10] A. Komuravelli, A. Gurfinkel, and S. Chaki, "Smt-based model checking for recursive programs," in *Computer Aided Verification - 26th International Conference, CAV 2014. Proceedings*, ser. LNCS, vol. 8559. Springer, 2014, pp. 17–34. [Online]. Available: https://doi.org/10.1007/978-3-319-08867-9_2

[11] H. Hojjat and P. Rümmer, "The ELDARICA horn solver," in *2018 Formal Methods in Computer Aided Design, FMCAD 2018*. IEEE, 2018, pp. 1–7. [Online]. Available: https://doi.org/10.23919/FMCAD.2018.8603013

[12] C. A. R. Hoare and N. Wirth, "An axiomatic definition of the programming language PASCAL," *Acta Informatica*, vol. 2, pp. 335–355, 1973.

[13] A. Miné, "The octagon abstract domain," *Higher-Order and Symbolic Computation*, vol. 19, no. 1, pp. 31–100, 2006. [Online]. Available: https://doi.org/10.1007/s10990-006-8609-1

[14] P. Cousot and N. Halbwachs, "Automatic discovery of linear restraints among variables of a program," in *Conference Record of the Fifth Annual ACM Symposium on POPL'78*. ACM Press, 1978, pp. 84–96. [Online]. Available: https://doi.org/10.1145/512760.512770

[15] N. S. Bjørner, K. L. McMillan, and A. Rybalchenko, "On solving universally quantified horn clauses," in *Static Analysis - 20th International Symposium, SAS 2013. Proceedings*, ser. LNCS, vol. 7935. Springer, 2013, pp. 105–125. [Online]. Available: https://doi.org/10.1007/978-3-642-38856-9_8

[16] G. Fedyukovich, S. Prabhu, K. Madhukar, and A. Gupta, "Quantified invariants via syntax-guided synthesis," in *Computer Aided Verification - 31st International Conference, CAV 2019, Proceedings, Part I*, ser. LNCS, vol. 11561. Springer, 2019, pp. 259–277. [Online]. Available: https://doi.org/10.1007/978-3-030-25540-4_14

[17] B. Jeannet and A. Miné, "Apron: A library of numerical abstract domains for static analysis," in *21st Int. Conf., CAV 2009. Proceedings*, ser. LNCS, vol. 5643. Springer, 2009, pp. 661–667. [Online]. Available: https://doi.org/10.1007/978-3-642-02658-4_52

[18] L. M. de Moura and N. Bjørner, "Z3: an efficient SMT solver," in *14th International Conference, TACAS 2008. Proceedings*, ser. LNCS, vol. 4963. Springer, 2008, pp. 337–340. [Online]. Available: https://doi.org/10.1007/978-3-540-78800-3_24

[19] I. Dillig, T. Dillig, B. Li, and K. L. McMillan, "Inductive invariant generation via abductive inference," in *Proceedings of the 2013 ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages & Applications, OOPSLA 2013*. ACM, 2013, pp. 443–456. [Online]. Available: https://doi.org/10.1145/2509136.2509511

[20] A. S. Dimovski, "Quantitative program sketching using lifted static analysis," in *25th Int. Conf. FASE 2022, Proceedings*, ser. LNCS, vol. 13241. Springer, 2022, pp. 102–122. [Online]. Available: https://doi.org/10.1007/978-3-030-99429-7_6

[21] A. Dimovski, "Quantitative program sketching using decision tree-based lifted analysis," *J. Comput. Lang.*, vol. 75, p. 101206, 2023. [Online]. Available: https://doi.org/10.1016/j.cola.2023.101206

[22] A. S. Dimovski, "A binary decision diagram lifted domain for analyzing program families," *J. Comput. Lang.*, vol. 63, p. 101032, 2021. [Online]. Available: https://doi.org/10.1016/j.cola.2021.101032

[23] A. S. Dimovski, S. Apel, and A. Legay, "Several lifted abstract domains for static analysis of numerical program families," *Sci. Comput. Program.*, vol. 213, p. 102725, 2022. [Online]. Available: https://doi.org/10.1016/j.scico.2021.102725

[24] A. S. Dimovski, C. Brabrand, and A. Wasowski, "Finding suitable variability abstractions for family-based analysis," in *FM 2016: Formal Methods - 21st International Symposium, Proceedings*, ser. LNCS, vol. 9995. Springer, 2016, pp. 217–234. [Online]. Available: https://doi.org/10.1007/978-3-319-48989-6_14

[25] A. S. Dimovski and A. Wasowski, "From transition systems to variability models and from lifted model checking back to UPPAAL," in *Models, Algorithms, Logics and Tools*, ser. LNCS, vol. 10460. Springer, 2017, pp. 249–268. [Online]. Available: https://doi.org/10.1007/978-3-319-63121-9_13

[26] B. Atanasovski, M. Bogdanovic, G. Velinov, L. Stoimenov, A. S. Dimovski, B. Koteska, D. Jankovic, I. Skrceska, M. Kon-Popovska, and B. Jakimovski, "On defining a model driven architecture for an enterprise

e-health system," *Enterp. Inf. Syst.*, vol. 12, no. 8-9, pp. 915–941, 2018. [Online]. Available: https://doi.org/10.1080/17517575.2018.1521996

[27] T. Dillig, I. Dillig, and S. Chaudhuri, "Optimal guard synthesis for memory safety," in *Computer Aided Verification - 26th International Conference, CAV 2014. Proceedings*, ser. LNCS, vol. 8559. Austria: Springer, 2014, pp. 491–507. [Online]. Available: https://doi.org/10.1007/978-3-319-08867-9_32

[28] A. S. Dimovski, "On synthesizing presence conditions in numerical software product lines," in *Proceedings of the 29th ACM International Systems and Software Product Line Conference - Volume A, SPLC 2025*. ACM, 2025. [Online]. Available: https://doi.org/10.1145/3744915.3748474

[29] A. Albarghouthi, I. Dillig, and A. Gurfinkel, "Maximal specification synthesis," in *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2016*. St. Petersburg: ACM, 2016, pp. 789–801. [Online]. Available: https://doi.org/10.1145/2837614.2837628

[30] A. S. Dimovski, "Weakest safe context synthesis by symbolic game semantics and logical abduction," in *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025*. ACM, 2025, pp. 1990–1997. [Online]. Available: https://doi.org/10.1145/3672608.3707849

[31] F. Bourdoncle, "Abstract debugging of higher-order imperative languages," in *Proceedings of the ACM SIGPLAN'93 Conference on Programming Language Design and Implementation (PLDI)*. ACM, 1993, pp. 46–55. [Online]. Available: https://doi.org/10.1145/155090.155095

[32] A. Miné, "Backward under-approximations in numeric abstract domains to automatically infer sufficient program conditions," *Sci. Comput. Program.*, vol. 93, pp. 154–182, 2014. [Online]. Available: https://doi.org/10.1016/j.scico.2013.09.014

[33] P. Cousot, R. Cousot, J. Feret, L. Mauborgne, A. Miné, D. Monniaux, and X. Rival, "The astreé analyzer," in *Programming Languages and Systems, 14th European Symposium on Programming, ESOP 2005, Proceedings*, ser. LNCS, vol. 3444. Springer, 2005, pp. 21–30. [Online]. Available: https://doi.org/10.1007/978-3-540-31987-0_3

[34] M. Greitschus, D. Dietsch, M. Heizmann, A. Nutz, C. Schätzle, C. Schilling, F. Schüssele, and A. Podelski, "Ultimate taipan: Trace abstraction and abstract interpretation - (competition contribution)," in *23rd International Conference, TACAS 2017, Proceedings, Part II*, ser. LNCS, vol. 10206, 2017, pp. 399–403. [Online]. Available: https://doi.org/10.1007/978-3-662-54580-5_31

[35] J. Henry, D. Monniaux, and M. Moy, "PAGAI: A path sensitive static analyser," *Electron. Notes Theor. Comput. Sci.*, vol. 289, pp. 15–25, 2012. [Online]. Available: https://doi.org/10.1016/j.entcs.2012.11.003

[36] X. Rival, "Understanding the origin of alarms in astrée," in *Static Analysis, 12th International Symposium, SAS 2005, Proceedings*, ser. LNCS, vol. 3672. Springer, 2005, pp. 303–319. [Online]. Available: https://doi.org/10.1007/11547662_21

[37] A. S. Dimovski and A. Legay, "Computing program reliability using forward-backward precondition analysis and model counting," in *Fundamental Approaches to Software Engineering - 23rd International Conference, FASE 2020, Proceedings*, ser. LNCS, vol. 12076. Springer, 2020, pp. 182–202. [Online]. Available: https://doi.org/10.1007/978-3-030-45234-6_9

[38] R. Alur, R. Bodík, G. Juniwal, M. M. K. Martin, M. Raghothaman, S. A. Seshia, R. Singh, A. Solar-Lezama, E. Torlak, and A. Udupa, "Syntax-guided synthesis," in *Formal Methods in Computer-Aided Design, FMCAD 2013*. IEEE, 2013, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/6679385/

[39] Y. Gu, T. Tsukada, and H. Unno, "Optimal CHC solving via termination proofs," *Proc. ACM Program. Lang.*, vol. 7, no. POPL, pp. 604–631, 2023. [Online]. Available: https://doi.org/10.1145/3571214

[40] T. A. Beyene, C. Popeea, and A. Rybalchenko, "Solving existentially quantified horn clauses," in *Computer Aided Verification - 25th International Conference, CAV 2013. Proceedings*, ser. LNCS, vol. 8044. Springer, 2013, pp. 869–882. [Online]. Available: https://doi.org/10.1007/978-3-642-39799-8_61

[41] T. Kuwahara, R. Sato, H. Unno, and N. Kobayashi, "Predicate abstraction and CEGAR for disproving termination of higher-order functional programs," in *Computer Aided Verification - 27th International Conference, CAV 2015, Proceedings, Part II*, ser. LNCS, vol. 9207. Springer, 2015, pp. 287–303. [Online]. Available: https://doi.org/10.1007/978-3-319-21668-3_17

[42] A. Gurfinkel, T. Kahsai, A. Komuravelli, and J. A. Navas, "The seahorn verification framework," in *Computer Aided Verification - 27th International Conference, CAV 2015, Proceedings, Part I*, ser. LNCS, vol. 9206. Springer, 2015, pp. 343–361. [Online]. Available: https://doi.org/10.1007/978-3-319-21690-4_20

[43] T. Kahsai, P. Rümmer, H. Sanchez, and M. Schäf, "Jayhorn: A framework for verifying java programs," in *Computer Aided Verification - 28th International Conference, CAV 2016, Proceedings, Part I*, ser. LNCS, vol. 9779. Springer, 2016, pp. 352–358. [Online]. Available: https://doi.org/10.1007/978-3-319-41528-4_19

[44] K. Hashimoto and H. Unno, "Refinement type inference via horn constraint optimization," in *Static Analysis - 22nd International Symposium, SAS 2015, Saint-Malo, France, September 9-11, 2015, Proceedings*, ser. Lecture Notes in Computer Science, vol. 9291. Springer, 2015, pp. 199–216. [Online]. Available: https://doi.org/10.1007/978-3-662-48288-9_12

[45] J. Leroux, P. Rümmer, and P. Subotic, "Guiding craig interpolation with domain-specific abstractions," *Acta Informatica*, vol. 53, no. 4, pp. 387–424, 2016. [Online]. Available: https://doi.org/10.1007/s00236-015-0236-z

[46] T. A. Beyene, S. Chaudhuri, C. Popeea, and A. Rybalchenko, "A constraint-based approach to solving games on infinite graphs," in *The 41st Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL '14*. ACM, 2014, pp. 221–234. [Online]. Available: https://doi.org/10.1145/2535838.2535860

[47] J. Kim, Q. Hu, L. D'Antoni, and T. W. Reps, "Semantics-guided synthesis," *Proc. ACM Program. Lang.*, vol. 5, no. POPL, pp. 1–32, 2021. [Online]. Available: https://doi.org/10.1145/3434311

[48] K. J. C. Johnson, A. Reynolds, T. W. Reps, and L. D'Antoni, "The semgus toolkit," in *Computer Aided Verification - 36th International Conference, CAV 2024, Proceedings, Part III*, ser. LNCS, vol. 14683. Springer, 2024, pp. 27–40. [Online]. Available: https://doi.org/10.1007/978-3-031-65633-0_2

[49] S. So and H. Oh, "Synthesizing imperative programs from examples guided by static analysis," in *Static Analysis - 24th International Symposium, SAS 2017, Proceedings*, ser. LNCS, vol. 10422. Springer, 2017, pp. 364–381. [Online]. Available: https://doi.org/10.1007/978-3-319-66706-5_18

[50] A. S. Dimovski, "Imperative program synthesis by abstract static analysis and SMT mutations," in *Proceedings of the 24th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences, GPCE 2025, Bergen, Norway, July 3-4, 2025*. ACM, 2025, pp. 27–40. [Online]. Available: https://doi.org/10.1145/3742876.3742884

[51] A. Dimovski, "Mutation-based lifted repair of software product lines," in *38th European Conference on Object-Oriented Programming, ECOOP 2024*, ser. LIPIcs, vol. 313. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, pp. 12:1–12:24. [Online]. Available: https://doi.org/10.4230/LIPIcs.ECOOP.2024.12

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 531–542
DOI: 10.15439/2025F5613
ISSN 2300-5963 ACSIS, Vol. 43

Machine-Readable by Design: Language Specifications as the Key to Integrating LLMs into Industrial Tools

Alexander Fischer*, Louis Burk*, Ramin Tavakoli Kolagari, Uwe Wienkop
0009-0001-1737-7395
0009-0003-0727-6456
0000-0002-7470-3767
0009-0000-4487-2458
Nuremberg Institute of Technology,
Faculty of Computer Science,
Kesslerplatz 12,
90489 Nuremberg, Germany
Email: {a.fischer, louis.burk, ramin.tavakolikolagari, uwe.wienkop}@th-nuernberg.de

***Abstract*—We propose a meta-language-based approach enabling Large Language Models (LLMs) to reliably generate structured, machine-readable artifacts referred to as Meta-Language-defined Structures (MLDS) adapted to domain requirements, without adhering strictly to standard formats like JSON or XML. By embedding explicit schema instructions within prompts, we evaluated the method across diverse use cases, including automated Virtual Reality environment generation and automotive security modeling. Our experiments demonstrate that the meta-language approach significantly improves LLM-generated structure compliance, with an 88% validation rate across 132 test scenarios. Compared to traditional methods using LangChain and Pydantic, our MLDS method reduces setup complexity by approximately 80%, despite a marginally higher error rate. Furthermore, the MLDS artifacts produced were easily editable, enabling rapid iterative refinement. This flexibility greatly alleviates the "blank page syndrome" by providing structured initial artifacts suitable for immediate use or further human enhancement, making our approach highly practical for rapid prototyping and integration into complex industrial workflows.**

I. Introduction

INDUSTRIAL tools have become increasingly powerful and can now handle vast amounts of data. However, as their range of functions continues to grow, operating these tools often overshadows the core task of content or problem management. This situation is further exacerbated by the common "blank page syndrome," [1], a common phenomenon in complex design or modeling tasks, where the absence of any initial structure or guidance leads to hesitation, inefficiency, or even creative paralysis. In such cases, users face difficulty starting from scratch, especially when formal specifications or architectural expertise are required. This not only slows down the initial design process but may also prevent domain experts from effectively contributing to model-driven workflows. Consequently, the focus moves away from the substantive content toward the tool itself, lowering productivity and widening the gap between content and tooling

At the same time, LLMs offer new opportunities by helping domain experts generate creative impulses and mitigate the "blank page syndrome." These assisting tools have evolved rapidly—especially through the integration of AI capabilities—that they have become even more complex, making proper integration of LLMs into content creation workflows difficult. In particular, when domain-specific languages are used that deviate from common formats like JSON or XML, producing reliably machine-readable artifacts remains a significant challenge.

Our approach addresses this issue by defining a language specification (meta-language) that is adapted to the requirements of each domain, and which likewise does not rely on existing standard formats. This enables us to capture the structural differences and specific relationships of a domain. By combining LLMs with this meta-language, it becomes possible to produce customized, machine-readable artifacts without requiring domain experts to possess extensive knowledge of architectural or formal specifications.

To clarify our terminology, we define two key concepts central to this paper. A Meta-Language-defined Structure (MLDS) refers to a structured, machine-readable artifact that is generated based on natural language input, using a set of schema rules defined in a meta-language which we call Meta-Language-defined Structure Instructions (MLDSI) . The meta-language itself is a domain-specific specification that defines the structure of the expected artifact. It serves as a structural blueprint that guides the generation of MLDS artifacts and ensures that the resulting outputs conform to the expected semantics of the target domain.

Although recent approaches have explored how LLMs can be guided to generate structured outputs, they typically rely on rigid, domain-agnostic formats such as JSON or XML. These

*Both authors contributed equally to this work.

solutions offer limited semantic adaptability and often require specialized tools or substantial manual intervention. This highlights a persistent research gap: the lack of a lightweight, domain-adaptable framework for generating structured outputs from natural language in a way that is both semantically meaningful and technically accessible. Our work addresses this gap by introducing the MLDS approach, which enables domain experts to produce customized, machine-readable content with minimal setup effort and without requiring expertise in schema tooling or formal languages.

Whether this meta-language approach is viable and beneficial is examined by focusing on the following central question:

What conditions must be met for meta-languages to function as a bridging element between natural content descriptions and machine-readable structures, enabling the integration of LLMs into industrial toolchains?

This central question is addressed through the following three sub-questions:

- **R1:** Which methods can be used to incorporate meta-language definitions into a model?
- **R2:** What criteria can be considered to validate the output?
- **R3:** To what extent can reference examples be used to adapt the output to specific needs?

To ensure transparency and facilitate reproducibility, all relevant research artifacts, including evaluation results, domain-specific meta-language definitions and generated Meta-Language-defined Structures (MLDS) artifacts, have been made publicly available in our GitHub repository [2].

In the remainder of this paper, we present the technical background and motivation for our meta-language approach in Section II. An overview of related research follows in Section III, which positions our work within the broader field. Next, Section IV introduces our proposed methodology and the design of the meta-language. We then provide a comprehensive evaluation in Section V, showcasing its applicability across various domains. Section VI discusses our key findings and highlights potential limitations, while Section VII outlines promising avenues for future research. Finally, Section VIII concludes the paper with a summary of our main contributions.

II. Background and Motivation

In many technical and economic fields, there are specialized challenges that can only be successfully addressed with the help of experts. This is especially true for the development and implementation of complex software and system architectures [3]. A classic example is the collaboration between a person with architectural expertise (Architectural Domain Expert, ADE) and someone with substantive or specialized expertise (Content Domain Expert, CDE). While an architect generally possesses deep knowledge of a specific, clearly defined domain, the transfer of architectural concepts to other domains is often difficult and time-consuming. This hurdle complicates efficient collaboration and leads to practical inefficiencies when the same fundamental principle is applied in different fields.

The goal, therefore, is to allow CDEs in a given domain to create structured artifacts directly processable by machines, even without extensive knowledge of software or system architecture. These artifacts, which we refer to as MLDS, make it possible to model specialized requirements in detail. By using MLDS, ideas and requirements expressed in natural language are transformed into formal structures that adhere to certain rules and constraints. Consequently, the ADE's role shifts to a higher-level task: defining a structural specification (the "meta-language") that ensures the resulting MLDS meet the desired quality and formatting requirements.

Currently, structured outputs from natural language models, such as JSON-based responses, typically require specialized tools and significant manual intervention. Tools like Pydantic and LangChain have emerged to facilitate structured outputs by validating and enforcing schemas, thereby ensuring JSON structures are accurately generated and adhered to [4], [5]. However, effective use of these tools demands substantial technical expertise and familiarity with their underlying frameworks [6]. Moreover, these tools require ongoing adaptation to handle differences in how newer LLM architectures generate structured output, which introduces additional maintenance overhead and complexity.

Furthermore, not all LLMs inherently support structured output generation effectively. This limitation restricts their applicability to scenarios requiring strictly formatted outputs [7], [8]. For models that do support structured outputs, outputs often need post-processing and manual corrections due to subtle inconsistencies, ambiguities, or deviations from predefined structures, increasing complexity and effort [9].

To circumvent these challenges, our approach introduces MLDS Instructions—structured prompts explicitly designed to guide LLMs in generating precisely defined structured artifacts. LLMs are powerful models that can interpret and produce natural language [8]. However, a key issue arises because LLMs generally do not generate precise, user-defined structures unless these are already standard formats. This is precisely where our solution comes into play: with so-called "MLDS instructions," we provide the LLM with a meta-language characterized by a clear structure and defined relationships, without relying on existing standard formats.

The MLDS instructions inform the model about both the desired output format and the rules needed to produce it. This distinguishes our approach from classic procedures that require extensive fine-tuning of the model or force the output into a fixed format. Through the meta-language and its meta-levels, complex relationships within the MLDS can also be represented. In this way, MLDS builds a bridge between the conceptual problem statements formulated by CDE and a strictly formatted output that can be processed by machines.

Our approach is inspired by established metamodeling architectures, such as those found in the Meta-Object Facility (MOF) [10] and the Unified Modeling Language (UML) [11], and is structured into four layers:

- At the M0 level, the focus is on concrete data or instance-level elements—the real-world objects or executions that

result from applying the MLDS. This level is where the model "comes to life" and is directly processed by software systems.

- The M1 level, known as the model level, is where the MLDS artifacts reside. These artifacts are formal representations that capture the concrete, structured requirements of a domain, serving as blueprints derived from the natural language input provided by domain experts.
- The M2 level, or metamodel level, contains the meta-language. This layer defines the rules, constraints, and abstract syntax and semantics required for constructing valid MLDS artifacts. The MLDS instructions used to guide the LLM in generating structured outputs are derived from this meta-language.
- At the highest M3 level, the meta-meta model is defined. This layer establishes the fundamental constructs and abstractions upon which the meta-language is built. Although users do not normally interact directly with this level, it is necessary to ensure that the meta-language (M2) is consistent, which in turn ensures that the rules governing the MLDS (M1) are sound.

By organizing our approach into these four interconnected layers, M0 for instances, M1 for MLDS artifacts, M2 for the meta-language, and M3 for the meta-meta model, we create a structured, hierarchical approach that ensures domain-specific requirements are transformed consistently into machine-processable formats. While content domain experts focus on formulating concrete requirements at the M1 level, the meta-language at the M2 level ensures that these requirements are implemented within a consistent, formal scope. The M3 layer supports the whole system by providing a basis for the meta-language.

In this paper, we will illustrate the functionality of our MLDS instructions and the resulting MLDS outputs using two application scenarios: (1) Automotive Software Security and (2) Virtual Reality in business applications. By "Automotive Software Security," we mean the specific security and reliability requirements in vehicle software engineering, while "Virtual Reality in business applications" refers to immersive and interactive digital learning and work environments used within companies. These examples highlight how varied the domains can be in which experts (CDE)—for instance, from the security or VR fields—wish to express their specialized requirements in MLDS, without being trained as architects in the sense of software system architecture.

An illustrative use case demonstrates this: a VR specialist wants to model a virtual learning space for internal corporate training. Although this expert (CDE) has a clear understanding of the requirements for the learning space, they lack the experience to describe it in a precise architectural form. The existing tool for creating VR environments requires a clearly defined structure in the form of MLDS. Until now, it was the architect's job to convert these specialized requirements into the corresponding formal structure. With our approach, a natural language description of the desired learning space and the MLDS instructions satisfy for an LLM to generate the appropriate MLDS. Thanks to the MLDS instructions, the LLM knows exactly how the resulting artifact should be structured. As a result, the CDE's specialized requirements are transformed into a formally processable data format, without the need for the CDE to possess deeper knowledge of software architecture or formal languages.

III. Related Work

In our approach, we propose a meta-language framework that allows the creation of formally defined structures from purely natural-language descriptions *without* relying on standard formats like JSON or UML, nor requiring extensive fine-tuning.

A variety of prior studies already employ LLMs for structured output generation, yet they pursue either *more specialized* or *fundamentally different* techniques. For instance, Wang et al. [12] demonstrate a *grammar prompting* procedure, where LLMs are constrained by a Backus–Naur Form (BNF) grammar to produce strictly formal outputs. While this approach works very well for specific languages that fit into a context-free grammar, it often requires explicit modifications to the decoding process (such as restricting token choices) or relies on well-defined, existing grammar rules. By contrast, our focus is on designing a *flexible meta-language* that need not be encoded in BNF.

A related but narrower method is illustrated by Scholak et al. [13] with their *PICARD* approach, where an autoregressive model is incrementally validated against the desired language (e. g., SQL). Similar to grammar prompting, token decoding is continuously constrained during generation. However, in contrast to our meta-language paradigm, PICARD involves *invasive modifications* to the decoding algorithm. We rely instead on clearly defined meta-language instructions and LLM prompt processing, yielding more latitude for *language structures*.

Further research in retrieval augmentation or fine-tuning comes from Bassamzadeh and Methani [14], comparing whether structured code is best produced via specialized fine-tuning or optimized prompting techniques. While these findings are valuable for reducing errors and hallucinations, they typically assume a *known* format (e. g. JSON/YAML). In our method, no specialized model training is needed; we *generate* all target structures solely via prompt-based instructions and meta-language rules.

In the context of model-driven software development, Netz et al. [15] show how UML-like class diagrams (CD4A) can be generated by GPT-3.5/GPT-4 and transformed directly into complete web applications. However, they rely on *UML notations* and existing generator frameworks (e. g. MontiGem). This differs from our approach, which does *not* require any model-driven development toolchain or standardized notation. Instead, we allow arbitrary domain-specific structures—for instance, automotive security or VR scenarios—specified with a freely definable meta-language.

Key Contribution of Our Approach: While the mentioned studies employ various forms of grammar-based decoding,

fine-tuning, or predefined formats, our method provides a universal meta-language definition. Thus, architects or tool developers can prescribe any target structure (from security modeling to 3D scenes) at the meta-level so that an LLM subsequently outputs *machine-readable* artifacts in exactly that format–without training, without invasive decoder modifications, and without restricting to JSON/UML. Because our approach does not fix a specific output format, it supports both machine-driven and human refinements of the same structure, fostering an iterative co-creation process. Domain experts bring in new scenario descriptions, the LLM generates or extends the MLDS, and humans can then fine-tune or correct the result in the meta-language immediately. This openness, combined with a clear metamodel architecture, constitutes the central distinguishing feature of our work.

IV. Methodology

In this Section, we explain our proposed approach (Section II) in detail, showing how we use our custom meta-language together with the MLDS framework to solve the problem. We begin by outlining how the meta-language is derived from the tool's requirements, providing a starting point for the process. Next, we describe how to structure the descriptions of tasks and scenarios, setting up a foundation for generating the MLDS using LLMs. The complete procedure is shown in Figure 1.

To keep the process transparent, we provide examples that demonstrate how prompts are combined with the meta-language. A notable feature of our approach is that it allows users to influence different parts of the MLDS in various ways, making the method adaptable. After fully explaining the MLDS, we show how it can be directly imported into tools, and we conclude by demonstrating how this method addresses the blank page syndrome.

A. Conception of the Meta-Language

The core of our approach lies in developing a meta-language that incorporates the specific requirements of a domain or tool and organizes them into a structured form that is not bound to common standards. In contrast to JSON or XML, the meta-language may deliberately allow specialized object types, relationships, and value ranges that thoroughly reflect the peculiarities of a domain. The meta-language thus pursues two primary objectives: on the one hand, it provides the LLM with explicit guidelines on which structures, attributes, and relationships should appear in the generated MLDS. On the other hand, it is designed to be easily understandable and extendable by human domain experts, allowing them to add new elements or rules when necessary.

1) Reasons for a Meta-Language: One of the main challenges in automatically generating structured output is that LLMs initially lack any knowledge of custom formats. Although models can be adapted to standard formats, they often reach their limits when specialized domains require formats that differ from these standards or go beyond their semantics. Many advanced tools employ proprietary definitions that cannot readily be mapped onto common standards. Our meta-language addresses this gap by:

- Allowing freely definable object and relationship types to represent specific domain concepts.
- Defining clearly delimited value ranges (e.g., discrete selection options, predefined numerical ranges, or boolean variables) for each attribute, so that the LLM is not forced to "improvise" unnecessarily but instead chooses from a set of valid options.
- Providing structural guidelines that define the "skeleton" of an MLDS, ensuring that the generated output follows a meaningful and machine-readable architecture.

In this way, it is not necessary to extensively fine-tune the LLM. Instead, we present the model with precise rules via the meta-language and corresponding prompt instructions, which it follows when generating the MLDS.

2) Analysis of the Required Input: Before implementing the meta-language, it is necessary to determine which information a particular tool or domain actually needs. A standard example is a VR development tool that manages various object types (e.g., *rooms*, *avatars*, or *interactions*), each with specific properties (size, access rights, animation logic, etc.). These aspects are collected, categorized, and eventually transferred into the meta-language. On this basis, one can also estimate the relationships between objects—an *avatar*, for instance, may be assigned to a *room*, and an *interaction* may be tied to an *object*.

3) Extracting Relevant Objects and Properties: Once the domain analysis is complete, relevant objects are defined. An *object* corresponds to a central concept in the domain (e.g., learning environment, avatar, security component). A *learning environment*, for example, might have properties such as *name*, *maximum number of users*, or *room configuration*. Each property is assigned a corresponding value range (e.g., a fixed set of room types, numerical ranges for the number of users) to ensure consistency and comprehensibility.

4) Defining Relations: Another key to understanding and modeling complex relationships is specifying the links between objects. For instance, a *human actor* object at a certain *resilience level* could refer to an *attack* object, or multiple *vehicle features* could form a single item. These relationships may be configured hierarchically within the meta-language or refer to the same level, each implying different semantic connections. Such a relational network dictates the subsequent structure of the MLDS, guiding the LLM in how individual elements should interrelate.

5) Domain-Specific Meta-Language Instructions: In the following, we present three excerpts from a possible domain-specific meta-language design, illustrating its key concepts. Although the example involves a 3D scene for VR applications, the same underlying mechanisms can be applied to virtually any domain.

a) Defining the Scene Structure: The first step is to establish a top-level object that acts as a global container for all subsequent elements (Listing 1). In the VR domain, we call

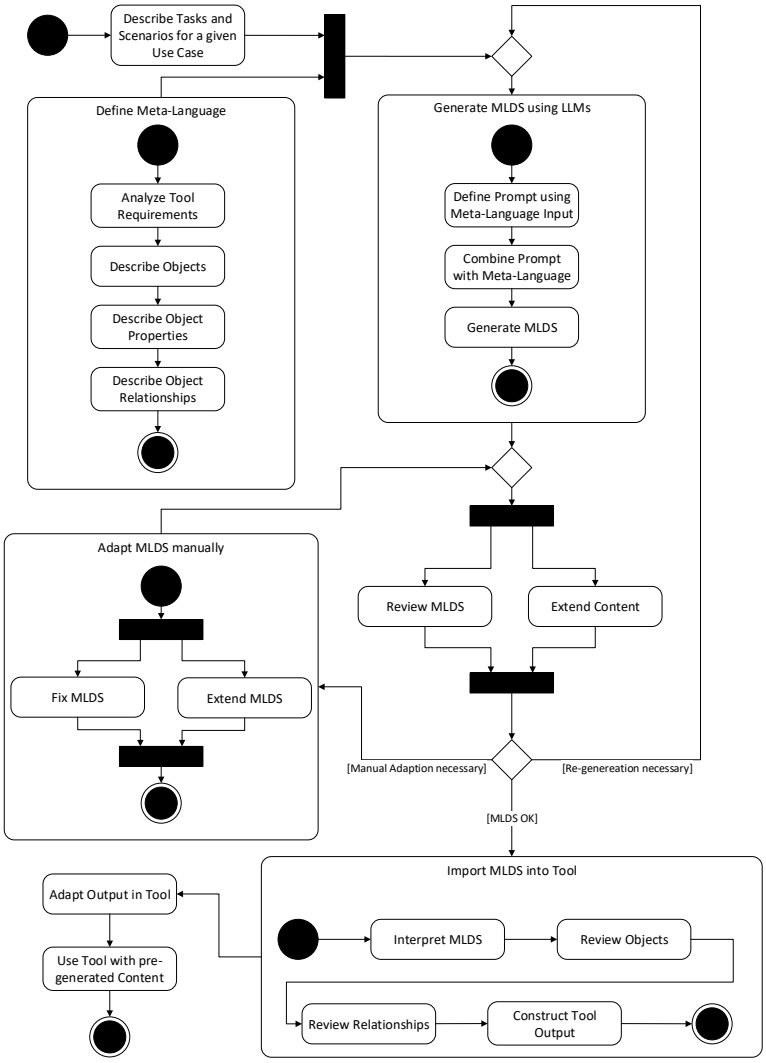


Fig. 1. Workflow for generating and refining MLDS from natural language input, combining meta-language specification, LLM-based structure generation, and iterative human-AI co-editing for tool integration.

this object `scene`, which holds basic environmental details and a list of objects.

```
{
  "scene":
  {
    "sceneName": "VR Training Room",
    "environment":
    {
      "type": "indoor",
      "dimensions": { "width": 10.0, "
          height": 3.0, "depth": 15.0 }
    },
    "objects":
    [
      /* Objects go here */
    ]
  }
}
```

Listing 1. Definition of the Scene Structure

In other domains, what is here called `"scene"` could be labeled `"system"` or `"project"` and might include completely different properties (e.g., `systemCategory` or `budget`). Nevertheless, the principle remains the same: a clearly defined parent object provides the framework in which all detailed elements and their attributes are nested.

b) Modeling Objects and Their Relationships: Each object has a unique `objectId` and is characterized by properties (e.g., `objectType`, `position`, `rotation`) as well as optional relational information. Listing 2 shows how to place an object (a table) relative to another object using `relativePositioning`.

```
{
  "objectId": "object_01",
  "objectType": "table",
  "position": { "x": 2, "y": 0, "z": 3 },
  "rotation": { "x": 0, "y": 45, "z": 0
      },
  "dimensions": { "width": 1.5, "height":
       0.75, "depth": 1.0 },
  "relativePositioning":
  {
    "referenceObject": "object_02",
    "relation": "next_to"
  },
  "offset": { "x": 0.2, "y": 0, "z": 0 }
}
```

Listing 2. Objects with Relative Positioning

In other domains, objects may have entirely different properties (for instance, `memoryLimit` or `location` in a cloud application). Similarly, a relationship like `next_to` could be replaced by something like `depends_on` if you want to describe software dependencies. By specifying such structures in a meta-language, you ensure that the LLM knows exactly which attributes exist and which values are permissible.

c) Hierarchical Nesting of Objects: To describe more complex scenarios, objects can themselves have *children* (Listing 3). This enables a hierarchical organization with subordinate elements.

```
{
  "objectId": "object_02",
  "objectType": "shelf",
  "relativePositioning":
  {
    "referenceObject": "object_03",
    "relation": "on_top_of"
  },
  "children":
  [
    {
      "objectId": "object_02_child_01",
      "objectType": "box",
      "relativePositioning": {
        "referenceObject": "object_02",
        "relation": "next_to"
      },
      "offset": { "x": 0.1, "y": 0, "z": 0
          }
    }
  ]
}
```

Listing 3. Object Hierarchies Using Children

This type of `children` property could represent, for example, submodules within a larger module in a software architecture or smaller processes within a bigger industrial workflow. Likewise, a relationship such as `"on_top_of"` is just an example of a hierarchical link; in other contexts, it might be `"contained_in"` or `"extends"`.

B. Generation of the MLDS

Once the meta-language is established, natural language descriptions or scenarios adapted to a specific use case can be created. As illustrated in Figure 1, the meta-language is developed as a separate module that, once finalized, can be reused for all subsequent generations.

In the most compact workflow, we concatenate the meta-language and the scenario into one composite prompt and pass it to a pretrained LLM via its API. Because the model already possesses broad world knowledge, it needs no task-specific fine-tuning; instead, the embedded meta-language rules channel the model towards a syntactically correct MLDS while the scenario text supplies the domain content. A prompt segment might therefore read:

```
<<META LANGUAGE>>
object room { width:float height:float
    depth:float ... }
...
<<SCENARIO>>
Create a training classroom with 20
    student desks, one teacher desk,
and a projector mounted to the ceiling.
```

Listing 4. Excerpt of a single-shot prompt that merges meta-language and scenario

For projects that demand stricter traceability or staged quality control, we separate the process into two calls. The first call feeds nothing but the meta-language to the LLM and receives back a condensed, token-optimised representation, the second call injects the scenario plus the condensed rules. This split makes it easy to inspect intermediate artifacts, attach unit tests to the generated schema, or swap in a different scenario without retransmitting an unchanged meta-language, an advantage when prompt length becomes a bottleneck.

Regardless of the single- or multi-step variant, the raw MLDS is fed through an automatic validator that flags missing attributes, illegal value ranges or broken references. Only after the MLDS passes this gatekeeper is it handed to the target tool, guaranteeing that the downstream pipeline never receives malformed input.

C. Adaption of the MLDS

In practice, generating the initial MLDS is only the first step in its full utilization. Because the MLDS can be produced in a format that is easily readable by humans, domain experts can readily validate its contents without a steep learning curve. Once generated, the MLDS may be scrutinized for correctness, and any missing information can be added. As illustrated in Figure 1, these iterative improvements can be carried out by the LLM itself, which can respond to feedback or correction prompts. At the same time, domain experts remain free to make manual adjustments, ensuring a high degree of precision where AI suggestions may not capture every nuance. This dual approach of automated and manual editing fosters a co-creative process, merging AI-driven speed and consistency with human expertise. Moreover, the MLDS's structured nature facilitates intuitive graphical interfaces, allowing content specialists to modify or enrich the MLDS without delving into the underlying code. By accommodating both fully automated and personalized interventions, the workflow remains flexible, transparent, and accessible to all involved stakeholders.

D. Integration into the Tool Environment

Once the finalized version of the MLDS is ready, it can be loaded directly into the target tool. The tool interprets the generated structure and creates the required output, representing all defined objects and their relationships according to the MLDS specifications. Additional modifications and enhancements can then be carried out directly within the tool, giving users the flexibility to refine or extend the setup as needed. This ensures that the produced content is fully leveraged, while minimizing both development effort and potential errors.

V. Evaluation

A. Objective and Approach

To demonstrate the developed Meta-Language approach, we selected two distinct domains: (1) the automated generation of Virtual Reality (VR) worlds and (2) the Security Modeling of embedded systems in the automotive context. The aim is to show that from a defined meta-language, a domain-specific MLDS can be generated that consistently follows the given specification. Furthermore, we highlight how this method mitigates the so-called "blank page syndrome," where a completely unstructured starting point often leads to high efforts in initial model or environment creation. By providing a machine-readable foundation, users can directly refine and adapt the generated MLDS artifacts.

B. Use Case: Automated VR World Generation

The first domain focuses on automatically generating VR worlds in Unity. A structured description is created to form a basic layout of the virtual environment. Designers provide a natural-language prompt (for instance: "Generate a classroom for 20 students"), which is then transformed into a JSON-based MLDS using our meta-language. Unity processes the resulting MLDS to position objects and define relationships in the 3D scene. Initially, placeholders (e.g., cubes) are placed to help visualize object arrangements. Designers can later replace these placeholders with final assets, ensuring a quick start for prototyping. This effectively reduces the blank page syndrome since each new scene already has a predefined structure.

Over 200 test scenarios (see Section IX) were conducted, demonstrating that LLMs are able to generate coherent object hierarchies and spatial relationships. For example, Listing 5 shows a parent-child relationship where a desk acts as the parent object, and a chair is placed relative to the desk. Positioning attributes and relational references (e.g., `in_front_of_positive_z`) are used to automate the placement in the 3D environment.

```
{
  "objectId": "studentDesk_r0_c0",
  "objectType": "student_desk",
  "assetName": "student_desk_asset",
  "position": {"x": -6, "y": 0.5, "z":
      8},
  "dimensions": {"width": 2, "height": 1
      ,"depth": 2},
  "group": "furniture",
  "children":
  [
    {
      "objectId": "studentChair_r0_c0",
      "objectType": "chair",
      "assetName": "student_chair_asset",
      "relativePositioning": {
        "referenceObject": "
            studentDesk_r0_c0",
        "relation": "in_front_of_positive_z
            ",
        "distance": 1
      },
      "dimensions": {"width": 0.5, "height
          ": 0.5, "depth": 0.75},
      "group": "furniture"
    }
  ]
}
```

Listing 5. VR Parent-Child Relation

By incorporating such parent-child relations, users benefit from an iterative co-creation process: the generated foundation can be refined either through extended prompts (in natural language) or direct modifications of the MLDS output.

C. Use Case: Security Modeling in Automotive

The second use case addresses Security Modeling for embedded systems in the automotive context. Building upon the metamodel proposed in [16], we define a meta-language for an MLDS that helps create a foundational security model. Security experts or systems engineers can outline potential threats or vulnerabilities in natural language (e.g., "unsecured wireless communication" or "man-in-the-middle attack"). An LLM then translates these descriptions into the MLDS, which is directly processable by a security modeling tool. Such a tool subsequently calculates risk levels or highlights vulnerabilities. Manual adjustments remain possible at any stage, ensuring a flexible co-creation process.

Listing 6 illustrates a JSON-based MLDS snippet with typical attributes for attack characterization, such as `AttackComplexity` and `UserInteraction`. By assigning initial values to these fields, the LLM provides a starting point for further risk assessments.

```
{
  "Name": "Man-in-the-middle:
      intercepting and manipulating
      wireless traffic",
  "AccessRequired": "Network",
  "AttackComplexity": "(H)igh",
  "PrivilegesRequired": "(N)one",
  "Urgency": "(L)ow",
  "UserInteraction": "(N)one",
  "AvailabilityImpact": "(L)ow",
  "ConfidentialityImpact": "(H)igh",
  "IntegrityImpact": "(L)ow",
  "SafetyRelevance": "(H)igh",
  "vulnerabilities":
  [
    {
      "Name": "Unsecured Wireless
          Communication Protocol"
```

```
16        }
17    ]
18 }
```

Listing 6. Automotive Security Attack Representation

Since risk or vulnerability scores are computed within the modeling tool itself, attributes assigned by the LLM can be refined later. This aligns with the cooperative design principle, where the model's first version is automatically generated and then iteratively improved.

D. Results and Insights

In both domains, we combined the language specification with natural-language prompts across several LLMs (including GPT 4o mini, GPT o3 mini, DeepSeek-R1, and DeepSeek-R1 14B running locally). Most generated MLDS files were syntactically correct and machine-readable. Occasional issues arose mainly from incomplete API responses or minor inaccuracies in the meta-language. Notably, even complex problem descriptions in natural language led to well-structured and valid MLDS artifacts. Key findings are:

- **High Machine-Readability:** The LLM outputs conform closely to the defined meta-language, minimizing parsing errors.
- **Blank Page Syndrome Mitigation:** In both VR world creation and security modeling, users start with a usable blueprint rather than a completely empty project.
- **Ease of Extension:** Generated models or VR scenes can be expanded using extended prompts or direct modifications.
- **Domain-Agnostic Feasibility:** The approach works effectively in multiple domains, provided a domain-specific meta-language is well-defined—for example, structuring content in education, automating routines in smart environments, modeling workflows in business, formalizing scenarios in healthcare, or generating logic in games and legal contexts.

We further evaluated our approach by systematically analyzing 132 generated MLDS files to quantify adherence to the prescribed meta-language schema. This analysis was automated through a custom Python-based parser validating each file against the defined schema.

E. Evaluation Setup

The evaluation dataset consisted of files categorized into two complexity levels: *easy* (n=72) and *medium* (n=60), selected from the VR dataset based on the required content complexity. The dataset aimed to enable an end-to-end pipeline for generating virtual environments. Furthermore, each complexity level had two variations: one employing semantic validators (*withValidator*) and one without validators (*withoutValidator*). These validators were applied downstream of the MLDS translation process and specifically assessed semantic correctness, such as checking for overlapping objects, rather than syntactic accuracy. As our analysis indicated no significant differences in syntactic correctness between these variations, we decided to merge them into a single unified dataset for evaluation purposes. Each file was evaluated based on structural validity, completeness of schema components, and the presence of specific parsing errors.

We defined the metrics as follows:

$$V = \sum \text{valid files}, \quad I = N - V \tag{1}$$

$$\text{Validity Rate } (\%) = \frac{V}{N} \times 100 \tag{2}$$

$$\text{Presence Rate } (\%) = \frac{\text{Files containing schema part}}{N} \times 100 \tag{3}$$

where N is the total number of evaluated files, V is the number of valid (correctly structured) files, and I is the number of invalid (incorrect) files.

F. Quantitative Results

Out of $N = 132$ files, $V = 116$ (87.9%) were valid, while $I = 16$ (12.1%) contained parsing errors (Fig. 2). Errors primarily arose from JSON parsing issues due to incorrect string formatting.

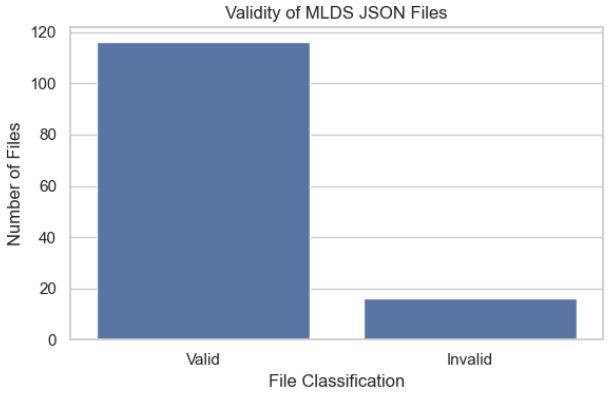


Fig. 2. Distribution of valid vs. invalid MLDS JSON files.

Out of the 16 invalid files, 15 exhibited a parsing error indicated by the message 'str' object has no attribute 'get'. This error occurred because the JSON parser expected a structured object but encountered a string instead. Only one file showed a different issue, specifically missing essential schema components: *environment* and *sceneName*.

Table I summarizes the presence rates of schema parts. Most elements showed high presence rates, indicating strong schema adherence, except for optional fields like `env.background`, which appeared less frequently (72.7%).

TABLE I
SCHEMA PART PRESENCE SUMMARY

Schema Part	Presence	Missing files
sceneName	87.9%	16
environment	87.9%	16
env.type	87.9%	16
env.dimensions	87.9%	16
env.lighting	85.6%	19
env.background	72.7%	36
objectGroups	87.9%	16
objects	88.6%	15
obj.objectId	87.9%	16
obj.objectType	87.1%	17
obj.position	87.9%	16
obj.rotation	87.1%	17
obj.dimensions	87.1%	17

Since we did not explicitly request the inclusion of background components and instead relied on the LLM's initiative to incorporate them as part of the MLDS, the observed number of omissions was anticipated. This could be improved by explicitly addressing background components within the input prompt, thus providing clearer guidance to the LLM.

Figure 3 visually presents these results and highlights areas for possible improvements.

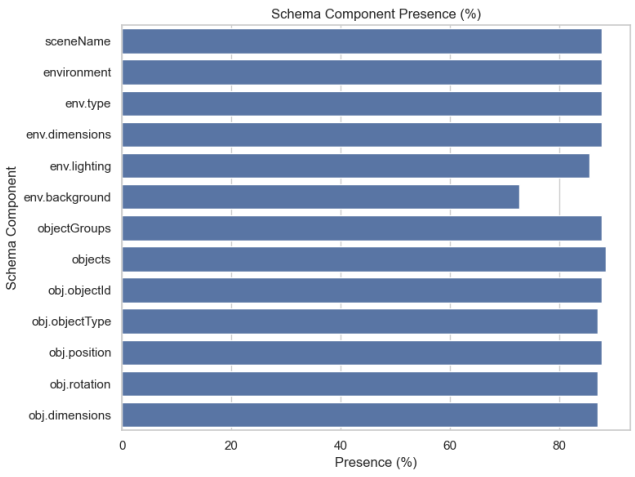


Fig. 3. Presence rates of schema parts in MLDS files.

Figures 4 and 5 highlight the complexity of generated files. Most scenarios had a moderate number of objects, suitable for realistic industrial applications, while nesting depth indicated well-structured JSON.

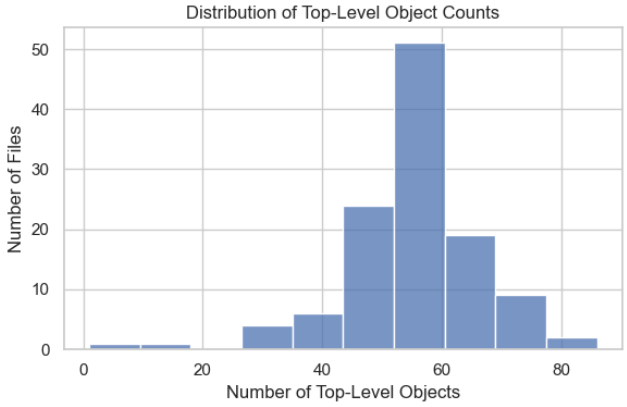


Fig. 4. Distribution of object counts per MLDS file.

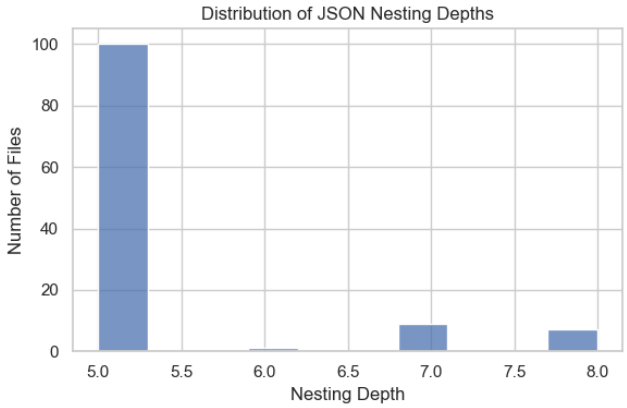


Fig. 5. Distribution of nesting depths in MLDS files.

The nesting depth was computed with the following Python function:

```
def compute_depth(obj, depth=0):
    if isinstance(obj, dict):
        return max([compute_depth(v, depth
            +1) for v in obj.values()] or [
            depth])
    if isinstance(obj, list):
        return max([compute_depth(i, depth
            +1) for i in obj] or [depth])
    return depth
```

Listing 7. Computing JSON nesting depth

G. Comparison with Traditional Approaches

We compared our MLDS approach directly against a traditional structured-output pipeline using LangChain and Pydantic [5], [4]. The implementation of the traditional approach took approximately five times longer than preparing our MLDS instructions, primarily due to the complexity and steep learning curve associated with LangChain and Pydantic. While the traditional approach consistently produced compilable results when successfully executed, approximately 7%

of pipeline runs failed entirely due to validation errors or incomplete responses. In contrast, our MLDS-based approach resulted in a slightly higher error rate (around 6% more errors), but significantly reduced the overall complexity, effort, and required expertise. Therefore, especially for rapid prototyping, iterative adjustments, or when specific expertise in Pydantic and LangChain is not readily available, our MLDS approach offers a clear practical advantage.

H. Implications

Explicitly embedding meta-language schema instructions within prompts enhances structural compliance and machine readability. This efficiency allows rapid prototyping and seamless integration into diverse industrial workflows, underscoring the practical effectiveness of our approach.

VI. Discussion

In this Section, we revisit the insights gained from our evaluation in light of the research questions posed in Section I. We then compare our approach to existing work (see Section III) and critically examine potential limitations as well as threats to validity.

A. Interpretation of Evaluation Results

1) Research Question R1: Methods for Incorporating Meta-Language Definitions into Models: Our experiments in the domains of Virtual Reality (VR) world generation and Automotive Security indicate that a well-specified meta-language can effectively convey structural and relational constraints to LLMs. The results suggest that this method scales to both creative (VR design) and safety-critical (automotive security) fields. We observed that generating complex MLDS artifacts worked particularly well when the meta-language was sufficiently fine-grained and included specific object definitions and enumerated attribute values (e.g., *relativePositioning*). This confirms that incorporating meta-language definitions does not require adherence to standardized formats, so long as the underlying structure is clear and domain-relevant.

2) Research Question R2: Criteria for Validating the Generated Output: Our evaluation revealed that, on one hand, formal aspects such as syntax and mandatory attributes can be automatically validated. On the other hand, certain domain-specific or semantic aspects (e.g., the correctness of spatial relationships in VR scenarios or the accuracy of security attributes in automotive models) often benefit from human review. Thus, a purely syntactic validation is a valuable first step but does not eliminate the need for *semantic* validation. A cooperative feedback loop—where experts refine machine-generated MLDS artifacts and tools provide automated format checks—leverages the strengths of both human expertise and automated structure enforcement.

3) Research Question R3: Adapting Output via Reference Examples: Including examples (e.g., prototype *scene* definitions for VR environments or sample *attack* entities in security modeling) improved both the consistency and format of the generated outputs. The LLM relies on these "exemplary" structures to align with domain-specific expectations. Notably, references can demonstrate not only value ranges but also relational patterns (*children* objects, hierarchy levels, dependencies). Moreover, the use of *templates* allows for targeted adjustments of the desired model detail, for instance by distinguishing between *minimal* or *extended* output variants.

B. Comparison with Related Work

Meta-languages for architecture and modeling are well-established in research, typically appearing in the form of DSMLs (Domain-Specific Modeling Languages) or UML profiles. Our approach extends these ideas by leveraging LLMs: instead of conforming strictly to standardized outputs (e.g., UML diagrams or XML), we demonstrate that "custom" meta-languages can be reliably processed by LLMs, provided the rules are clearly specified. This addresses a gap often left by other methods: reliable *and* flexible structured modeling without extensive fine-tuning of the model or proprietary export layers.

C. Implications and Practical Benefits

Our findings suggest that the proposed MLDS approach offers a flexible means of translating domain-centric (natural-language) descriptions into formally processable artifacts, even for users lacking deep architecture knowledge. In both domains, the technique reduced the "blank page syndrome," as users begin with a ready-to-use framework rather than an empty project. At the same time, the generated MLDS remains open to subsequent modifications, either by further prompting or by direct model editing.

Looking ahead, this could reshape workflows in software architecture and engineering. Architects may focus on designing and maintaining the meta-language, while domain experts continue to think in natural-language scenarios. This division of responsibilities can foster more productive, model-based processes and raise the adoption of structured modeling methods.

D. Limitations and Threats to Validity

While our results are encouraging, some constraints remain:

- **Domain-Specific Complexity:** As domains grow more detailed, the meta-language itself can become large and complex. An insufficiently granular metamodel may produce vague or incomplete MLDS artifacts.
- **Dependence on Prompt Quality:** The clarity of the natural-language input significantly affects the generated output. Ambiguous or insufficiently detailed prompts may result in incomplete models.
- **Technical Limitations of LLMs:** Issues such as model confusion or output truncation can arise with especially lengthy or nested structures. While example MLDS patterns help mitigate these risks, LLMs are not foolproof.
- **Generalizability:** Although we explored two notably different domains (VR and automotive security), other domains could impose even more specialized requirements that warrant extended validation.

VII. Future Directions

The integration of LLMs and meta-languages opens up several research areas. One is the development of automated *quality checks* that address both syntactic and semantic coherence (e.g., inconsistencies in hierarchical relationships). Another is exploring multi-prompting strategies or incremental fine-tuning to further improve MLDS consistency. There is also potential value in establishing a cross-domain *vocabulary* for reusable structural elements (e.g., `children`, `parent`, `referenceObject`).

Lastly, a more collaborative platform could be developed, where architects design new meta-languages while domain experts contribute reference MLDS examples. Such a shared knowledge base could be continuously updated and utilized by LLMs, enabling more efficient and accurate generation of structured domain artifacts.

VIII. Conclusion

We presented a domain-adaptable method for generating machine-readable, structured outputs from natural language prompts using Large Language Models. Our approach combines a custom-defined meta-language with schema-aware prompting to enable the creation of Meta-Language-defined Structures (MLDS). Evaluation results across two distinct use cases demonstrated a high validation rate (88%) and a substantial reduction in setup complexity (approx. 80% compared to LangChain and Pydantic). The resulting artifacts were interpretable and easy to refine, supporting rapid iterations in industrial workflows.

This approach is particularly valuable in domains where structured modeling is required but formal schema design presents a bottleneck. In addition to the use cases analyzed in this paper, we have successfully applied MLDS-based methods in a parallel study on immersive learning environments in Virtual Reality. There, natural language descriptions were used to generate didactically structured 3D scenes, allowing instructors to prototype entire VR labs without 3D design expertise, significantly reducing creation time while maintaining structural consistency.

These examples illustrate the broader potential for MLDS-based modeling in industries such as digital twin development, technical education, process automation, and safety-critical system design. Here, our method enables faster prototyping, reduces reliance on technical tooling, and facilitates interdisciplinary collaboration between domain experts and system architects.

Future work may explore the integration of MLDS with execution engines, real-time simulation tools, or graphical editors, as well as an expansion of the approach to additional modeling paradigms.

To reproduce our findings, please consult the readme file in our repository as referenced in Section IX.

IX. Data Availability

We have made all relevant research artifacts publicly available in the interest of fostering open science. Our repository contains two domain-specific meta-languages as well as additional MLDS artifacts generated with the help of these meta-languages. All materials can be accessed at the following GitHub repository: https://github.com/CoNaLaDe-MLDS/Artifacts.

Statement Concerning Generative AIs

We would like to clarify that this paper does not include substantial Sections of text directly produced by Generative AI, including LLMs.

However, since the primary objective of our work is to explore effective integration and utilization of LLMs in structured data generation and domain-specific modeling, we have naturally employed GenAI-generated outputs as part of our technical approach. In particular, LLMs were utilized to automatically generate structured MLDS files, code snippets, and example artifacts crucial to our experiments. Additionally, in selected cases, GenAI models assisted in translating specific text segments from German into English to enhance readability and accuracy.

We have clearly indicated within the paper all instances where GenAI-generated content was integrated, and provided an explicit analysis of its impact, including both benefits and limitations.

References

[1] C. K. Joyce, "The blank page: Effects of constraint on creativity," Social Science Research Network (SSRN), 2009.

[2] A. Fischer, L. Burk, R. Tavakoli Kolagari, and U. Wienkop, "Conalade mlds research artifacts," https://github.com/CoNaLaDe-MLDS/Artifacts.

[3] J. A. Díaz-Pace, A. Tommasel, and R. Capilla, "Helping novice architects to make quality design decisions using an llm-based assistant," in *Software Architecture*, M. Galster, P. Scandurra, T. Mikkonen, P. Oliveira Antonino, E. Y. Nakagawa, and E. Navarro, Eds. Cham: Springer Nature Switzerland, 2024, pp. 324–332.

[4] S. Colvin, "Pydantic: Data validation and settings management using python type annotations." [Online]. Available: https://pydantic.dev/

[5] H. Chase, "Langchain: Building applications with large language models," 2023. [Online]. Available: https://github.com/langchain-ai/langchain

[6] E. Filipovska, A. Mladenovska, M. Bajrami, J. Dobreva, V. Hillman, P. Lameski, and E. Zdravevski, "Benchmarking openai's apis and other large language models for repeatable and efficient question answering across multiple documents," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024, pp. 107–117.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2025. [Online]. Available: https://arxiv.org/abs/2303.18223

[9] S. Geng, H. Cooper, M. Moskal, S. Jenkins, J. Berman, N. Ranchin, R. West, E. Horvitz, and H. Nori, "JSON-SchemaBench: A rigorous benchmark of structured outputs for language models," *arXiv preprint arXiv:2501.10868*, 2025.

[10] Object Management Group (OMG), "Meta object facility (MOF) 2.0 core specification," 2006. [Online]. Available: http://www.omg.org/spec/MOF/2.0

[11] ——, "Unified modeling language (UML) specification version 2.5.1," 2017. [Online]. Available: https://www.omg.org/spec/UML/2.5.1/

[12] B. Wang, Z. Wang, X. Wang, Y. Cao, R. A Saurous, and Y. Kim, "Grammar prompting for domain-specific language generation with large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 030–65 055, 2023.

[13] T. Scholak, N. Schucher, and D. Bahdanau, "Picard: Parsing incrementally for constrained auto-regressive decoding from language models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9895–9901.

[14] N. Bassamzadeh and C. Methani, "A Comparative Study of DSL Code Generation: Fine-Tuning vs. Optimized Retrieval Augmentation," *arXiv preprint arXiv:2407.02742*, 2024.

[15] L. Netz, J. Michael, and B. Rumpe, "From Natural Language to Web Applications: Using Large Language Models for Model-Driven Software Engineering," in *Modellierung*, 2024.

[16] A. Fischer, J.-P. Tolvanen, and R. Tavakoli Kolagari, "Automotive cybersecurity engineering with modeling support," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024, pp. 319–329.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 543–552
DOI: 10.15439/2025F4427
ISSN 2300-5963 ACSIS, Vol. 43

Forecasting Consent in Organ Donation: Early Assessment of Machine-Learning Techniques

Arianna Freda
ORCID: 0009-0006-8856-9630
Roma Tre University
Rome, Italy
Email: arianna.freda@uniroma3.it

Davide Maestosi
Roma Tre University
Rome, Italy
Email: davide.maestosi@stud.uniroma3.it

Maurizio Naldi
ORCID: 0000-0002-0903-398X
LUMSA University
Rome, Italy
Email: m.naldi@lumsa.it

Gaia Nicosia
ORCID: 0000-0003-4043-2812
Roma Tre University
Rome, Italy
Email: gaia.nicosia@uniroma3.it

Andrea Pacifici
ORCID: 0000-0001-6144-0024
"Tor Vergata" University
Rome, Italy
Email: andrea.pacifici@uniroma2.eu

***Abstract*—Accurately predicting whether consent for organ donation will be granted is essential for optimizing timing and resource use in donor management. This study develops and evaluates machine learning models to estimate the likelihood of obtaining consent based on donor and contextual factors. The goal is to support early clinical decision-making by identifying cases where consent is more or less likely. Using real-world data from a regional transplant center operating under an opt-in system, we conduct data preprocessing, feature selection, and model training with various algorithms. Model performance is assessed using standard classification metrics, and key predictors of consent outcomes are identified. Results show accuracy levels exceeding 80%, highlighting the importance of including information about the relatives responsible for the decision. We also find that prediction accuracy varies with donor nationality, being higher for non-Italian donors. These findings demonstrate the value of predictive analytics in improving organ procurement efficiency and reducing unnecessary costs.**

***Index Terms*—Health care, Modeling and prediction, Machine learning, Decision support**

I. Introduction

ORGAN transplantation remains the most effective and economically viable treatment for patients with end-stage organ failure. However, despite ongoing efforts to raise awareness and increase donor registrations, a persistent and critical gap remains between the demand for and supply of transplantable organs [1], [2]. Among the key barriers contributing to this shortage is the high rate of opposition to organ donation, particularly when consent must be obtained post-mortem from next-of-kin.

This work has been partially funded by the European Union–Next Generation EU within the framework of PRIN 2022 Project "MEDICINE+AI, Law and Ethics for an Augmented and Human-Centered Medicine" (2022YB89EH) – CUP E53D23007020006

The acquisition of consent is a pivotal step in the organ donation process. In systems where explicit authorization is required—such as the *opt-in* framework in place in many regions—the inability to secure timely consent can lead to lost donation opportunities, unnecessary delays, and avoidable costs. Even in jurisdictions with *opt-out* legislation, in practice, healthcare professionals often still seek family confirmation, making the consent process a practical and ethical bottleneck.

In this paper, we propose a data-driven approach to support and optimize the consent acquisition phase through predictive modeling. Specifically, we develop and evaluate machine learning models aimed at forecasting the likelihood of obtaining consent for organ donation. By leveraging available demographic, clinical, and contextual data, our models seek to provide early estimates of consent probability, offering valuable decision support for clinicians involved in donor management.

As it is better discussed in Section III-B, this predictive capability has the potential to significantly enhance operational efficiency. For example, if a high probability of consent is anticipated, preparatory clinical activities can be initiated earlier, minimizing the time to transplantation and reducing organ deterioration risks. Conversely, in cases where consent is unlikely, healthcare teams may choose to delay or forego certain cost-intensive procedures, thus conserving resources.

Our study contributes to the field of healthcare information systems by introducing a predictive tool for estimating the likelihood of consent or opposition in organ donation. The practical relevance of such predictions lies in their integration into transplant logistics, where they can support informed decision-making and improve both clinical efficiency and cost-effectiveness.

The main contributions of the paper are as follows.

- We develop predictive models based on established machine learning algorithms to estimate the likelihood of family consent (or opposition) to organ donation, with a focus on the Italian opt-in system.
- We present a context-specific analysis using real-world data from the Lazio region, contributing the first known predictive study on consent acquisition in Europe.
- We compare various machine learning algorithms and identify the most effective approaches for consent prediction under data scarcity conditions.
- We highlight differences in predictive performance across different nationality groups.

The remainder of the paper is organized as follows. In Section II, we review the relevant literature. In Section III, we describe the donor management process, with a particular focus on the critical role of consent acquisition. The input data used in this study, along with the preprocessing and transformation procedures (including feature deletion, grouping, and refinement), are presented in Section IV. In Section V, we outline the predictive modeling methodology, detailing the algorithms, performance metrics, and selected features. The results of the predictive analysis are reported and discussed in Section VI. Finally, conclusions and future research directions are presented in Section VII.

II. Related literature

Organ donation consent is a critical component of the transplantation process, with direct implications for timing, cost, and overall success rates. As highlighted in the introduction, delays in consent acquisition can lead to suboptimal organ preservation and increased process costs [3]–[6]. In this section, we review the literature across three main areas: (i) barriers and refusal patterns in consent acquisition, (ii) strategies and policy frameworks aimed at improving consent rates, and (iii) the use of artificial intelligence and predictive analytics in organ transplantation.

A number of studies have investigated the sociocultural and procedural factors that hinder the acquisition of post-mortem consent. For example, the refusal rate in South Korea was examined in [7], where the authors identified structural and cultural barriers to obtaining consent. The intersection of ethics, religion, and policy was found to influence organ donation practices in South Asia [8]. Similarly, Lewis et al. discussed general barriers to organ donation across different populations [9]. In the Italian context, Grossi et al. found that refusal rates were significantly higher among immigrant populations compared to native citizens, highlighting the importance of demographic and cultural variables in consent prediction [10].

Efforts to proactively increase donation rates and consent have also been widely studied. In [11], hospitals were benchmarked based on their consent rates, with key influencing factors including donor age and ethnic background. A number of studies [9], [12], [13] have compared opt-in and opt-out systems, with opt-out systems generally associated with higher donation rates. However, these studies also acknowledge the complex ethical and political considerations surrounding such transitions. The role of family members in the final consent decision is emphasized in [14]–[16], even in jurisdictions with presumed consent. Additionally, many studies have revealed that one of the most influential issues is the expertise of the person approaching the family to obtain consent for organ donation. For instance, [17] highlighted the crucial influence of healthcare professionals' engagement and communication skills in facilitating positive consent outcomes, while [18] reviews studies on differences in consent rates based on the type of professional doing the asking. Expanding on these findings, [19] presents the practices adopted by Spanish transplant coordinators during consent requests for Irreversible Cessation of Circulatory and Respiratory Functions (ICOD). The study outlines the sequence of preparatory steps coordinators follow prior to engaging families in decision-making. Many of these steps are not only rooted in a structured and empathetic communication approach but are also recognized as best practices in broader organ donation contexts. This method has been associated with a notably low rate of family refusals, reinforcing the importance of professional training and protocol in achieving favorable consent outcomes. Additionally, some studies explore effective strategies for obtaining consent from families. For example, [19] presents the practices adopted by Spanish transplant coordinators when requesting consent for Irreversible Cessation of Circulatory and Respiratory Functions (ICOD). The study outlines the sequence of steps coordinators follow before engaging the family in decision-making. Many of these steps are recognized as best practices in broader organ donation contexts and are associated with a notably low rate of family refusals.

Broader applications of AI in organ donation and transplantation have also been explored. In [20], a comprehensive review was conducted on the use of artificial intelligence across various stages of the transplant process, including organ allocation, process optimization, and clinical decision support. However, the application of AI specifically to predict consent outcomes remains largely unexplored.

While the importance of timely consent acquisition in organ donation is widely acknowledged, research employing predictive modeling in this context remains limited. Notable exceptions include the studies by Khan et

al. [2] and Tutun et al. [21], both of which apply machine learning techniques to forecast consent outcomes. In particular, the methodologies proposed by Khan et al. [2] demonstrate high levels of accuracy, indicating that such approaches are both viable and effective. Ultimately, they have the potential to increase consent rates and thereby save more lives among individuals on organ transplant waiting lists.

In [2], the authors develop a predictive framework that integrates machine learning and network science to estimate the likelihood of family consent in organ donation. Their model is designed not only to predict outcomes but also to uncover the most influential factors driving family decisions. The basic goal is to provide Organ Procurement Organizations (OPOs) and hospital staff with actionable insights that support more effective communication strategies during consent discussions. This work emphasizes the role of predictive models as expert systems to improve consent rates and, by extension, help mitigate the persistent mismatch between organ supply and demand.

Similarly, Tutun et al. [21] propose a responsible AI framework that merges network science with machine learning to enhance consent prediction. Their study leverages a large dataset collected from approximately 1,500 consent discussions across 92 hospitals in New York, recorded between January 2016 and March 2018. Results indicate that incorporating network-based features significantly improves the performance of traditional ML algorithms. The study highlights how algorithmic insights can support ethical and effective approaches to organ procurement.

Despite their relevance, both studies above differ from our work. First, their primary objective is to reduce the rate of family refusal, with a strong focus on optimizing the approach strategies of healthcare professionals. In contrast, our goal is to develop a predictive framework that provides a probabilistic estimate of consent likelihood, which can be integrated into transplant logistics and used to inform cost-effective and timely decision-making throughout the donor management process. Second, the methodological approaches vary considerably. While [2] and [21] rely on extensive sets of features and large datasets—enabling more complex models and improved predictive performance—our study operates on a significantly smaller dataset collected in Italy. This naturally constrains model complexity but provides a novel, context-specific perspective.

This context-specific perspective builds on previous investigations of the Italian case study [22], [23]. In [22], a probabilistic model was developed to estimate the time required to obtain consent, using real-world data from successful organ donation cases in the Lazio region, where an opt-in consent system is in effect. However, that study focused on modeling the timing of consent acquisition, rather than predicting the binary outcome of whether consent would be granted. The same case study was further explored in [23], where the pre-transplant process was modeled and simulated to identify time-critical activities through critical path analysis. Building on that work, [24] conducted a cost-benefit analysis to evaluate management strategies aimed at balancing time efficiency and cost-effectiveness throughout the pre-transplant phase.

To the best of our knowledge, our study is the first to address consent prediction using real-world data from Italy, and more broadly, among the first of its kind in Europe. This regional focus enables us to explore consent dynamics within the specific legal and procedural framework of an opt-in system, offering valuable insights for local policy and process optimization.

III. Overview of Donor Management and Consent Dynamics

Transplant centers worldwide generally follow a hierarchical structure, with a national coordinating authority and regional and local healthcare facilities. This is also the case in the center we consider here. Namely, we focus on the Italian Transplant System, involving the National Transplant Center (NTC), Regional Transplant Centers (RTCs), and local hospital units. Organ transplantation is a complex process involving many phases and actors. In this paper, we are interested in the process leading to consent acquisition. In this section, we describe the donor management process and the relevance of the timeliness of consent acquisition for the overall efficiency of the process.

A. The organ donor management process

The organ donor management process is a structured set of activities that begins with the identification of a potential donor and ends with the retrieval of organs and their allocation to compatible recipients. Its primary aim is to ensure the quality and availability of transplantable organs while safeguarding recipient health.

This work focuses on the process currently implemented at the RTC of the Lazio region in Italy. The UML (Unified Modelling Language) diagram in Figure 1 represents a sketch of the donation process, which begins when a potential donor is identified. The identification of a donor starts with the recognition of the brain death status of a patient, which qualifies him/her as a potential donor. A six-hour period is required for Brain Death Assessment (BDA). After BDA, two parallel streams of activities are triggered:

- Acquisition of consent for organ donation;
- Initial clinical and diagnostic evaluations.

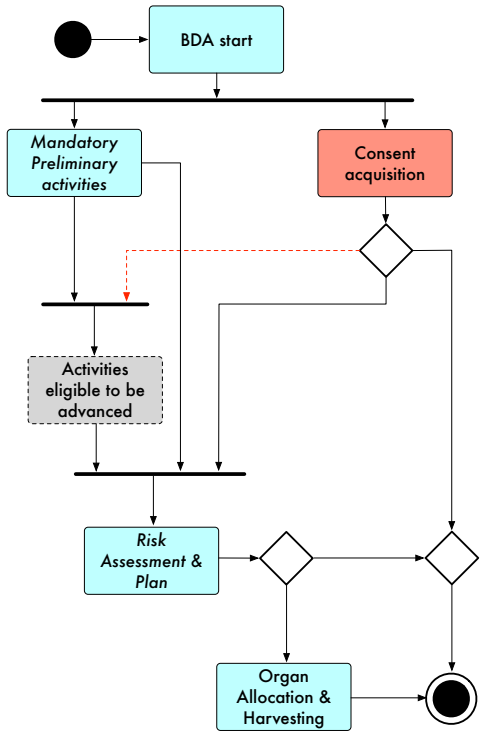


Fig. 1: Compact UML diagram of the process.

The following phases involve a comprehensive assessment of donor suitability, including laboratory analyses, virological testing, and instrumental diagnostics. Although there are no strict technical constraints, higher-cost exams are generally performed only after consent is obtained, in order to avoid unnecessary expenses when the transplant cannot proceed due to a lack of consent. If consent arrives in due time, it allows the medical staff to proceed with all the medical tests needed to properly carry out the transplant. If consent is denied, any further medical test for transplant is cancelled. Some activities are susceptible to be advanced, i.e., carried out even if consent has not arrived yet, so as to reduce the overall process time. Those activities are shown as *advance-eligible* in Fig. 1. Dashed arrows indicate precedence relations that are not mandatory. After the examinations are carried out to assess the clinical suitability of the potential donor, a risk level is defined—sometimes following consultation with the Second Opinion team appointed by the National Transplant Center (NTC). If the evaluations indicate that the donor presents an unacceptable risk to recipients, the process is immediately stopped. Once the donor's risk level has been defined, the process moves into the Organ Allocation and Harvesting phase. In this stage, potential recipients are identified and matched based on clinical compatibility and logistical considerations, through coordinated efforts involving the NTC, Regional Transplant Centers (RTCs), local transplant teams, and logistics services, ensuring timely and efficient organ retrieval and distribution.

B. Impact of Consent activity on process time and cost

If a potential donor has not expressed their wishes regarding organ donation while alive, the responsibility for releasing consent is transferred to their relatives. The Consent acquisition phase becomes a critical crosspoint in the donation process—regardless of whether consent is ultimately granted or denied by those legally empowered to decide. This phase significantly affects both the duration and cost of the process.

To illustrate, imagine that the family or other authorized individuals communicate their decision at some unknown time point T after the donation process begins. By that time, a possibly empty subset S of activities (eligible to be initiated before the consent is given) may already have been started or completed. Let c_S denote the cumulative cost of these activities. If consent is not granted, then a cost c_S is effectively wasted.

This would suggest a conservative strategy: delaying costly actions until consent is confirmed to keep S, and thus c_S, minimal. However, such caution comes at a price. Postponing action until the decision time T prolongs the process—especially in cases where consent is eventually given. The more activities completed before T, the more progress has been made by the time consent is secured, allowing transplantation to proceed more quickly. Since organs are perishable and their viability deteriorates over time, any delay can compromise the success of the transplant.

This tension calls for a cost-benefit analysis to guide decision-making, weighing the trade-off between minimizing wasted costs and maximizing time efficiency in cases where consent is ultimately granted (see [24]). In short, cost-efficiency favors deferring activities until consent is certain, while time-sensitivity and organ viability argue for initiating tasks early—even at the risk of incurring unrecoverable costs if consent is later denied, particularly when T is large. If we indicate the process times as T_{adv} and $T_{\mathrm{del}} > T_{\mathrm{adv}}$, respectively, when activities are advanced or postponed with respect to consent acquisition, we can summarize the time-cost trade-offs in the matrix shown in Fig. 2, where each bracket contains the time needed to complete the process and the wasted cost.

In this context, a predictive tool that can estimate the likelihood of consent with reasonable accuracy would be highly valuable. High confidence in a positive outcome justifies starting preparations sooner, improving both patient outcomes and resource utilization. Conversely, if the predicted probability of denial is high, delaying

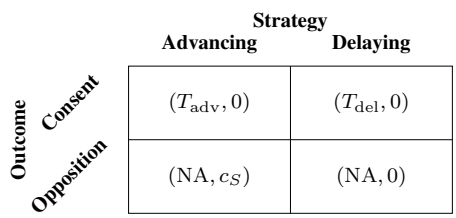


Fig. 2: Trade-offs in donor management process strategy

costly tasks becomes the prudent course, helping to avoid unnecessary expenditure.

IV. Dataset

We collected and analysed a dataset concerning potential donors at the Regional Transplant Center for the Lazio region in Italy. In Italy, the opt-in regulation applies where the potential donor may have given or denied his/her consent to transplant when they were alive. In the case the potential donor did not express their decision concerning organ transplant during their life, it is up to the family members to decide after the potential donor's brain death. The decision about donation as well as the personal details of the potential donor and their relatives are recorder. In this section, we describe the dataset including that information.

Our data collection took place over four years, from 2021 to 2024. We can first see how the consent information is distributed over the years in Table I. A significant growth year-on-year can be observed, with a CAGR (Compound Annual Growth Rate) of 6.7%. The overall number of instances is 867 (the data for 2024 do not cover the whole year). However, the dataset is heavily imbalanced, with the *Silent* group (i.e., those not expressing their opinion during their lifetime) being the large majority (80.2% of the total). The overall dataset has been processed through Bootstrap and 10-fold cross-validation to get the final dataset to be fed to the algorithms described in Section V-A for training and testing. Since our focus is to predict consent for potential donors after brain death (DBD), we are not interested in those donors expressing their consent during their lifetime. Actually, for those potential donors, the donor management process either starts with all the medical tests to be carried out as soon as possible, or does not start at all. Hence, in the following, we will focus on the silent potential donors. After examining the dataset, we removed those instances where most features are absent. The overall number of silent donors has then reduced to 653, which is the reference number we will consider in the following.

In addition, in order to examine how the nationality of the donor may influence the decision and may be predicted with larger or lesser accuracy, we considered the following subsets of potential donors:

- European donors;
- Italian donors;
- non-Italian donors.

TABLE I: Consent expression in the dataset

Year	Consent	Opposition	Silent	Total
2021	18	18	196	232
2022	32	14	202	248
2023	34	28	202	264
2024	23	5	95	123
Total	107	65	695	867

TABLE II: Number of silent donors by nationality

Nationality	Numerosity
European	585
Italian	542
non-Italian	111

The numerosity of the three subsets is shown in Table II. As can be seen, Italians make up roughly 90% of the whole set of European potential donors, as expected.

We can see in Fig. 3 how nationality influences the decision about organ donation during lifetime. While we see a balanced distribution for European and South American potential donors, the distribution is heavily imbalanced for the other continents. The culture of donation is quite widespread in North America, where just a small minority opts for opposing donation. On the other end of the spectrum, we find Asia and Africa, where oppositions largely outnumber consents.

For each potential donor, 67 features have been collected. Those features may be subdivided into six groups:

1) General;
2) Personal data;
3) Information on relatives;
4) Risk level;
5) Organ information;
6) Consent.

Group 1 contains information related to the death of the potential donor, including time and cause of death. Group 2 includes all personal data pertaining to the donor: name, surname, age, gender, date of birth, place of birth, place of residence, height, and weight. Group 3 provides data necessary to determine the degree of kinship of the relatives involved in the consent process (e.g., parents or siblings). Group 4 comprises information regarding the donor's risk profile, based on second opinions obtained by the medical staff during the organ compatibility assessment. Finally, Group 5 is

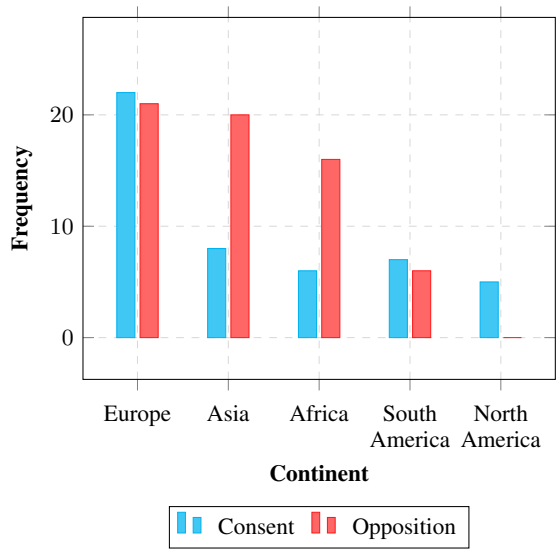


Fig. 3: Frequency of support and opposition by continent

included only if consent has been obtained, and contains specific details about the organ that may be considered for transplantation.

It is to be noted that Group 6 actually includes the target variable (whether Consent has been released), but also information on the source of consent (the donors themselves or their relatives).

V. Methodology

The problem of consent prediction has been modelled as a binary classification one, where the two classes are Consent or Opposition (Consent denial) and the potential features are the 67 mentioned in Section IV. In this section, we describe the classification algorithms that we have employed, the performance metrics that we have adopted to assess the algorithms, and the feature selection process.

A. Algorithms and tools

We have employed the following established classification algorithms:

- Decision Tree;
- Logistic Regression;
- AdaBoost;
- Feedforward neural network.

Each method in this selection has its own characteristics.

Decision Trees split data into subsets using feature thresholds, creating a tree-like structure where each node represents a feature condition, and leaves represent class labels. They are easy to interpret, but are prone to overfitting, especially on small datasets like ours.

Logistic Regression is a linear model where the probability of a class is predicted by using a logistic function. Regularization is achieved through LASSO (L1 Regularization) by adding a penalty term to reduce overfitting and performing feature selection by shrinking some coefficients to zero. It is simple, interpretable, good for high-dimensional data with irrelevant features. It has been chosen since we have a relatively large number of features (we are considering non-image data) many of which may prove to be not relevant. It has limited capacity for complex relationships.

AdaBoost (Adaptive Boosting) is an ensemble method that combines multiple weak learners (often decision stumps) by iteratively focusing on previously misclassified examples. It should be more accurate than single models like decision trees or logistic regression and works well with less data (as in our case). However, it is sensitive to noisy data and outliers.

Finally, feedforward Neural Networks represent a very simple type of deep learning model where data flows through multiple layers of interconnected nodes (neurons) from input to output. They are powerful for modeling complex, non-linear relationships. However, they require large datasets, which we do not have at present, and are less interpretable.

In the following we describe the configuration details for each classification algorithm.

For decision trees we have employed a general (non-binary) structure, where splitting was carried out just for nodes with at least five instances. The maximum tree depth was kept at 50 and pruning was employed for nodes where the majority class includes more than 95% of the instances.

Logistic Regression was employed with LASSO regularization with L1 norm and the classification threshold set at 0.5 probability..

In AdaBoost 60 trees were used as base estimators, with the learning rate set to 1.

A multilayer network with 100 neurons per hidden layer was employed with a feedforward structure. A logistic sigmoid function was chosen as the activation function used, and ADAM was selected as solver, based on the gradient descent algorithm for calculating weights during the training phase [25]. The maximum number of iterations (stopping condition) was set to 200.

B. Performance metrics

Our aim is to identify the cases where consent to donation is obtained. We have then a binary classifier, where we label as Positive the cases where consent is actually obtained. With the usual 4-cell confusion matrix, we have then the following possible outcomes: True and False Positives (TP and FP, respectively) as well as True

TABLE III: MOST SIGNIFICANT FEATURES

Feature	Overall	European	Italian	non-Italian
Donor's hospital	✓	✓	✓	✓
Nationality	✓	✓		
Birthplace region	✓	✓	✓	
Continent	✓			✓
Cause of death	✓	✓	✓	✓
Region of residence	✓	✓	✓	✓
Donor's age	✓	✓	✓	✓
Size of birth city	✓	✓	✓	
Donor gender	✓	✓	✓	✓
Year of death	✓	✓	✓	✓
Donor blood type	✓		✓	
Size of city of residence	✓	✓	✓	✓
Part of Italy (residence)	✓	✓	✓	✓
Country of birth				✓
Country of residence				✓

and False Negatives (TN and FN, respectively). We adopt the following extablished performance metrics [26]:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 Score = $2\frac{1}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}$

C. Features

Though we have a large number of potential features (67), most of them have proved to be of small significance for the performance. For that reason, we have chosen a subset of significant features by employing three metrics: the information gain, the gain ratio, and the Gini index. The choice of features depends on the subset of interest, so that we have selected the most relevant features for each data subset. In Table III, we show the selection for the full dataset and the three subsets investigated in this paper. Some features appear to be significant for all the sets, namely: Donor's hospital, Cause of death, Location of residence, Donor's age, Year of death, and Size of city of residence.

VI. RESULTS

We have examined our dataset by employing all the four algorithms described in Section V-A. We first consider the results obtained for the whole set of silent donors, i.e., those potential donors who did not express their consent (or denial) during their lifetime, so that the decision has been left to their relatives. Then we examine the results for the nationality subsets defined in Section IV.

In Table IV, we show the major performance metrics. Linear Regression is by far the worst performer, exhibiting figures below 70% on all four metrics. We see a significant improvement by roughly ten percentage points with decision trees. However, if we abandon the intrinsic explainability of decision trees, both ensemble techniques (AdaBoost in this case) and neural networks

TABLE IV: MODEL PERFORMANCE USING BOOTSTRAP ON THE SILENT DATASET

Technique	Accuracy	F1 Score	Precision	Recall
Tree	0.76	0.76	0.76	0.76
LASSO	0.66	0.64	0.65	0.66
AdaBoost	0.85	0.85	0.85	0.85
Neural Net	0.82	0.82	0.82	0.82

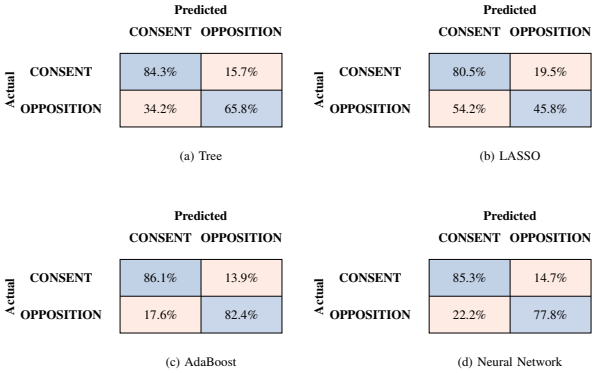


Fig. 4: Confusion matrices (percentages) using Bootstrap on the silent dataset

(though in their simple feedforward configuration) allow us to reach good performances achieving figures well beyond 80% on all metrics. AdaBoost appears the best overall performer.

We can take a closer look at the performance by examining the confusion matrices in Fig. 4. Linear regression (LASSO) shows the worst imbalance in performance, being good in recognizing Consent decisions, but failing by misclassifying a major portion of Opposition decisions (performing even worse than a random classifier). The imbalance in performance decreases but is still significant in decision trees (though the majority of oppositions is now classified correctly). The per-class performance improves significantly with AdaBoost and neural networks. The imbalance across classes is very small for AdaBoost and a bit larger for neural networks. In both cases, the imbalance works against oppositions, which are classified correctly in less than 80% of the instances with neural networks.

Summing up, Adaboost appears as the best classifier, both overall and over either class, exhibiting well-balanced performances.

We can now examine the performances on the subsets described in Section IV, considering European, Italian, and non-Italian potential donors.

In Table V, we show the performance metrics for European donors. The performance is quite similar to what we have obtained on the overall dataset, though a bit better. This may suggest that European donors may

TABLE V: MODEL PERFORMANCE USING BOOTSTRAP ON EUROPEAN DONORS

Technique	Accuracy	F1 Score	Precision	Recall
Tree	0.78	0.78	0.78	0.78
LASSO	0.67	0.66	0.66	0.67
AdaBoost	0.84	0.84	0.84	0.84
Neural Net	0.85	0.85	0.85	0.85

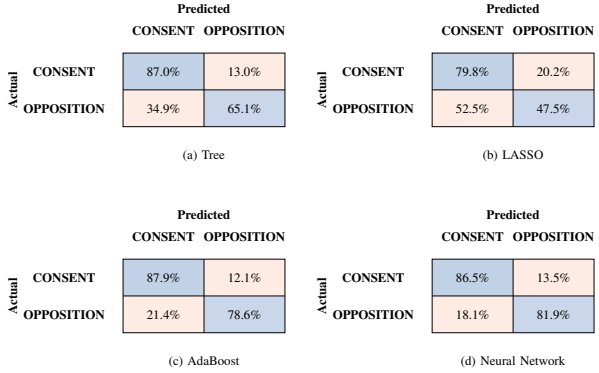


Fig. 5: Confusion matrices (percentages) using Bootstrap on the European dataset

be predicted more reliably than non-European donors.

Similarly to what we did for the whole dataset, we can examine the performance metrics on either class by looking at the confusion matrices in Fig. 5. We see again that decision trees have similar performances as the best performing algorithms for the Consent class. It is the performance on the Opposition class that brings the overall accuracy down. The imbalance between the performance for the two classes is high for the LASSO regression. Neural networks appear as the most balanced algorithm for this subset.

If we focus on Italian donors, which are, as expected, the largest nationality subgroup by far, we obtain the results shown in Table VI. We do not see significant changes with respect to the results that we have already shown. The pertaining confusion matrix is shown in Fig. 6.

What removes some doubts about the reliability of predictions for nationals is the examination of Table VII, where the performance improves significantly. We can

TABLE VI: MODEL PERFORMANCE WITH BOOTSTRAP ON ITALIAN DONORS

Technique	Accuracy	F1 Score	Precision	Recall
Tree	0.77	0.76	0.77	0.77
LASSO	0.69	0.67	0.68	0.69
AdaBoost	0.81	0.81	0.81	0.81
Neural Net	0.83	0.83	0.83	0.83

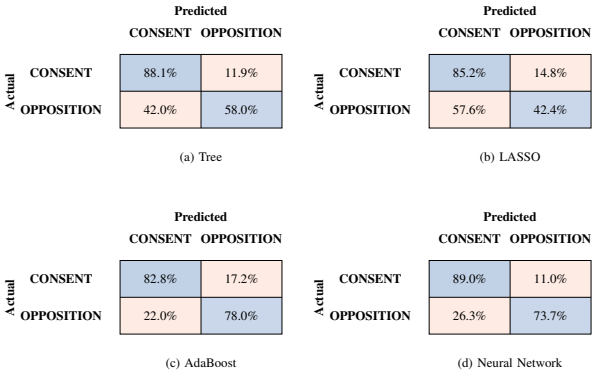


Fig. 6: Confusion matrix (percentages) for Neural Network using Bootstrap on the Italian dataset

TABLE VII: MODEL PERFORMANCE WITH BOOTSTRAP ON NON-ITALIAN DONORS

Technique	Accuracy	F1 Score	Precision	Recall
Tree	0.71	0.71	0.71	0.71
LASSO	0.71	0.71	0.71	0.71
AdaBoost	0.86	0.86	0.86	0.86
Neural Net	0.85	0.85	0.85	0.85

conclude that, though AdaBoost and neural networks provide accuracy values over 80% in any case, predictions are more accurate for non-Italian potential donors than for Italian ones. though the size of the non-Italian subsample is not very large, it appears that the features that we have employed give much clearer indications for non-Italian donors.

Finally, we have examined the possibility of improving the forecasting accuracy by introducing features related to family members. The results in Table VIII clearly show a significant improvement for the worst performing algorithms (decision trees and LASSO regres-

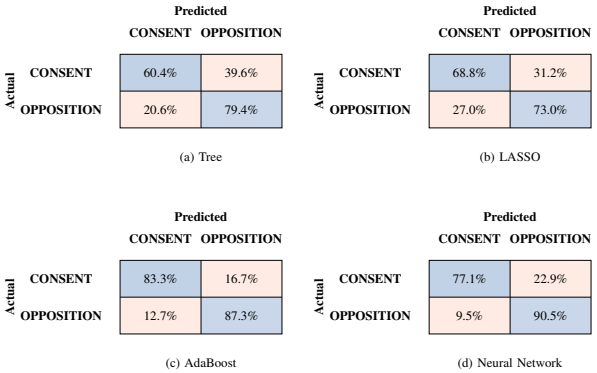


Fig. 7: Confusion matrices (percentages) using Bootstrap on the non-Italian dataset

TABLE VIII: MODEL PERFORMANCE WITH BOOTSTRAP BY ADDING INFO ON FAMILY MEMBERS

Technique	Accuracy	F1 Score	Precision	Recall
Tree	0.80	0.79	0.80	0.80
LASSO	0.79	0.79	0.79	0.79
AdaBoost	0.88	0.88	0.88	0.88
Neural Net	0.86	0.86	0.86	0.86

sion) and a small improvement for the best-performing algorithm (AdaBoost and neural networks). For the latter algorithms, the accuracy gets closer to 90%. The addition of that information, though notoriously difficult to get, is then highly recommended.

VII. CONCLUSIONS

This study has demonstrated the feasibility and effectiveness of using machine learning techniques to forecast consent outcomes in the organ donation process. By leveraging a comprehensive set of socio-demographic, clinical, and contextual features, our models—particularly AdaBoost and neural networks—achieved strong predictive performance, with accuracies exceeding 85% and well-balanced classification across both Consent and Opposition cases. These results affirm the value of predictive modeling in addressing one of the most critical and uncertain phases of organ donation: the acquisition of consent.

The practical implications of this work are substantial. In a process where timing is critical and resources are constrained, early estimation of the likelihood of obtaining consent enables healthcare professionals to tailor their actions. High predicted probabilities of consent support the early initiation of cost-intensive diagnostic and preparatory procedures, thereby minimizing delays and reducing the risk of organ deterioration. Conversely, low predicted probabilities can help avoid unnecessary expenditures and optimize allocation of staff and equipment. The integration of such predictive tools into clinical decision support systems can therefore enhance both operational efficiency and cost-effectiveness in transplant logistics.

Looking ahead, several promising directions could further extend this research on consent prediction in solid organ transplantation. First, integrating additional qualitative and behavioral data—such as prior interactions with healthcare staff or known family attitudes—may enhance predictive accuracy and provide deeper contextual understanding. Second, the development of real-time, interpretable models tailored for clinical settings would improve usability and foster trust among healthcare professionals, potentially leveraging more advanced and explainable forecasting architectures [27].

Third, conducting cross-regional studies with data from multiple transplant centers would allow for broader validation and support the generalization of findings, thus laying the groundwork for national or even international implementation of consent prediction tools within organ donation systems.

Finally, we believe that embedding the proposed model within a comprehensive decision-support framework—capable of optimizing various components of the donor management process, including logistical challenges such as organ transportation [28]–[31]—could significantly improve the overall performance and efficiency of the system (see, e.g., [32]).

ACKNOWLEDGMENT

The authors wish to express their sincere gratitude to the staff of the Regional Transplant Center of Lazio for their invaluable support. Their detailed explanations of the donor management process, particularly the procedures and challenges related to consent acquisition, as well as their essential contributions to data collection and analysis, were fundamental to the development of this study.

REFERENCES

[1] R. Girlanda, "Deceased organ donation for transplantation: challenges and opportunities," *World Journal of Transplantation*, vol. 6, no. 3, p. 451, 2016. doi: doi: 10.5500/wjt.v6.i3.451

[2] M. E. Khan and S. Tutun, "Understanding and predicting organ donation outcomes using network-based predictive analytics," *Procedia Computer Science*, vol. 185, pp. 185–192, 2021. doi: 10.1016/j.procs.2021.05.020

[3] G. Fleming and E. M. Thomson, "Organ donation and management of the potential organ donor," *Anaesthesia & Intensive Care Medicine*, vol. 22, no. 8, pp. 475–481, 2021. doi: 10.1016/j.mpaic.2021.06.007

[4] R. M. Kotloff, S. B. G. J. Fulda, D. Malinoski, V. N. Ahya, L. Angel *et al.*, "Management of the potential organ donor in the ICU: society of critical care medicine/American college of chest physicians/association of organ procurement organizations consensus statement," *Critical Care Medicine*, vol. 43, no. 6, pp. 1291–1325, 2015. doi: 10.1097/CCM.0000000000000958

[5] R. Vanholder, B. Domínguez-Gil, M. Busic, H. Cortez-Pinto, J. C. Craig, K. J. Jager, B. Mahillo *et al.*, "Organ donation and transplantation: a multi-stakeholder call to action," *Nature Reviews Nephrology*, vol. 17, no. 8, pp. 554–568, 2021. doi: 10.1038/s41581-021-00425-3

[6] C. Dictus, B. Vienenkoetter, M. Esmaeilzadeh, A. Unterberg, and R. Ahmadi, "Critical care management of potential organ donors: our current standard," *Clinical Transplantation*, vol. 23, pp. 2–9, 2009. doi: 10.1111/j.1399-0012.2009.01102.x

[7] W.-H. Cho, "Causes of donation failure and improvement measures analyzed based on data from domestic deceased donors in 2019," *Korean Journal of Transplantation*, vol. 34, no. 4, pp. 219–230, 2020. doi: 10.4285/kjt.20.0040

[8] A. S. Sakarwala, A. Dejwani, M. F. R. Moinuddin, W. Ahmad, and H. Furqan, "Bridging ethics, religion, and policy: Transforming organ donation in south asia," *INNOVAPATH*, vol. 1, no. Q2, pp. 6–6, 2025. doi: 10.63501/aczjpq11

[9] A. Lewis, A. Koukoura, G.-I. Tsianos, A. A. Gargavanis, A. A. Nielsen, and E. Vassiliadis, "Organ donation in the US and Europe: The supply vs demand imbalance," *Transplantation Reviews*, vol. 35, no. 2, p. 100585, 2021. doi: 10.1016/j.trre.2020.100585

[10] A. A. Grossi, F. Puoti, U. Maggiore, and M. Cardillo, "Refusal rates to organ donation in intensive care units among immigrant populations in Italy," *Transplant International*, vol. 36, p. 11674, 2023. doi: 10.3389/ti.2023.11674

[11] R. Matesanz, E. Coll, B. Domínguez-Gil, G. De La Rosa, R. Marazuela, V. Arráez, P. Elorrieta *et al.*, "Benchmarking in the process of donation after brain death: a methodology to identify best performer hospitals," *American Journal of Transplantation*, vol. 12, no. 9, pp. 2498–2506, 2012. doi: 10.1111/j.1600-6143.2012.04128.x

[12] N. J. Williams and N. C. Manson, "'take my kidneys but not my corneas'—selective preferences as a hidden problem for 'opt-out' organ donation policy," *Bioethics*, vol. 36, no. 8, pp. 829–839, 2022. doi: 10.1111/bioe.13052

[13] A. Rosenblum, L. Horvat, L. Siminoff, V. Prakash, J. Beitel, and A. Garg, "The authority of next-of-kin in explicit and presumed consent systems for deceased organ donation: An analysis of 54 nations," *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association*, vol. 27, pp. 2533–46, 2011. doi: 10.1093/ndt/gfr619

[14] F. Ghorbani, H. Khoddami-Vishteh, O. Ghobadi, S. Shafaghi, A. Rostami Louyeh, and K. Najafizadeh, "Causes of family refusal for organ donation," *Transplantation Proceedings*, vol. 43, no. 2, pp. 405–406, 2011. doi: 10.1016/j. transproceed.2011.01.031

[15] A. Molina-Pérez, J. Delgado, M. Frunza, M. Morgan, G. Randhawa, J. R.-V. de Wijdeven, S. Schicktanz, E. Schiks, S. Wöhlke, and D. Rodríguez-Arias, "Should the family have a role in deceased organ donation decision-making? A systematic review of public knowledge and attitudes towards organ procurement policies in europe," *Transplantation Reviews*, vol. 36, no. 1, p. 100673, 2022. doi: 10.1016/j.trre.2021.100673

[16] D. J. Molina-Pérez A, Rodríguez-Arias D, "Differential impact of opt-in, opt-out policies on deceased organ donation rates: a mixed conceptual and empirical study," *BMJ Open*, 2022. doi: 10.1136/bmjopen-2021-057107

[17] F. Ruta, C. Somma, M. Lusignani, G. Vitobello, L. Gesualdo, F. Dal Mas, A. Peloso, and L. Cobianchi, "The healthcare professionals' support towards organ donation. An analysis of current practices, predictors, and consent rates in Apulian hospitals," *Annali dell'Istituto Superiore di Sanità*, vol. 57, pp. 291–299, 2021. doi: 10.4415/ANN_21_04_04

[18] A. L. Simpkin, L. C. Robertson, V. S. Barber, and J. D. Young, "Modifiable factors influencing relatives' decision to offer organ donation: Systematic review," *BMJ (Online)*, vol. 338, no. 7702, p. 1061 – 1063, 2009. doi: 10.1136/bmj.b991

[19] R. García-Sánchez, M. Soria-Oliver, J. S. López, J. M. Martínez, M. J. Martín, A. Barceló-Soler, E. Coll, J. Roldán, D. Uruñuela, and A. Fernández-Carmona, "Requesting relatives' consent for intensive care for organ donation: An empirical analysis of spanish transplant coordinators' practices," *Transplantation*, vol. 109, no. 5, p. e237 – e247, 2025. doi: 10.1097/TP.0000000000005210

[20] D. B. Olawade, S. Marinze, N. Qureshi, K. Weerasinghe, and J. Teke, "The impact of artificial intelligence and machine learning in organ retrieval and transplantation: A comprehensive review," *Current Research in Translational Medicine*, vol. 73, no. 2, 2025. doi: 10.1016/j.retram.2025.103493

[21] S. Tutun, A. Harfouche, A. Albizri, M. E. Johnson, and H. He, "A responsible ai framework for mitigating the ramifications of the organ donation crisis," *Information Systems Frontiers*, vol. 25, no. 6, p. 2301 – 2316, 2023. doi: 10.1007/s10796-022-10340-y

[22] M. Naldi, A. Freda, G. Nicosia, A. Pacifici, G. Teti, and M. Feccia, "Modelling the consent acquisition time in organ donor management through clustering and mixture probability models," in *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 2024, pp. 1213–1218.

[23] M. Feccia, A. Freda, M. Naldi, G. Nicosia, and A. Pacifici, "Modelling and simulating the organ donation process using bootstrap and event-driven process chain representation," *Journal of Simulation*, 2025. doi: 10.1080/17477778.2025.2486709

[24] A. Freda, M. Naldi, G. Nicosia, and A. Pacifici, "A trade-off analysis of costs and time in organ transplantation management," 2025, submitted for publication.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[27] P. Fantozzi and M. Naldi, "The explainability of transformers: Current status and directions," *Computers*, vol. 13, no. 4, p. 92, 2024. doi: 10.3390/computers13040092

[28] A. Agnetis, M. Cosmi, G. Nicosia, and A. Pacifici, "Two is better than one? Order aggregation in a meal delivery scheduling problem," *Computers and Industrial Engineering*, vol. 183, 2023, 109514. doi: 10.1016/j.cie.2023.109514

[29] J. Owsiński, J. Stańczak, A. Barski, K. Sęp, and P. Sapiecha, "Graph based approach to the minimum hub problem in transportation network," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2015. doi: 10.15439/2015F218 pp. 1641–1648.

[30] K. Stoilova and T. Stoilov, "Bi-level optimization application for urban traffic management," in *2020 15th Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2020. doi: 10.15439/2020F18 pp. 327–336.

[31] S. Fidanova, O. Roeva, and M. Ganzha, "Ant colony optimization algorithm for fuzzy transport modelling," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020. doi: 10.15439/2020F15 pp. 237–240.

[32] V. Caruso and P. Daniele, "A network model for minimizing the total organ transplant costs," *European Journal of Operational Research*, vol. 266, no. 2, pp. 652–662, 2018. doi: 10.1016/j.ejor.2017.09.040

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 553–563
DOI: 10.15439/2025F6496
ISSN 2300-5963 ACSIS, Vol. 43

The Power of Preemptions in Scheduling on Shareable Resources

Omer Lapidot, Tami Tamir
0009-0002-8688-7687
0000-0002-8409-562X
School of Computer Science
Reichman University, Herzliya, Israel
Email: omer2372@gmail.com, tami@runi.ac.il

***Abstract*—Many combinatorial optimization problems arise in the context of resource allocation. In this paper, we study the problem of allocating shareable resources of different types to jobs, where each job consists of multiple tasks, and each task has a demand of a given duration to a single resource type. All resources are available over a common time interval. Several copies from each resource may be allocated. In a valid solution, at any given time, each job may be processed by at most one resource, and each resource may process at most one job. The objective is to complete all jobs while minimizing the total cost of the allocated resources.**

We focus on the power of preemptions in this model, analyzing how much the total cost can be reduced when jobs are allowed to be preempted — that is, when the processing of tasks can be split into multiple intervals.

We present both theoretical and experimental results, distinguishing between environments where jobs may be preempted but all intervals of a task must be processed on the same resource copy (weak preemptions), and environments where jobs may split the processing of a task among different resource copies (strong preemptions). Without preemptions, the problem is clearly NP-hard, as it generalizes the classical Bin Packing problem. We provide an optimal polynomial-time algorithm for the strong-preemption model, as well as a polynomial-time algorithm for the non-preemption model under a restricted class of task durations. Our empirical evaluation investigates the performance of several greedy heuristics, showing that even simple methods can achieve near-optimal results.

I. Introduction

SUSTAINABILITY plays a crucial role in various aspects of our future on Earth. Growing awareness of the need to protect the environment, combined with the ability to communicate easily, has led to new trends in resource consumption. Users have become more flexible about consumption and are willing and able to share resources in new ways. The new ways to consume resources give rise to new combinatorial problems related to allocation and utilization of shareable resources. In this work, we define and study problem of this kind.

We consider a *resource sharing* model, where each resource is capable of being used or leased for any duration within its availability window. For example, consider Amazon drivers who perform deliveries in their assigned regions, using trucks of different sizes tailored to specific delivery needs. A similar setting arises in large organizations, where departments might share meeting rooms of various capacities, or in workplaces where part-time or hybrid employees share desks, cubicles, or workstations. Likewise, in residential communities, neighbors may collaboratively use equipment such as electric vehicle charging stations [18], lawn mowers or other landscaping tools.

In our resource sharing problem, every user (job) is associated with a demand for several resources. For every user and resource, it is known for how long the user needs the resource. Without sharing, every user is allocated one copy from each resource. However, different users can share the same resource if they use it at different times. Assume that every copy of a resource is associated with a given cost. The goal is to schedule the users on different copies of the resources such that the total cost of allocated resources is minimal. In the basic setting of this model, every user is using each resource consequently and with no interruptions. When preemptions are allowed, a user may split the use of a resource into several segments. It is well known that in many scheduling problems, preemptions significantly affect the objective function value and the computational complexity of the problem. For example, the classical load balancing problem of minimizing the completion time of a schedule (the makespan) is NP-hard if preemptions are not allowed [5] and is solved by a linear-time algorithm if preemptions are allowed

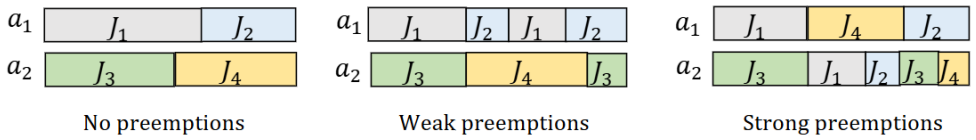


Fig. 1. Possible schedules of four jobs on two copies of resource type a.

[11]. In this work, we study the power of preemption in the resource sharing model. Practically, the power of preemption measures how the flexibility of users in the way they are ready to use the resource affects the total cost.

The resource sharing problem is dual to *open-shop scheduling*, in which we have one copy of each resource type, and the goal is to minimize the schedule's completion time (the makespan). For a single resource type, our problem is identical to the well-known NP-hard *bin-packing* problem in which items of different sizes need to be packed in a minimal number of identical bins. In the bin-packing problem, if items can split among several bins, the problem becomes trivial. In our setting, such splits correspond to job-preemptions. Thus, our problem can be viewed as a generalization of bin-packing. In this generalization with several resource types, the power of preemption has not been studied in the past.

In our analysis we distinguish between four types of resource consumption: (i) no resource sharing, (ii) resource sharing without preemptions, (iii) resource sharing with weak preemptions, where the processing of a task can be paused and resumed on the same resource copy, and (iv) resource sharing with strong preemptions, where tasks may pause and resume, and also migrate across different resource copies. We present an optimal algorithm for the problem with strong preemptions allowed, and analyze theoretically and by experiments the performance of some natural greedy heuristics, in general and for restricted classes of instances. As we show, even simple heuristics can achieve near-optimal performance.

A. Notation and Problem Statement

An instance of the *resource sharing* problem is given by a tuple $\langle \mathcal{J}, M, (p_{ij})_{i\in M, j\in\mathcal{J}}, (c_i)_{i\in M}, T\rangle$, where $\mathcal{J}$ is a set of $n \geq 1$ jobs, M is a set of $m \geq 1$ resource-types, p is an $n \times m$ matrix of processing requests, $c_i > 0$, is the cost of each copy of resource i, and $[0, T]$ is the availability interval of each of the resource copies. We use the term resource i to denote resource-type i. Each job $j \in \mathcal{J}$ consists of several tasks. Formally, each job is associated with a set of resource types $S_j \subseteq M$ corresponding to the tasks composing it (for which $p_{ij} > 0$). For each $i \in S_j$, job j should be processed on a resource of type i for p_{ij} time units. In a valid instance $\sum_i p_{ij} \leq T$. Let $J_i = |\{j \in \mathcal{J} \mid i \in S_j\}|$ be the number of jobs that need resource i.

The idea in the sharing model is that several jobs can share a single copy of a resource, as long as each copy services at most one job in each time slot. Therefore, the number of copies required from resource i may be lower than J_i. It is easy to see that the actual number of copies required from resource i is between $\left\lceil \sum_j p_{ij}/T \right\rceil$ and J_i.

A *non-preemptive schedule* of an instance is denoted by $\sigma = (\sigma_j)_{j\in\mathcal{J}}$. For every job j, $\sigma_j = (\sigma_{ji})_{i\in S_j}$ is the schedule of job j, given by the set of intervals in which j uses the resources it needs. Specifically, $\sigma_{ji} = [start(\sigma_{ji}), end(\sigma_{ji}))$ is the interval in which job j uses resource i in σ. In a feasible solution, $end(\sigma_{ji}) - start(\sigma_{ji}) = p_{ij}$; that is, job j is assigned to the resource for p_{ij} time units. The intervals in which a job uses different resources must be disjoint; that is, for all j and any i_1, i_2, it holds that $\sigma_{ji_1} \cap \sigma_{ji_2} = \emptyset$.

For a given schedule σ, for any time $0 \leq t \leq T$, let $N_i(\sigma, t) = |\{j \in \mathcal{J} \mid t \in \sigma_{ji}\}|$ be the number of jobs $j \in \mathcal{J}$ that use resource i at time t. Let $N_i(\sigma) = \max_{0\leq t<T} N_i(\sigma, t)$. That is, $N_i(\sigma)$ is the maximal number of jobs who simultaneously use resource i throughout the schedule. It holds that $N_i(\sigma)$ is the number of copies of resource i needed to satisfy all job requirements. The cost of a schedule σ is the total cost of the resources used in σ, given by $cost(\sigma) = \sum_{i=1}^{m} N_i(\sigma) \cdot c_i$. The goal is to satisfy all job requirements at a minimum total resource cost, i.e., find a schedule σ such that $cost(\sigma)$ is minimal.

When preemptions are allowed, the processing of a task on one resource may split into several disjoint intervals. We distinguish between two types of preemptions (see Figure 1).

1) Weak Preemptions - for every job j and type $i \in S_j$, the processing of the task of type i of job j may split into several intervals, all of them are on the same copy of a resource of type i.
2) Strong Preemptions - for every job j and type $i \in S_j$, the processing of the task of type i of job j may split into several intervals, which can be processed on different copies of resources of type i.

For an instance I, let $OPT_{nopmtn}(I)$, $OPT_{Wpmtn}(I)$, $OPT_{Spmtn}(I)$ denote the minimal cost of a feasible schedule of I without preemptions, with weak preemptions, and with strong preemptions, respectively.

Definition 1.1: Let I be an instance of a resource sharing problem. The power of weak preemption for I is $\mathrm{PoWP}(I) = \frac{OPT_{nopmtn}(I)}{OPT_{Wpmtn}(I)}$. The power of strong preemption for I is $\mathrm{PoSP}(I) = \frac{OPT_{nopmtn}(I)}{OPT_{Spmtn}(I)}$.

Definition 1.2: For a class $\mathcal{I}$ of instances, let $\mathrm{PoWP}(\mathcal{I}) = \max_{I\in\mathcal{I}} \mathrm{PoWP}(I)$, and $\mathrm{PoSP}(\mathcal{I}) = \max_{\mathcal{I}\in\mathcal{I}} \mathrm{PoSP}(\mathcal{I})$.

When weak preemptions are allowed, a job may split its processing but must remain in the same copy of a resource. The next definition, of the power of migration, isolates the benefit from allowing jobs to split a task among several different copies of a resources.

Definition 1.3: Let I be an instance of a *Resource Sharing* problem and $\mathcal{I}$ a class of instances. The power of migration for I is $\mathrm{PoM}(I) = \frac{\mathrm{PoWP(I)}}{\mathrm{PoSP(I)}}$, and the power of migration for $\mathcal{I}$ is $\mathrm{PoM}(\mathcal{I}) = \max_{I\in\mathcal{I}} \frac{\mathrm{PoWP}(I)}{\mathrm{PoSP}(I)}$.

We conclude the introduction with an example, demonstrating the power of preemption. Consider an instance I with $n = 6$ jobs and $m = 5$ resource types. The jobs' demands are given in the table in Figure 2. Assume that all resources have the same unit cost per copy. Assume also that $T = 5$. Figure 2 presents two possible schedules. The left schedule is an optimal schedule with weak preemptions. In this schedule exactly one copy of each resource is allocated, therefore $OPT_{Wpmtn}(I) = 5$. The right schedule is an optimal non-preemptive schedule. It holds that $OPT_{nopmtn}(I) = 8$. We can see that the left schedule uses exactly $\left\lceil \sum_j p_{ij}/T \right\rceil$ copies of each resource type i, which is clearly the minimal possible. We conclude that $\mathrm{PoWP}(I) = \mathrm{PoSP}(I) = \frac{8}{5} = 1.6$.

The above example can be generalized to show that for every $n \geq 2$, there exists an instance I_n with n jobs and $m = n+1$ resource types of unit cost such that $\mathrm{PoWP}(I_n) = \mathrm{PoSP}(I_n) = \frac{2n}{n+1} = 2 - \frac{2}{n+1}$. In this construction, one job needs all m resources for a short duration ε, while each of the remaining $n-1$ jobs needs a distinct resource for $T - \varepsilon$ time units. In the non-preemptive schedule, the first and last resources can be shared and require a single copy each, while each of the remaining $m-2$ resources requires two copies, yielding a total cost of $2(n-1)+2 = 2n$. In contrast, a preemptive schedule allows all jobs to be assigned efficiently with only one copy per resource, giving a cost of $m = n+1$, and thus the claimed ratio.

$T=5$	J_1	J_2	J_3	J_4	J_5	J_6
a	1	4	0	0	0	0
b	1	0	4	0	0	0
c	1	0	0	4	0	0
d	1	0	0	0	4	0
e	1	0	0	0	0	4

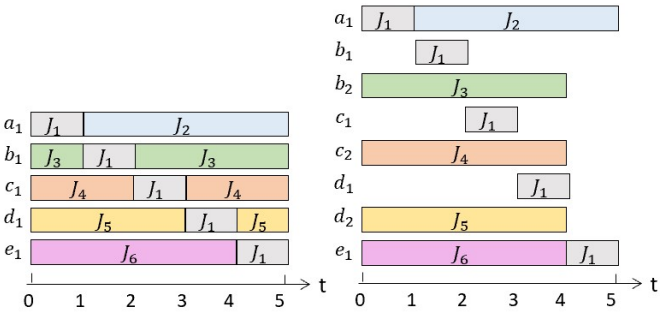


Fig. 2. An example of an instance I for which $\mathrm{PoP}(I) = \frac{8}{5}$.

B. Related Work

There is an extensive body of literature dedicated to job scheduling on parallel machines of various types. As previously mentioned, the resource sharing problem is dual to the well-studied open-shop problem. An instance of *open-shop* scheduling consists of $m \geq 1$ machines (or processors). Each of these machines is capable to perform a single task type. The machines are available at time $t = 0$, and, unlike our model, do not have limited active time, that is, $T = \infty$. In the problem $O||C_{max}$ the goal is to minimize the makespan of the schedule, given by the last completion time of a job. When preemptions are allowed, the problem $O|pmtn|C_{max}$ can be solved in linear-time [6], [10] for any number of machines. In contrast, for the non-preemptive variant, a poly-time algorithm is known only for $m \leq 2$.

For $m \geq 3$, $O||C_{max}$ is NP-complete [6]. A known approximation algorithm guarantees an approximation to the optimal solution that is dependent on the ratio between the maximal total demand for a single machine and the maximal demand of a single job on a single machine [4]. The hardness of the approximation for $O||C_{max}$ has also been studied. It is known that, unless P=NP, there does not exist a polynomial-time approximation algorithm that constructs a schedule with length guaranteed to be strictly less than $\frac{5}{4}$ times the optimal length [17]. When m is fixed, there exists a $(1+\epsilon)$-approximation scheme for the problem that is polynomial in the size of the instance, but exponential in m and $\frac{1}{\epsilon}$ [14]. Optimal algorithm

or better approximal factors were given for restricted classes, such as unit-length jobs [12], [1], or limited number of jobs or machines [3].

The *Power-of-Preemption* (PoP) has been analyzed mainly for the problem $R||C_{max}$ of minimizing the makespan on unrelated parallel machines. In [7] it is shown that the PoP is exactly 4, by showing a lower bound to the PoP, after an upper bound as well as approximation algorithms for restricted classes were given in [2], [13], [15]. The paper [16] studies the PoP and the power of limited preemptions (where only a limited number of preemptions is allowed) for multiple variants of machine environments, job characteristics and objective functions.

Besides the resemblance to the shop-scheduling models, our resource sharing problem with $m = 1$ is identical to the *Bin-Packing* problem (BP). An instance of BP is given by n items, each has a size in (0,1], and the objective is to pack the items into a minimum number of bins, subject to the constraint that the total item sizes in each bin is at most 1. The bin-packing problem has been studied extensively. It is known to be NP-hard and has an FPTAS whose approximation ratio is $(1+\epsilon)$, with a run time of $O(n^{\frac{4}{\epsilon^2}+1})$ [8], [9].

C. Our Results

In this work, we make several contributions that address the scheduling of shared resources under preemptive and non-preemptive settings, and examine how preemptions may reduce the total resource cost. Our results are listed below.

a) Optimal Algorithm with Strong Preemptions: We present a polynomial-time optimal algorithm for the resource sharing problem when strong preemptions are allowed. Our solution is based on a reduction to open-shop scheduling with preemptions ($O|pmtn|C_{\max}$), and reveals that, under strong preemptions, a minimal-cost schedule, in which for all i, there are exactly $\left\lceil \sum_j p_{ij}/T \right\rceil$ copies of resource i, can be computed efficiently.

b) Power of Preemption in the Class $p_{ij} \in \{0, p\}$: For the restricted class of instances in which every task is either of length p or zero, we showed that both non-preemptive and strong-preemptive scheduling can be computed in polynomial time. Yet, the benefit from preemptions may be significant. We prove that the Power of Strong Preemption (PoSP) for this class is bounded by 2, and this is tight.

c) Greedy Heuristics and Empirical Evaluation: Since non-preemptive resource sharing is NP-hard in the general case, we design and evaluate several heuristic algorithms. Our experiments measure how these heuristics perform as problem parameters (number of jobs, resource costs, variation of processing times) vary, and quantify the benefit of preemptions in practice.

Together, these results offer both theoretical and empirical insights into the power of preemptions in shared resource scheduling. We identify tractable cases, establish tight bounds on preemption benefits, and demonstrate through experiments how simple heuristics behave across realistic settings. Our work highlights the fundamental trade-offs between scheduling flexibility, computational complexity, and solution cost.

II. Resource sharing With Strong Preemptions Allowed

In this section we consider settings with strong preemption allowed. We first present and analyze natural greedy algorithms. Similar greedy algorithms are known to be optimal for makespan minimization and for the Bin Packing problem [11]. However, as we show, the greedy approach fails in the resource sharing problem. We then present an optimal, yet more elaborate, polynomial-time optimal algorithm for resource sharing with strong preemptions allowed.

Greedy Algorithm: Algorithm 1 considers the resources one after the other. For every resource i, it assigns the jobs for which $i \in S_j$ one after the other. Every task of length p_{ij} is assigned to the earliest set of disjoint intervals of total length p_{ij} available on open copies of resource i, or on a new copy of resource i if required. The interval must be disjoint with intervals already allocated to previous tasks of job j. Figure 3 presents an execution of Algorithm 1, and demonstrates that it is sub-optimal.

Algorithm 1 - A greedy algorithm for computing a schedule with strong preemptions.

```
All jobs are initially unassigned.
for i = 1 to m do
    for j = 1 to n do
        if i ∈ S_j then: then
            Assign task i of job j on the earliest feasible
            disjoint intervals of total length p_ij on copies of
            resource i. Open a new copy of resource i if
            needed.
        end if
    end for
end for
```

It is easy to see that additional greedy heuristics, which apply a different assignment order (e.g., the outer loop considers the jobs one after the other), are sub-optimal as well,

$T=5$	J_1	J_2	J_3
a	3	3	3
b	2	1	2

Resource demand table

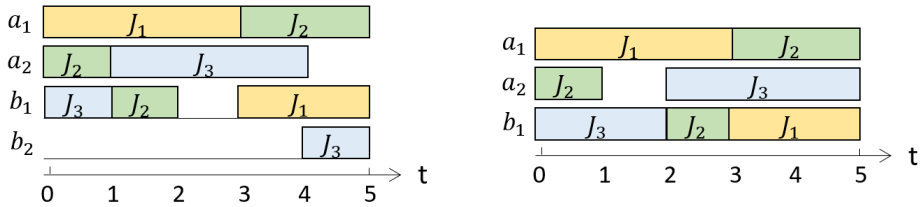


Fig. 3. An example of an instance I with unit-cost resources, demonstrating the suboptimality of a greedy strong-preemptive algorithm. Left: demand table. Center: greedy schedule σ_{alg} with cost 4. Right: optimal schedule $\sigma^\star$ with cost 3.

hinting that an optimal algorithm for resource sharing with strong preemptions, should be more involved.

Optimal Algorithm: Our algorithm for scheduling with strong preemptions allowed employs an optimal algorithm for the problem $O|pmtn|C_{\max}$ as a black box [6]. Recall that in an Open-Shop $O|pmtn|C_{\max}$ problem, we are given a set of n jobs, a set of m machines, and the processing time, p_{ij} of task i of job j. The jobs should be assigned to the machines (a single copy of each machine) such that different tasks of job j are processed in disjoint time intervals. The goal is to minimize the makespan – defined as the last completion time of a job. Let $A_j = \sum_i p_{ij}$ be the total length of job j's tasks. Let $L_i = \sum_j p_{ij}$ be the length of tasks that need to be performed on machine i.

The optimal algorithm for $O|pmtn|C_{\max}$ presented in [6] computes a schedule whose makespan is $\alpha = \max(\max_j A_j, \max_i L_i)$. Since each of these values is a lower bound for the optimal makespan, the algorithm is clearly optimal.

Given an instance of resource sharing, we construct an instance of an open-shop problem as follows:

For every $1 \leq i \leq m$, we define and allocate $\lceil L_i/T \rceil$ machines that will process the tasks of type i. Denote by M_i the set of machines of type i. For every $1 \leq j \leq n$, there is a single job, whose tasks are determined as described in Algorithm 2. The tasks are defined such that for all $i \in S_j$, the total length of tasks on machines in M_i is p_{ij}, and for every machine in M_i, the total length of tasks that need to be processed on M_i is at most T. An example of our reduction is given in Figure 4.

Algorithm 2 - An optimal algorithm with strong preemptions allowed.

1: Given a set J of $n \geq 1$ jobs and a set M of $m \geq 1$ resource types, $p_{ij}, \forall i, j$, and T.
2: For $1 \leq i \leq m$, allocate a set M_i of $\lceil L_i/T \rceil$ machines.
3: Let $m' = \sum_{1 \leq i \leq m} |M_i|$.
4: Construct an instance of $O|pmtn|C_{\max}$ with n jobs and m' machines.
5: **for** $i = 1$ to m **do**
6: **for** $j = 1$ to n **do**
7: **while** $p_{ij} > 0$ **do**
8: Let $i^\star$ be the lowest index of a machine in M_i with total load $l_i^\star < T$.
9: Let $p_{i^\star j} = min(p_{ij}, T - l_i^\star)$.
10: $l_i^\star = l_i^\star + p_{i^\star j}$
11: $p_{ij} = p_{ij} - p_{i^\star j}$
12: **end while**
13: **end for**
14: **end for**
15: Solve $O|pmtn|C_{\max}$ on the resulting instance.

$T=5$	J_1	J_2	J_3
a	3	3	3
b	2	1	2

(a)

	J_1	J_2	J_3
a_1	3	2	0
a_2	0	1	3
b_1	2	1	2

(b)

Fig. 4. An example of our reduction. (a) Resource sharing demand table. (b) Corresponding open-shop instance with $M_a = \{a_1, a_2\}$ and $M_b = \{b_1\}$.

Theorem 2.1: Algorithm 2 computes in polynomial time an optimal solution to the resource sharing problem with strong preemptions allowed.

Proof: Let $\sigma^\star$ be the schedule produced by Algorithm 2. The algorithm reduces the resource sharing instance to an instance of the open-shop problem $O|pmtn|C_{\max}$ by allocating, for each resource i, exactly $\lceil L_i/T \rceil$ machines, and distributing the tasks of each job j across these machines such that the

total processing time per resource equals p_{ij}. This guarantees that the constructed open-shop instance satisfies the time constraints of resource sharing and respects the job-resource requirements.

The schedule $\sigma^\star$ is computed using the optimal algorithm for $O|pmtn|C_{\max}$ from [6], which produces a schedule of makespan $\alpha = \max(\max_j A_j, \max_i B_i)$, where $A_j = \sum_i p_{ij}$ and $B_i = \sum_j p_{ij}$. Our construction guarantees (steps 8-9) that in the open-shop instance, each machine is allocated at most T time units, that is $\forall i, B_i \leq T$. In addition, by definition of the resource rental problem, the total length of each job is at most T, that is, $\forall j, A_j \leq T$. This implies that $\alpha \leq T$, and the resulting makespan is at most T.

Since each machine in the open-shop solution corresponds to one copy of a resource in resource sharing, and all jobs are scheduled such that their tasks are disjoint on each machine, the resulting schedule is feasible in resource sharing under strong preemptions. Furthermore, the number of machines used for each resource is the minimal possible given the total load and capacity T, meaning the total cost (i.e., number of machines) is minimized. Hence, $\sigma^\star$ is a feasible resource sharing schedule with minimal cost. The time complexity of the algorithm is clearly polynomial and is dominated by the time complexity of the algorithm for $O|pmtn|C_{\max}$ [6], which is $\mathcal{O}(r^2)$, where r is the number of non-zero tasks - typically smaller than $\mathcal{O}(n^2)$. ∎

III. The Class $\mathcal{I}_{0,p}$: instances in which $p_{ij} \in \{0, p\}$

In this section we analyze the class $\mathcal{I}_{0,p}$ of resource sharing instances with $p_{ij} \in \{0, p\}$. That is, all the non-empty tasks have the same length p. As a restricted class of the general problem, our optimal algorithm for strong-preemptive applies for this class as well, making the strong-preemptive problem polynomially solvable.

We show that this class is polynomially solvable even when preemptions are not allowed. Given an instance in $\mathcal{I}_{0,p}$, let r_i denote the number of the non-zero tasks which require resource i. Thus, the total demand for resource i is $p \cdot r_i$. Our optimal solution reduces the problem to $O|p_{ij} \in \{0,1\}|C_{\max}$, that is, minimum makespan in an open-shop environment with unit processing times. An optimal poly-time solution for this problem, based on a reduction to minimum edge coloring in a bipartite graph, is presented in [17].

Overview of our algorithm: In the first step, we determine how many copies are required from each resource type. It is easy to see that for all $1 \leq i \leq m$, $\lceil r_i / \lfloor T/p \rfloor \rceil$ is a lower bound for this number. We define an instance of $O|p_{ij} \in \{0,1\}|C_{\max}$ with $\lceil r_i/\lfloor T/p \rfloor \rceil$ machines of type i. We then use the algorithm in [17] to get an optimal schedule for the resulting $O|p_{ij} \in \{0,1\}|C_{\max}$ problem, and conclude an optimal solution to our resource rental problem. Formally:

Theorem 3.1: Computing an optimal solution for a resource sharing problem in $\mathcal{I}_{0,p}$ can be done in polynomial time.

Proof: Let I be an instance of resource sharing in $\mathcal{I}_{0,p}$, with availability time T. For each resource i, let r_i denote the number of jobs j for which $p_{ij} = p$. Since each job requires either 0 or p units on each resource, and each copy of i can process at most $\lfloor T/p \rfloor$ tasks of length p, it follows that at least $\lceil r_i/\lfloor T/p \rfloor \rceil$ copies of resource i are required.

We reduce this instance to an instance of open-shop scheduling $O|p_{ij} \in \{0,1\}|C_{\max}$ with unit-length tasks. Note that since $p_{ij} \in \{0,1\}$, such an instance is defined by determining a set of machines, a set of jobs, and the subset of machines required by each job. For each resource i in the resource rental problem, we introduce $\lceil r_i/\lfloor T/p \rfloor \rceil$ machines of type i, and for every job j with $p_{ij} = p$, we add one of these machines to the subset of machines required by job j. This assignment is done such that each machine is required by at most $\lfloor T/p \rfloor$ jobs. In this way, we guarantee that the load on every machine in the resulting open-shop instance is at most $\lfloor T/p \rfloor$.

The constructed instance is a valid $O|p_{ij} \in \{0,1\}|C_{\max}$ problem, which is known to admit a polynomial-time optimal solution via a reduction to minimum edge coloring in bipartite graphs [17]. The solution is the maximum between the load on a single machine and the total demand of a job. Our construction guarantees that this value is at most $\lfloor T/p \rfloor$. The resulting open-shop schedule induces a feasible resource sharing schedule by stretching each unit-length task to a length p. The resulting makespan is T - that fits the availability time of the resources.

Since our construction uses the minimal number of machines per resource (matching the lower bound), and produces a feasible schedule, it is optimal. We conclude that resource sharing with $p_{ij} \in \{0, p\}$ can be solved optimally in polynomial time. ∎

Next, we bound the power of strong preemption for this class. It is interesting to note that $PoSP > 1$ even though the problem is polynomially solvable in both the preemptive and the non-preemptive models. Thus, preemptions do not affect the computational complexity of the problem but reduce the cost of an optimal solution.

Theorem 3.2: $\text{PoSP}(\mathcal{I}_{0,p}) = 2$.

Proof: Let $I \in \mathcal{I}_{0,p}$ be a resource sharing instance with $p_{ij} \in \{0, p\}$, time T, and resource costs c_i. Let r_i denote

the number of jobs that require resource i. The total load on resource i is therefore $r_i \cdot p$.

As discussed in Section II, the cost of an optimal solution with strong preemptions allowed is

$$\text{cost}_{\text{pre}}(I) = \sum_i \left\lceil \frac{\sum_j p_{ij}}{T} \right\rceil c_i = \sum_i \left\lceil \frac{r_i \cdot p}{T} \right\rceil c_i = \sum_i \left\lceil \frac{r_i}{T/p} \right\rceil c_i,$$

and the cost of an optimal solution *without* preemptions is:

$$\text{cost}_{\text{non-pre}}(I) = \sum_i \left\lceil \frac{r_i}{\lfloor T/p \rfloor} \right\rceil c_i.$$

We first note that if p divides T, then an optimal solution with preemptions allowed, does not use any preemptions, implying that $PoSP(I) = 1$. If p does not divide T, then since $p \leq T$, it holds that $\frac{T}{p} \geq 1$, implying $(T/p)/\lfloor T/p \rfloor < 2$. Therefore,

$$PoSP(I) = \frac{\sum_i \lceil r_i / \lfloor \frac{T}{p} \rfloor \rceil c_i}{\sum_i \left\lceil r_i / \frac{T}{p} \right\rceil c_i} < 2.$$

We show that the above analysis is tight, that is, we can get arbitrarily close to a $PoSP$ of 2. Specifically, given $0 < \delta \leq \frac{1}{2}$, we present an instance I such that $PoSP(I) \geq 2 - \delta$. As we show, a tight example exists already for instances with a single resource type.

Given $0 < \delta \leq \frac{1}{2}$, let $\epsilon \leq \delta$ such that $\frac{2-\epsilon}{\epsilon}$ and $\frac{1}{\epsilon}$ are integers. Let $T = 1$, $m = 1$, $n = \frac{2-\epsilon}{\epsilon}$, $p = \frac{1}{2-\epsilon}$ and $p_{1,j} = p$ for any $1 \leq j \leq n$.

Since $\frac{1}{2} < p \leq 1$, without preemptions no two jobs can share a copy of the resource, therefore, n copies of resource i are required, implying that the cost on any solution without preemptions is exactly $n = \frac{2-\epsilon}{\epsilon} = \frac{2}{\epsilon} - 1$.

On the other hand, the number of copies required with preemptions allowed is

$$\left\lceil \frac{r_1}{\frac{T}{p}} \right\rceil = \left\lceil \frac{n}{\frac{T}{p}} \right\rceil = \left\lceil \frac{(2-\epsilon)/\epsilon}{1/\frac{1}{2-\epsilon}} \right\rceil = \left\lceil \frac{(2-\epsilon)/\epsilon}{2-\epsilon} \right\rceil = \left\lceil \frac{1}{\epsilon} \right\rceil = \frac{1}{\epsilon}.$$

Therefore,

$$PoSP(I) = \frac{2/\epsilon - 1}{1/\epsilon} = \frac{2/\epsilon}{1/\epsilon} - \frac{1}{1/\epsilon} = 2 - \epsilon \geq 2 - \delta.$$

■

IV. Greedy Heuristics and Empirical Evaluation Experiment

A. Key Questions

In our experiments, we addressed the following questions regarding the performance of some natural greedy heuristics and the impact of preemptions in the resource sharing setting:

1) **Evaluate the performance of natural greedy algorithms**
   Since non-preemptive resource sharing is NP-hard, we rely on greedy heuristics as approximate solutions. Our goal is to test whether even simple heuristics produce significant cut down in resource costs.
2) **Investigate how parameter variations influence the efficiency of allocations with no preemptions, weak, and strong preemptions.**
   We systematically vary the number of jobs (n), the number of resources (m) and the variance of distributions of job lengths (p_{ij}). Our goal is to see how these changes amplify or diminish the empirical advantages of preemptions compared to non-preemptive allocations.
3) **Estimating the Power of Preemption.** For the restricted class $p_{ij} \in \{0, p\}$, we proved that PoSP cannot exceed 2 (see Section III). For general instances, we lack a theoretical upper bound for either PoSP or PoWP. To explore them, we compare the performance of our three algorithms (non-preemptive, weak-preemptive, and strong-preemptive). Even while not optimal, these heuristics provide a practical view of preemption benefits across diverse instances.

B. Experiment Design

In order to address the key questions outlined in Section IV-A, we built a simulator that generates a variety of resource sharing instances and evaluated the schedules produced for these instances by several greedy heuristics under different preemptive settings.

We measured the performance of each heuristic by comparing the total cost of its resulting schedule against two baseline values:

- The **optimal cost**, achieved when strong preemptions are allowed, and given by

$$C_{\text{opt}}(I) = \sum_{i \in M} c_i \cdot \left\lceil \frac{\sum_{j \in \mathcal{J}} p_{ij}}{T} \right\rceil.$$

- The **no-sharing cost**, which assumes that each non-empty task $p_{ij} > 0$ is assigned its own dedicated copy. This cost represents a naive allocation strategy and is given by:

$$C_{\text{no-share}}(I) = \sum_{i \in M} c_i \cdot |\{j \in \mathcal{J} \mid p_{ij} > 0\}|.$$

When evaluating the efficiency of various heuristics, we used the no-sharing cost as a normalization baseline.

For simplicity, all instances used in our experiments assume unit resource costs; that is, $c_i = 1$ for all $i \in M$. This

allows us to focus solely on the structural differences between scheduling strategies without the added complexity of cost heterogeneity.

1) Heuristics: We implemented three heuristics to resource sharing, corresponding to non-preemptive, weak-preemptive, and strong-preemptive scheduling. All three heuristics share the same core idea: iteratively assign each job's tasks to available resource copies using a first-fit strategy.

For scheduling with strong-preemptions we have used Algorithm 1, augmented with the following resource-sorting step: Before assigning a job, we sort its required resources in descending order of task duration. That is, for each job j, we reorder its set of required resources S_j according to p_{ij}, scheduling longer tasks first. This helps accommodating shorter tasks in the remaining gaps, and by that improving success rate of sharing the resources. For scheduling with weak-preemptions we used the same algorithm; however, in Step 5, we only considered assignments on a single copy.

Algorithm 3 below describes our heuristic for non-preemptive scheduling. In the non-preemptive setting, partial assignments may block later tasks of the same job and prevent any feasible solution. For example, if $T = 6$ and a job consists of two tasks of length 3, then the second task cannot be performed non-preemptively if the first task is not processed in $[0, 3)$ or in $[3, 6)$. To address this, we keep for each job a variable $left_j$ that stores the leftmost time-point in $[0, T]$ in which it is assigned. We schedule the job using a left-aligned strategy: tasks are scheduled one after the other, with each task placed as early as possible (starting at a leftmost time) and within a carefully restricted interval to ensure that it ends early enough -enabling the remaining tasks to be assigned such that the job completes by time T. If no such interval exists on an already-open copy of the corresponding resource, then a new copy is activated.

All three algorithms share the same Python implementation framework, managing available time segments in each machine's schedule and placing tasks accordingly. The final cost is computed by summing up the number of machines used for each resource.

2) Dataset Generation: Our method for creating benchmark instances generates task durations for each job in a consistent, yet randomized manner. Every instance is characterized by four parameters: number of jobs, n; number of resources, m; task-length variance $0 \leq \texttt{variance} \leq 1$; and job length as a fraction of T, $\texttt{span} \leq 1$.

For a given set of parameters, we construct a set of n jobs each consisting of m tasks, such that (i) for each job $1 \leq j \leq n$ and resource $1 \leq i \leq m$, with probability 0.2 the task of job j corresponding to resource i is empty, that is, $p_{ij} = 0$. (ii) Each of the remaining tasks has a length drawn uniformly from the range $[1-\texttt{variance}, 1+\texttt{variance}]$, and then scaled such that the total tasks length is $\texttt{span}$. We assume that $T = 1$. This is w.l.o.g. since in the last step, the task durations are scaled such that the total job length is $\texttt{span} \cdot T$.

Algorithm 3 Greedy Algorithms for Non-Preemptive Scheduling

1: **for** each job $j \in \mathcal{J}$ **do**
2: Sort S_j (the required resources of j) by non-increasing order of p_{ij}
3: Let $left_j \leftarrow 0$, $p_j \leftarrow \sum_{i \in S_j} p_{ij}$, $scheduled \leftarrow 0$
4: **for** each $i \in S_j$ where $p_{ij} > 0$ **do**
5: Let $remaining \leftarrow p_j - scheduled$
6: Let $latest_end \leftarrow T - (remaining - p_{ij})$
7: Assign task (i, j) as a single interval of length p_{ij} on an existing copy, or a new copy of resource i within $[left_j, latest_end]$
8: Let $left_j \leftarrow$ end time of assigned interval
9: Let $scheduled \leftarrow scheduled + p_{ij}$
10: Update current solution and resource state
11: **end for**
12: **end for**

To evaluate how different structural parameters affect scheduling performance, we generated four large sets of instances, each containing 1000 samples. In each set, we fixed three parameters and systematically varied the fourth: number of jobs, number of resources, variance, or span. This design isolates the influence of each factor on the heuristics' relative performance. When fixed, we used the following parameters: $n = 10$ jobs, $m = 3$ resources, variance 0.9, and span 1. Our experiments show that these values best capture the characteristics of each heuristic.

V. Experimental Results

We begin by presenting the raw results of our experiments. Figures 5- 8 summarize the efficiency of each of the heuristics with respect to four parameters: number of jobs, number of resources, variance, and span. For each experiment (of 1000 samples) we calculated the average solution cost of the three heuristics, as well as the optimal cost. These costs are normalized relative to the no-sharing baseline, giving the *efficiency* of each heuristic, which is depicted in the figures.

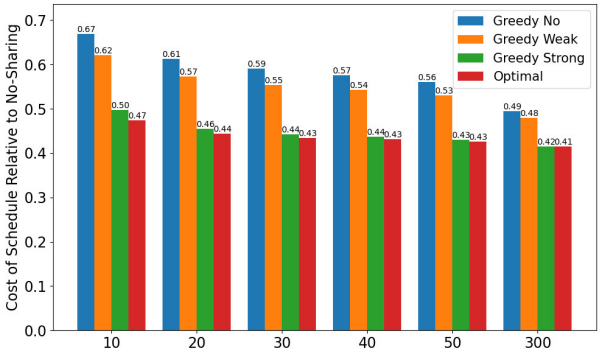


Fig. 5. Efficiency for variable the number of jobs.

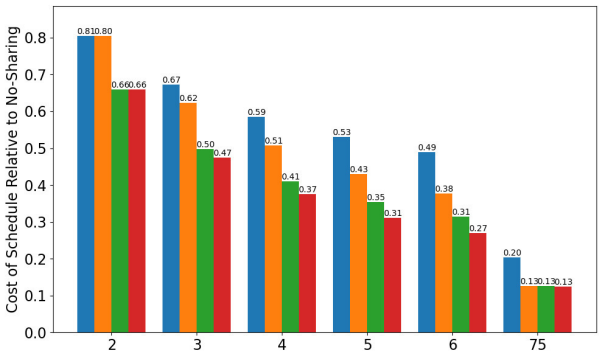


Fig. 6. Efficiency for variable number of resources.

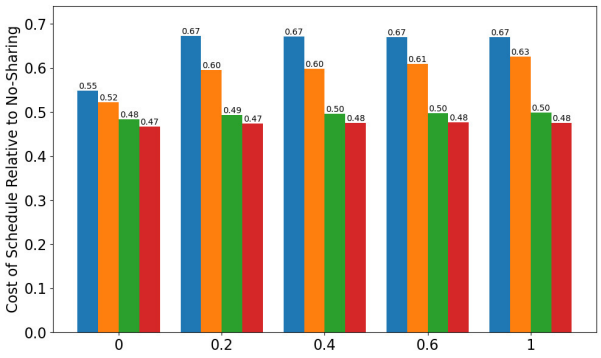


Fig. 7. Efficiency for variable `variance`.

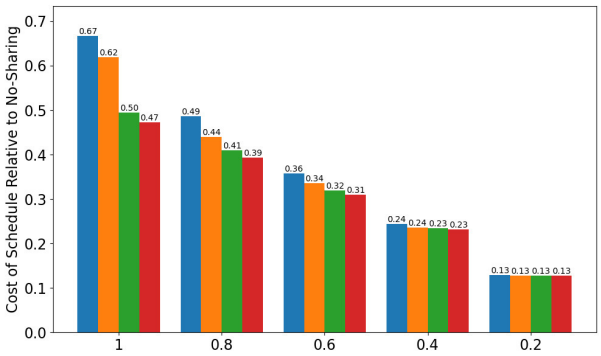


Fig. 8. Efficiency for variable `span` (total job length as a fraction of T).

We now turn to interpreting these results in light of the three research questions introduced earlier.

A. (Q1) How well do the greedy algorithms approximate optimal schedules?

Our results show that all greedy algorithms remain within a constant factor of the optimal strong-preemptive schedule across all tested parameters. In favorable cases, greedy schedules approximate the optimum within ratio 1.1. In more constrained scenarios (e.g., with many resources), the gap may increase, but none of the heuristics exceeded a factor of 2. This empirical observation supports the practical use of these heuristics even in the absence of theoretical guarantees.

B. (Q2) What is the effect of parameter variations on preemption benefits?

We observed how each structural parameter shaped the cost gap between non-preemptive, weak-preemptive, and strong-preemptive settings:

a) Number of Jobs (n): As n increases, all algorithms achieve improved efficiency relative to the no-sharing baseline. The strong-preemptive heuristic continues to closely approximate the optimal cost. Meanwhile, the non-preemptive heuristic gets increasingly closer to the weak-preemptive approach, and both converge to approximately 1.25 times the cost of the optimal solution.

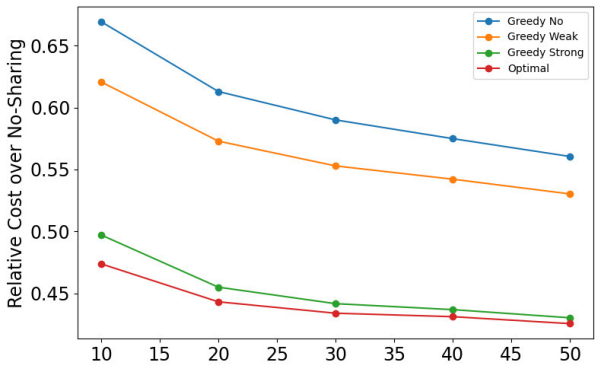


Fig. 9. Relative cost over no-sharing cost as number of jobs increases.

b) Number of Resources (m): As m increases, all of our algorithms achieve improved efficiency relative to the no-sharing baseline. The strong-preemptive heuristic continues to produce near-optimal results, and the weak-preemptive algorithm closely follows it, with the gap between them narrowing as m grows. The non-preemptive heuristic also benefits from increased resources, but to a lesser extent - its relative cost remains around 1.5 times the optimal.

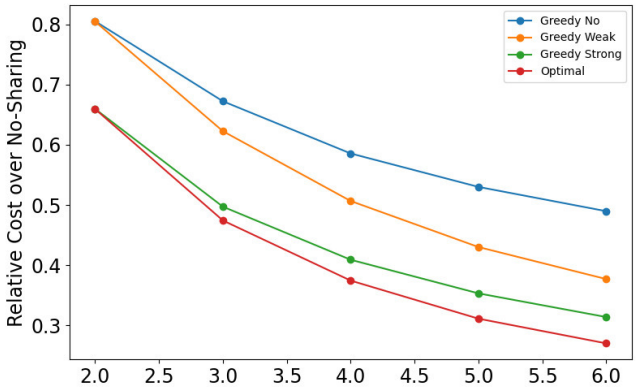


Fig. 10. Relative cost over no-sharing cost as number of resources increases.

c) Variance of Task Durations: Our results indicate that variance in task durations significantly impacts scheduling performance when it first appears (transitioning from zero variance to some positive value). However, beyond this initial effect, the magnitude of variance has minimal influence on the algorithm performance. Once some level of variance exists, increasing it further does not substantially change the gap between the heuristics, suggesting that it is the presence of variance, rather than its magnitude, that defines the difficulty of the instances.

d) Span: Recall that the span determines the total length of the jobs' tasks. As expected, decreasing the span consistently makes the problem easier, since tasks can be more flexibly scheduled without conflicts, leading to lower overall costs. Our results show that beyond a certain point, the problem becomes nearly trivial, with all algorithms, in particular the non-preemptive one, achieving near-optimal results.

C. (Q3) Estimating the Power of Preemption in General Cases

In the general case, across a wide range of tested instances with relatively low values of n and m, the three approaches—non-preemptive, weak-preemptive, and strong-preemptive - are relatively comparable. In those values, we find that non-preemptive schedules cost on average approximately 50% more than strong-preemptive ones, while weak-preemptive schedules cost about 25% more than strong-preemptive schedules. Thus, on average, the *power of strong preemption* is $PoSP(I) = 1.5$, and the *power of migration* is $PoM(I) = 1.25$. While these results are based on smaller parameter settings, the trends observed in earlier sections indicate similar or even stronger effects for larger instances. In particular, as the number of jobs increases, the non-preemptive and weak-preemptive heuristics converges toward a relative cost of approximately 1.25 compared to the optimal. While as the number of resources increases, the weak-preemptive algorithm achieves near-optimal performance, closing the gap with the strong-preemptive approach.

VI. Summary and Future Work

In this paper we analyzed a scheduling problem arising in environments in which resources may be shared by several users. The goal is to minimize the total cost of allocated resources. We summarize below our contribution and the main takeaways from our theoretical and empirical study of the resource sharing problem.

a) Optimal Algorithm for Scheduling with Preemptions Allowed: We developed an algorithm for solving the problem optimally when strong preemptions are allowed. Our optimal algorithm is suitable for all classes of instances, and achieves the lowest possible cost, given by

$$C_{\text{opt}}(I) = \sum_{i \in M} c_i \cdot \left\lceil \frac{\sum_{j \in \mathcal{J}} p_{ij}}{T} \right\rceil .$$

b) Tight analysis of the class $\mathcal{I}_{0,p}$: For the class $\mathcal{I}_{0,p}$, we developed an optimal algorithm for scheduling without preemptions and established a tight bound of 2 for the power of strong preemptions. We conjecture that this bound is valid for additional classes.

c) Greedy Approaches and Approximation: We developed three greedy heuristics for resource sharing: non-preemptive, weak-preemptive, and strong-preemptive, and found that all provide efficient approximations. As expected, the weak and strong preemptive variants consistently outperform the non-preemptive approach. In all parameter regimes, the heuristics remained within a constant factor of the optimal cost.

Our work leaves open several directions for future work:

1) *Theoretical analysis of the general case:* A deeper understanding of the general case is still needed. Two specific directions stand out:
   - Determining tight upper bounds for the power of weak and strong preemptions (PoWP and PoSP), as well as the power of migration (PoM).
   - Deriving theoretical worst-case approximation guarantees for the proposed heuristics, complementing the observed empirical performance.
2) *Analysis of additional restricted classes of instances:* Beyond the $\mathcal{I}_{0,p}$ class, it would be interesting to analyze additional classes of instances, arising in real-world applications.

- Instances with a constant number of job types, where all jobs of the same type have the same resource requirements.
- Instances with job-dependent or resource-dependent processing times, e.g., $p_{ij} = p_j$ or $p_{ij} = p_i$.
- Instances in which each job consists of a bounded number of tasks (e.g., $|S_j| = 2$ for all j).

For each such class, future work should aim to determine whether an optimal solution can be computed in polynomial time, or alternatively prove hardness and develop approximation algorithms.

3) *Modeling costs for preemptions and migrations:* Our model assumes that preemptions and migrations are not associated with a cost. In real-world settings, however, splitting or migrating tasks may involve setup or transition costs. Extending the model to incorporate these costs may yield richer structural insights and lead to new trade-offs between flexibility and efficiency. Analyzing the effect of such costs on both the complexity and the power of preemptions is a promising direction for future work.

REFERENCES

[1] H. Bräsel, D. Kluge, and F. Werner, A polynomial algorithm for the $[n/m/0, t_ij = 1, tree/C_{max}]$ open shop problem, European Journal of Operational Research, 72(1):125–134, 1994.

[2] R. Canetti and S. Irani. Bounding the power of preemption in randomized scheduling. *SIAM Journal on Computing*, 27(4):993–1015, 1998.

[3] I. G. Drobouchevitch. Three-machine open shop with a bottleneck machine revisited. *Journal of Scheduling*, 24(2):197–208, 2021.

[4] T. Fiala. An algorithm for the open-shop problem. *Mathematics of Operations Research*, 8(1):100–109, 1983.

[5] M.R. Garey and D.S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness, 1979.

[6] T. Gonzalez and S. Sahni. Open shop scheduling to minimize finish time. *J. ACM*, 23(4):665–679, 1976.

[7] J.R.Correa, M.Skutella, and J.Verschae. The power of preemption on unrelated machines and applications to scheduling orders. *Mathematics of Operations Research*, 37(2):379–398, 2012.

[8] N. Karmakar and R. Karp. An efficient approximation scheme for the one-dimensional bin packing problem. In *Proc. 23rd IEEE Symp. on Foundations of Computer Science*, pages 312–320, 1982.

[9] A. Levin. Approximation schemes for the generalized extensible bin packing problem. *Algorithmica*, 84(2):325–343, 2022.

[10] C. Liu and R. Bulfin. Scheduling open shops with unit execution times to minimize functions of due dates. *Operations Research*, 36(4):553–559, 1988.

[11] R. McNaughton. Scheduling with deadlines and loss functions. *Manage. Sci.*, 6:1–12, 1959.

[12] B. Naderi, M. Zandieh, and M. Yazdani. Polynomial time approximation algorithms for proportionate open-shop scheduling. *International Transactions in Operational Research*, 21(6):1031–1044, 2014.

[13] A. S. Schulz and M. Skutella. Scheduling unrelated machines by randomized rounding. *SIAM Journal on Discrete Mathematics*, 15(4):450–469, 2002.

[14] S. Sevastyanov and G. Woeginger. Makespan minimization in open shops: A polynomial time approximation scheme. *Math. Program.*, 82:191–198, 1998.

[15] H. Shachnai and T. Tamir. Multiprocessor scheduling with machine allotment and parallelism constraints. *Algorithmica*, 32(4):651–678, 2002.

[16] B. Takand. Towards power of preemption on parallel machines. *MPhil thesis, University of Greenwich*, 2016.

[17] D. P. Williamson, L. A. Hall, J. A. Hoogeveen, C. A. J. Hurkens, J. K. Lenstra, S. V. Sevast'janov, and D. B. Shmoys. Short shop schedules. *Operations Research*, 45(2):288–294, 1997.

[18] R. Zhang, N. Horesh, E. Kontou, and Y, Zhou, Electric vehicle community charging hubs in multi-unit dwellings: Scheduling and techno-economic assessment, *Transportation Research Part D: Transport and Environment*, Vol. 120, 2023.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 565–576
DOI: 10.15439/2025F3271
ISSN 2300-5963 ACSIS, Vol. 43

From Agents to Copilots: A Systematic Review of Digital Assistant Technology Adoption in Proprietary Productivity Software

Niklas Preiß
University of Sheffield Hallam,
Sheffield, United Kingdom
Email: niklas.J.Preiss@student.shu.ac.uk

Markus Westner
0000-0002-6623-880X
OTH Regensburg,
Regensburg, Germany
Email: markus.westner@oth-regensburg.de

***Abstract*—This study provides a systematic review of how the impact and adaptation of digital assistant technologies (DATs) are defined, operationalized, and studied, synthesizing key domains where DATs generate or are expected to generate value. Based on an analysis of 61 articles published since 2013, it identifies five main areas of impact: productivity and efficiency, business development, resource optimization, quality enhancement, and the promotion of learning and creativity. The review highlights DAT adoption across various disciplines and industries, while revealing limited longitudinal research on benefits and adaptation. Key gaps remain in understanding strategic use and sustained impact. Future research should explore longitudinal comparisons of recently introduced generative AI-driven DATs and their organizational implications. This review contributes to information systems research by structuring current knowledge on DAT adoption and outcomes, and by proposing a research agenda to support deeper exploration of their value and long-term integration.**

***Index Terms*—DATs, SLR, generative artificial intelligence, assistant, chatbot, agent, Copilot, Gemini, ChatGPT, productivity software.**

I. Introduction

DIGITAL assistant technologies (DATs)—such as intelligent agents, chatbots, and voice assistants—hold significant transformative potential for large enterprises by automating routine tasks, enhancing data-driven decision-making, and enriching user interactions across various business contexts [1]. The rapid breakthroughs in artificial intelligence (AI), particularly in generative artificial intelligence (GAI), have further amplified the capabilities of DATs sparking a competitive race among organizations to identify and leverage optimal use cases [2], [3], [4]. As a result, substantial investments are being directed toward AI-driven DATs aimed at enhancing productivity, efficiency, and strategic value [5]. These DATs enable automation, reduce non-value-adding or repetitive tasks, and improve user interactions across diverse business contexts. Despite their growing adoption and capabilities, particularly with the emergence of GAI, the specific mechanisms enabling organizations to capture productivity benefits from DATs remain insufficiently understood [1]. This research gap is particularly evident in widely adopted productivity platforms such as Microsoft 365 and Google Workspace, where DAT integration is in its early stages [6], [7]. These platforms dominate the productivity software market: as of February 2024, Google Workspace held over 44% of the global market, and Microsoft 365 approximately 30%, with some estimates exceeding 48%, representing over 345 million active paid users [8], [9], [10]. Given their market dominance, future academic research into these technologies is expected to yield substantial insights regarding their impact on productivity. Therefore, a critical—yet underexplored—area within IS research involves understanding how these technologies are embedded into everyday work processes and what factors influence their successful adoption.

Given these developments, it is also essential to examine how DATs were applied in the pre-GAI era, what impacts were studied, and how these insights can inform the integration of GAI-enhanced DATs within productivity software contexts today.

This systematic literature review (SLR) aims to synthesize and contextualize existing research on DAT adoption, identify key areas of impact, and articulate their operational and strategic implications for organizations. Beyond mapping the current knowledge base, it advances a research agenda that addresses pressing questions arising from the emergence of GAI in DAT-focused IS research. The following research questions guide the focus of this SLR:

- RQ1: How does existing literature evaluate and characterize the impact of DATs in organizational settings, particularly in terms of measurement, intended outcomes, and the reallocation of freed-up capacity?
- RQ2: How does existing literature examine the use of DAT in conjunction with proprietary productivity software in organizational contexts?

The study begins with background information, followed by methodology, findings (including bibliometric and data analyses), discussion addressing the research questions with critical reflections and recommendations. It concludes with key findings, implications, limitations, and a research agenda outlining thematic and methodological directions.

II. Research background

The recent proliferation of AI-related concepts—including machine learning, GAI, and transformer architectures—reflects their growing relevance across organizational contexts [1], [11]. Central to these developments are DATs, which encompass chatbots, intelligent personal assistants, virtual assistants (VA), and intelligent agents [12]. These technologies increasingly manifest in branded forms such as Gemini, Claude, and ChatGPT, illustrating the convergence of GAI

and DATs in organizational environments [13], [14], [15], [16]. GAI, as an advanced form of machine learning, generates novel content (e.g., text, images, audio, and video) by analyzing large-scale datasets [17]. GAI an advanced form of machine learning, has evolved with large language models (LLMs) like ChatGPT and Google Gemini, enabling the creation of novel text, images, audio, and video from large-scale data sets [6], [17], [18]. Unlike traditional predictive AI, it emphasizes creative generation and signals progress toward artificial general intelligence, enabling a wide range of potential use cases for organizations. With expanding multimodal capabilities, LLMs have become key drivers of innovation across sectors and are increasingly deployed through DATs in organizational contexts [1], [19]. Microsoft Copilot for 365 and Google Gemini for Workspace serve as VAs, highlighting the convergence of GAI and workplace productivity—reshaping workflows, enhancing collaboration, and transforming enterprise ecosystems [4], [12], [20], [21].

Productivity software—also known as office productivity software [22], [23]—extends beyond traditional tasks like text writing, spreadsheets, and presentations. It now encompasses intelligent collaboration, automation, and workflow orchestration—tools that are increasingly infused with GAI to support knowledge work at scale [24], [25]. The integration of GAI into productivity software represents a major paradigm shift [22]. Consequently, this study adopts the term "productivity software" to reflect this expanded functionality by products such as Microsoft 365 and Google Workspace.

III. Research Methodology

This section outlines the methodological framework guiding the SLR. The review systematically examines the literature within DAT research [27]. By aggregating and analyzing existing research, it provides an in-depth overview of the current scholarly discourse [28], [29]. Following the structure proposed by [30], the process unfolds in three phases: planning (i), review (ii), and dissemination (iii).

In the planning phase, the study's objectives and research questions—introduced in Chapter 1—are articulated to establish a clear conceptual basis. This phase also involves adopting structured, transparent, and replicable methods to identify, select, and evaluate relevant articles [30]. ensuring methodological rigor. The subsequent phases involve systematic data extraction and in-depth analysis [31], [32].

An initial search in the Scopus database yielded 5,469 articles. During the review phase (ii), iterative keyword and sample refinement using the PRISMA method narrowed the sample to 61 articles, prioritizing relevance and rigor [31]. In the dissemination phase (iii), data were extracted and synthesized, enabling critical interpretation of the findings.

ChatGPT-4o was employed to "identify synonyms, related terms, and language variations" [30, p. 87] for constructing the initial review sample, ultimately resulting in the manual creation of the final search string incorporating findings from [1].

IV. Planning the Review (I): Initial Review Sample

The initial search string was intentionally broad, focusing on keywords related to the three DAT concepts: assistant, agent, and chatbot [1]. Additional keywords such as "digital," "virtual," and "artificial intelligence" were included to define the scope and direction of the research [12]. Adjectives describing broader potential benefits of DAT applications were deliberately excluded to ensure neutrality, avoid bias, and minimize the risk of overlooking relevant research streams or topics. These keywords were specifically targeted within the title, keywords, or abstract of articles.

The search string was structured into two keyword buckets: one representing the DAT concept and the other indicating digital characteristics. The buckets were linked using AND for a systematic search, while keywords within each bucket were combined with OR (see Table I). The first bucket was designed to address jingle-jangle fallacies in the DAT domain, ensuring comprehensive coverage of all relevant research streams and preventing the exclusion of pertinent DATs due to inconsistent terminology [1]. We included journal articles for their recognized credibility [34], [35], [36], [37] and peer-reviewed conference proceedings for their role in providing early insights into emerging knowledge, particularly relevant to IS and GAI research [38]. To ensure comprehensive coverage, our search targeted German and English publications across all subject areas, capturing a wide range of industries benefiting from VA. Our focus was on articles and proceedings published after 2013, restricting the scope to completed research publications. To ensure the inclusion of all relevant and representative IS outlets in the initial sample, we utilized the '3XL' filter from Litbaskets.io, accessed on January 19, 2025, which is a curated set of 847 ranked IS journals [39]. Litbaskets.io effectively supports literature searches via Scopus, the largest multidisciplinary database [40], resulting in a total of 5,469 articles.

TABLE I
INITIAL SEARCH STRING

Basket 3XL / 847 essential IS journals	
searched in title, keywords, or abstract of a article or proceeding	
published after 2013 until January 19th 2025	
English / German only	
Keywords	
Technology **AND**	characteristic
agent* **OR** assistant* **OR** chatbot*	"artificial intelligence" **OR** digital **OR** virtual
5,469 articles	

We utilized VOSviewer 1.6.20 to examine the initial search results, creating visualization maps that depict articles and keywords as circles, where the size and proximity reflect activity levels and the strength of relationships [41]. A thesaurus file was used to standardize variations in spelling and plural forms of keywords (e.g., GAI/generative artificial intelligence, assistant/assistants). 37,415 unique keywords, grouping 1,552 (with a minimum occurrence of 10) into five distinct clusters based on co-occurrence patterns, as shown in Fig. 1. The analysis focuses on keywords related to RQ1—operational efficiency, workforce productivity, and business performance.

Cluster 1 (red) contains 778 items. The most prominent keyword in this cluster is "artificial intelligence," which appears 2,011 times with a total link strength of 19,020, indicating its central role in the research landscape.

A deeper examination of the keyword distribution highlights several terms associated with quality and efficiency. Notable terms include "quality of service" (569, occurrences: 65), "quality of life" (1,194, occurrences: 46), "health care quality" (416, occurrences: 16), "high quality" (97, occurrences: 13), "data quality" (238, occurrences: 11), and "image quality" (266, occurrences: 11). Efficiency is reinforced by "efficiency" (453, occurrences: 46) and "business process" (87, occurrences: 11), underscoring AI's role in optimizing operations and performance.

Cluster 2 (green) contains 442 items and is centered around the keyword "humans" (link strength: 25,335), which appears 1,180 times. Notably, despite having nearly half the occurrences of "artificial intelligence", this keyword exhibits a stronger link strength to other clusters. This extensive connectivity may stem from the focus of research on human-machine interaction.

Rather than presenting a unified theme, Cluster 2 consists of keywords commonly associated with human-machine interaction and healthcare digitization. The intensive linkage between Cluster 1 and Cluster 2 further emphasizes the high research focus and the growing significance of DATs, particularly in healthcare. This is reflected in key terms such as "chatbots" (link strength: 5,839, occurrences: 641), "conversational agents" (2,888, occurrences: 251), "personal digital assistant" (2,959, occurrences: 180), and "artificial intelligence chatbots" (797, occurrences: 62). Supporting this healthcare-driven focus are keywords like "health care" (2,226, occurrences: 132), "digital health" (2,432, occurrences: 101), and "communication" (2,118, occurrences: 112), all of which reinforce the relevance of AI-driven conversational technologies in the medical domain. The keyword "productivity"—appearing 11 times with a link strength of 156—is particularly relevant to the RQs and is positioned between Clusters 1 and 2 (within cluster 2). Its placement indicates the significance of productivity implications across both fields, further emphasizing DATs' potential impact.

Cluster 3 (blue), which contains 183 keywords related to research articles on molecular medicine and drug research.

Similarly, Cluster 4 (yellow) consists of 106 items and is centered around "algorithms" (link strength: 4,676, occurrences: 245). The primary theme of this cluster revolves around computer-aided or assisted diagnosis, as illustrated by keywords such as "computer-aided diagnosis" (446, occurrences: 24) and "computer-assisted diagnosis" (677, occurrences: 23). However, this cluster does not directly relate to the benefits of DATs in the context of large organizations.

Cluster 5 (purple), the smallest cluster with only 31 items, focuses on diabetes and potential treatments. However, its scattered position in the network indicates weak linkage to Clusters 1 and 2 and no significant relevance to the RQs. As a result, Clusters 3–5 were considered peripheral and filtered out.

The initial keyword analysis identifies terms for the final search query. With 5,469 articles, no content analysis was conducted, so keyword interpretations should be considered with caution. Still, three key insights emerge for refinement and expansion in the final search query:

- AI-related terms dominate, emphasizing its role in efficiency, quality, and business operations. Keywords such as "efficiency," "quality of service," and "business process" highlight AI's impact on optimizing workflows and enhancing productivity.
- Cluster 2 focuses on human-machine interaction, particularly in healthcare. Terms like "chatbots," "conversational agents," and "digital health" reflect the presents of DATs. Context analysis is needed to filter out peripheral medical topics.
- To ensure relevance, "quality," "efficiency," "productivity," and "business process" should be included in the next search string. These terms align with the RQs, help exclude irrelevant Clusters 3–5 and will help capture all relevant research streams.

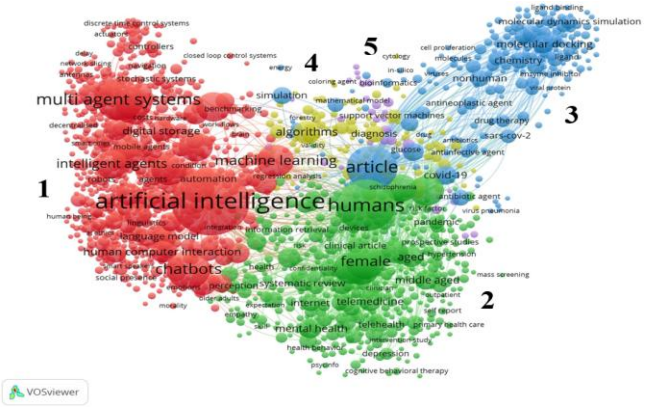


Fig. 1. Keyword co-occurrence network visualization - initial review sample (n= 5,469)

V. Performing the Review (II): Final Search Query

Based on the initial sample analysis, the final search string was refined with additional keywords, resulting in the query: ((agent* OR assistant* OR chatbot*) AND ("artificial intelligence" OR digital OR virtual) AND (quality OR efficiency OR productivity OR "business process*")). This refinement yielded a comprehensive dataset of 1,141 articles, exported on February 1, 2025. Following the PRISMA framework [31] (see Fig. 2), we performed a systematic selection process, including: (a) removal of duplicate or withdrawn articles, (b) screening of titles and abstracts based on predefined exclusion criteria, (c) a comprehensive full-text review to identify articles meeting the inclusion criteria, and (d) backward and forward citation searches to identify additional relevant articles [35], [37]. Following [34] we repeated the process for newly added articles until the sample achieved depth, consistency, and clarity.

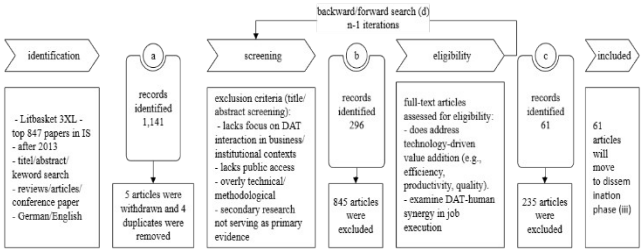


Fig. 2. Steps of the SLR with PRIMSA method according to [31]

VI. Dissemination (iii): Key findings & synthesis

A. Bibliometric Examination of the included dataset

The refinement process (see chapter V) yielded 61 articles containing 723 keywords, visualized in Fig. 3.This analysis identified three primary clusters: (1) DAT technologies and sales/customer service, (2) AI-related functional patterns and DATs, and (3) DAT and AI in research. The keyword 'artificial intelligence' emerged as the central node, serving as the nucleus and primary connector within this dataset.

This significance is driven by the pervasive integration of AI and GAI across various domains, which is "generating significant interest due to its potential to enhance personal efficiency" [4, p. 5]. The bibliometric analysis of the included sample reveals a steady increase in DAT adoption within corporate and institutional contexts. The number of relevant publications has grown from a single publication in 2019 to four in 2020, followed by seven in 2021 and six in 2022. Notably, a sharp increase was observed following the introduction of ChatGPT in November 2022 [18], with 13 publications in 2023 and 26 in 2024. Interestingly, in January 2025 alone, four articles have already been published, suggesting the potential for continued significant growth if this trend persists.

Keyword analysis. Conducting a keyword and co-occurrence analysis on the inclusion sample further reinforces a homogeneous pattern, as illustrated in Fig. 3. A total of 40 keywords met the threshold of three or more co-occurrences, with AI being the most prevalent term (40 occurrences). Notably, six distinct DATs were identified: chatbot (30), ChatGPT (14), artificial intelligence chatbots (4), intelligent assistants (4), conversational agents (4), VA (4), and intelligent agents (3). Copilot for M365 and Gemini for Google Workspace were not present in the key-word co-occurrence analysis. Collectively, these DAT terms account for 23.8% of the full co-occurrence sample. In relation to RQ1, the analysis further identified four keywords representing distinct measures for evaluating the potential impact of DAT integration within the final sample: efficiency (4), accuracy (3), high quality (3), and quality of service (3). These terms emphasize key dimensions in assessing the effectiveness and performance of DAT adoption across different contexts. Notably, productivity no longer appeared as a keyword in this sample.

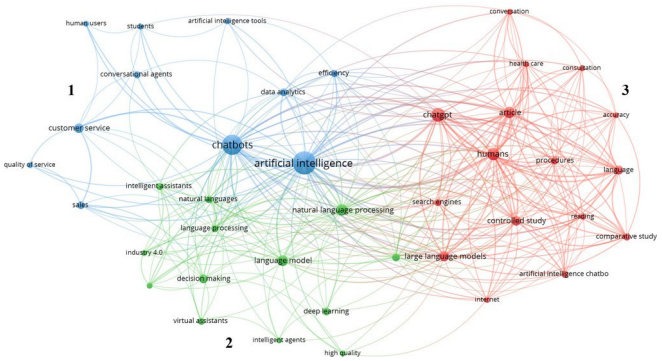


Fig. 3 Keyword co-occurrence network visualization (n=61 articles)

Documents by outlets. The examination of document types revealed the following distribution: journal articles (47; 77%) and conference papers (14; 23%). The prevalence of journal articles suggests that the topic has been extensively studied within a short period of time (since 2019), indicating strong scholarly engagement in the field [42]. The 61 publications analyzed were spread across 47 distinct outlets (see Table II). The top three publication outlets were the Journal of Medical Internet Research (6 articles; 15.4%), IFIP Advances in Information and Communication Technology (3 articles; 7.7%), and CEUR Workshop Proceedings (3 articles; 7.7%). This distribution underscores the interdisciplinary nature of the research, with contributions spanning medical informatics, information and communication technology, and other IS outlets. This emphasis likely accounts for the human-centered cluster 3 in Fig. 3, which highlights extensive research on human-machine interaction and healthcare.

Table II

Distribution of journals by total count and citation sum in the final sample (count >2)

Journal	Count		Sum citations/journal	
Journal of Medical Internet Research	6	9.8%	87	11.9%
IFIP Advances in Information and Communication Technology	3	4.9%	9	1.2%
CEUR Workshop Proceedings	3	4.9%	4	0.5%
Lecture Notes in Business Information Processing	2	3.3%	12	1.6%
International Journal of Medical Informatics	2	3.3%	16	2.2%
International Conference on Software Business	2	3.3%	2	0.3%
Expert Systems with Applications	2	3.3%	74	10.1%
Electronic Commerce Research and Applications	2	3.3%	4	0.5%
Other 39 journals count < 2	39	63.9%	734	77.9%
Total	61	100.0%	942	100.0%

Co-Authorship and author analysis. The analysis identified the key contributors within this sample, revealing that only 13 researchers had at least two co-authorships, defined as collab-

orative efforts between scholars contributing to the same publication. This group represents less than 12% of the total 285 authors in the dataset. Notably, these 13 researchers accounted for 43.7% of the total citation count, with Yu Chen having two co-authorships and 200 citations. The small number of scholars with >1 article underscores a potential gap indicating that relatively few researchers are consistently contributing within the domain. Among the 285 contributors, only one network is linked through six publications, primarily due to the work of Massimo Mecella. His research encompasses conversational agents in business process models [43], chatbots in digital factories [44], and an assistant for legal users [45]. The scarcity of strong interconnections and collaborative efforts among re-searchers highlights a highly segmented field with limited cooperation [46]. From an article perspective, the top 15 most influential publications in the sample account for 811 citations, representing 86.1% of the total 942 citations. Online appendix Table I highlights these key contributions, illustrating their impact on the overall sample through substantial citation counts. Overall, the absence of strong connections between authors, as indicated by link strength, may suggest a lack of collaborative research efforts within the DAT domain.

B. Data Analysis of existing literature

Data analysis – type of research. We analyzed the sample and present an overview of the research design and approach in Table III, with article-level details illustrated in Fig. 4 [47]. The findings reveal a strong emphasis on empirical research, with 56 articles (91.8%) adopting empirical approaches. Among these, quantitative designs are most common (26/42.6%), typically aligned with positivist (21/34.4%) and post-positivist (11/18.0%) worldviews (see Fig. 4), reflecting a focus on objectivity, measurement, and causal inference.

TABLE III

OVERVIEW OF RESEARCH DESIGN AND APPROACH

Design	count	Research approach	count
Quantitative	26	experimental	14
		survey	6
		comparative assessment	5
		interviews	1
Qualitative	15	experimental	4
		design science res.	3
		case study	2
		interviews	2
		co-design study	1
		comparative assessment	1
		survey	1
		systematic mapping	1
Mixed method	15	design science res.	6
		experimental	4
		convergent parallel	2
		exploratory sequential	2
		comparative assessment	1
Non-empir.	5	design science res.	3
		conceptual analysis	1
		practise-based	1

Notably, not a single study employed longitudinal research design. Despite the dominance of empirical methods, all 61 articles relied on alternative research approaches, as detailed throughout this chapter. This absence of longitudinal inquiry highlights a significant gap in understanding the sustained use, adaptation, and long-term impact of DAT within real-world contexts [48]. It underscores the importance of future research that extends beyond immediate performance metrics to investigate how users, systems, and organizational environments co-evolve over time.

Qualitative designs (15/24.6%) show stronger ties to interpretivist perspectives (7/11.5%), aiming to capture contextual depth and user experiences, though still secondary in representation. Mixed methods articles (15/24.6%) often reflect a pragmatist worldview (22/36.1%), blending quantitative and qualitative paradigms to pursue practical, problem-solving goals. The dominance of pragmatist and positivist orientations (shaped by inclusion criteria) reflects a focus on outcome-driven research, emphasizing quantifiable data and replicable cause-and-effect relationships [49]. The remaining non-empirical research (5/8.2%)—including design science, conceptual analysis, and practice-based work—supports theoretical exploration and methodological innovation, though it remains limited in volume. Overall, the analysis reflect a field focused on solution-oriented inquiry, emphasizing real-world performance, usability, and iterative refinement of DAT [50], [51].

While empirical inquiry dominates, the limited use of established IS theories reinforces the field's developmental stage. Only four articles explicitly employed IS theories, further underscoring this empirical orientation [52]. These include the Technology Acceptance Model (TAM) and DeLone & McLean IS Success Model [53], the Decomposed Theory of Planned Behavior [54], Fit-Viability Theory [55], and Attribution Theory [55]. The limited use of theory reflects the field's exploratory, application-driven nature, where technical feasibility often takes precedence. However, as DAT become more embedded in daily life, the growing uptake of IS theories since 2021 signals a timely shift toward deeper conceptual integration and a stronger focus on adoption. [52].

Notably, purely secondary research (e.g., SLRs, taxonomies, or frameworks) was excluded in the PRISMA process. In sum, research on DAT is empirical, pragmatic, and impact-driven, with a clear focus on performance and usability. However, the limited use of IS theories and interpretive lenses signals a need for deeper conceptual grounding and contextual insight [50].

Data Analysis – Concepts and DAT used in final sample. We further analyzed the concepts and technologies referenced in each publication to understand their application across various domains and use cases. Our review found that the assistant concept appears 34 times (55.7%), making it the most prevalent, followed by the agent concept, which is featured in 22 articles (36.1%). Some articles discuss multiple concepts or DATs, which is why the total does not sum to 61 in the more detailed Fig. 4. For example, Copilot and Gemini fall

| Reference | Entity: corporate | Entity: institutional | Segment/industry: academia | automotive | process management | construction | e-commerce | education | finance | general | healthcare | insurance | justice | logistics | manufacturing | marketing | public sector | social media | software development/coding | User group: authorities | bunker | business decision maker | customer service | customer service/sales agents | dentistry staff | engineers | knowledge worker | legal users | librarian | manufactoring worker | market maker | medical staff | mental health staff | patients | platform providers/content moderation | process designer | programmer | purchaser/buyer | questioner | researcher | students | teaching staff | trader | DAT concept studied: assistant | agent | chatbot | assistant Bard | assistant ChatGPT | assistant Claude | assistant Copilot | assistant Gemini | assistant Bloomz | AI technology: DL | GAI | ML | NLP | RL | SL | unspecified | Research setup: Philosopical worldview | Research design | Reserach approach | # Total quant sample size |
|---|---|---

into how each segment adopts and prioritizes DATs—particularly valuable for future research focused on specific sectors, such as large enterprises. 17 industries or segments were also identified, listed alphabetically by frequency. Fig. 4 presents the coded matrix of user groups within their respective application domains. The distribution in Fig. 4 underscores the multidisciplinary scope of DAT applications, with healthcare and general applications as the dominant domains, followed by software development/coding, education, and manufacturing. The term general (14/23%) encompasses publications with a broad scope, spanning multiple industries, general applications, or lacking a distinct focus. The most prominent industry-specific fields are healthcare (15/24.6%) and software development/coding (5/8.2%). Together, these three fields account for 34 (55.7%) of the total sample, illustrating the primary domains of the researched domain. However, a significant portion of the re-search remains broad and use-case unrelated, categorized under "general." This imbalance may suggest a need for more specialized research to generate deeper insights across various fields and broaden the understanding beyond the dominant healthcare domain.

Fig. 4 also illustrates the distribution of user groups addressed by the DAT, along with their respective counts and percentages. The most present user groups overall are knowledge workers (11/18.0%), medical staff (11/18.0%), and customer service (7/11.5%) which includes the single-occurrence customer service and sales agent group. In the corporate sector, knowledge workers dominate (10/16.4%), followed by customer service roles (7/11.5%), which exist exclusively in the corporate environment in this sample. The institutional sector is primarily driven by healthcare-related professionals (9/14.8%), though healthcare also appears in the corporate sector with 4 occurrences (6.6%), indicating its cross-sector relevance. Interestingly, DATs are also well established in supporting both institutional and corporate users in programming and software development (6/9.8%) across multiple domains, highlighting their significance in technological and digital advancements.

From a technology perspective, corporate research primarily focuses on assistants (18/45%) and chatbots (17/4%), which together represent 88% of articles in this category. One study examined both concepts jointly. Agents (4/10%) play a minor role in corporate research. A similar but more pronounced trend is observed in institutional research, where assistants account for 76% of articles and chatbots for 24%, primarily in healthcare, academia, and education. Agents are absent in institutional research, as they typically operate within clear and closed boundaries, handling structured tasks [5]. A detailed list of DAT occurrences in each article and its corresponding domain is provided in Fig. 4.

VII. Main findings

This section represents the core contribution of this SLR, synthesizing key findings on DAT applications, their intended functions, and the metrics and objectives guiding their implementation. It directly addresses RQ1 and RQ2 by examining the distinct characteristics of DAT utilization across the 61 analyzed articles. Table IV provides a structured overview of the five focus clusters identified through open coding, which were subsequently categorized into five application goal clusters using axial coding [60]. Each of these clusters is further explored and described in the following sections. The most prevalent subcategory, efficiency/quality, was examined in 13 articles and is analyzed in greater detail in online appendix Fig. 1. The categories will be briefly outlined, with an emphasis on the key measures that define them. Since some subcategories combine multiple measures within individual articles, not every specific combination will be discussed separately. However, their occurrence has been systematically coded within the analyzed articles.

Productivity & efficiency. Productivity and efficiency form the most dominant cluster (32/52.5%), serving as key drivers in business and institutional settings by optimizing resource use, accelerating operations, and availability [28, 65]. A major focus is efficiency and quality (13/21.3%), where DATs streamline processes while maintaining high standards, reducing manual effort and improving accuracy [61]. For example, Calisto et al. (2021) examined the impact of DATs on workflow efficiency and diagnostic quality in breast cancer [62], while Belhaj et al. (2021) demonstrated how such technologies can enhance the quality and efficiency of student-oriented services [63]. Productivity (9/14.8%) is closely linked, as DATs boost output by minimizing inefficiencies and accelerating task execution [5]. Brachten et al. (2021) argue that "chatbots should be positioned as personal technologies designed to support users and enhance productivity" [49 p. 11]. Efficiency (5/8.2%) reflects the role of DATs in eliminating bottlenecks and optimizing workflows, particularly in enterprise automation [28]. Some articles highlight a combined impact on efficiency and productivity (2/3.3%), where DATs simultaneously reduce resource waste and enhance overall performance [66]. Beyond speed and optimization, efficiency and effectiveness (1/1.6%) emphasize DATs' role in enhancing goal achievement and adaptive decision-making [67]. Finally, productivity and health protection (1/1.6%) highlight how DATs support both output and worker well-being [68].

Business growth & development. Business growth and development emerges as the second most prominent cluster (14/23%), underscoring the pivotal role of DATs strategic business expansion, decision-making, and revenue optimization. By facilitating value-driven tasks, business scalability, and cost efficiency, DATs enhance organizational competitiveness [44], [64], [65]. A key application goal within this cluster is enhancing focus on value-added tasks (7/11.5%) by reducing redundant activities, thereby improving overall performance, quality, and work-load efficiency. De Luzi (2023) finds that ChatGPT reduces processing time in legal work, freeing up time for higher-value tasks [45]. By automating mundane tasks, DATs enable employees to prioritize strategic and creative functions over routine administrative work, fostering innovation and workforce optimization, particularly in

knowledge-intensive industries [45], [66]. Closely linked, decision-making support, efficiency, and quality (3/4.9%) reflects DATs' capacity to enhance analytical capabilities, enabling organizations to navigate dynamic market conditions with optimized strategies [56], [67]. Kudryavtsev (2024) states that AI assistants enhance business development by accelerating design processes, lowering decision-making risks, and offering accessible, professional-level support [56]. Business expansion and development (2/3.3%) highlights DATs' contributions to scalability, market entry, and operational agility [56], [68]. The financial impact is evident in revenue growth and cost reduction (2/3.3%), where AI-driven insights and automation maximize profitability while minimizing expenses [54], [69].

Cost & resource optimization. Cost & resource optimization (7/11.5%) remains a critical objective, ensuring organizations maximize output while minimizing financial and operational inefficiencies. DATs support these goals through automation, predictive analytics, and streamlined processes, leading to enhanced cost efficiency and resource utilization [12]. Capacity increase (3/4.9%) highlights how AI-driven automation enables organizations to handle higher workloads without proportionally increasing resources, fostering scalability in customer service, operations, and production [14], [21], [70]. Cost reduction (3/4.9%) underscores DATs' role in eliminating inefficiencies, automating repetitive tasks, and optimizing resource allocation, significantly lowering operational expenses [71]. Bird and Lotfi (2024), for example, strengthen the narrative that chatbots in customer service are more cost-effective than fully human-operated models[72]. Organizations integrating intelligent automation in HR, finance, and customer service achieve substantial reductions in labor costs and time-intensive processes (cost-)effectiveness/process automation (1/1.6%) further demonstrates how DATs enhance efficiency through intelligent workflows, reducing manual interventions and improving process standardization across business functions [64].

Quality. The quality cluster (4/6.6%) highlights the role of DATs in enabling and maintaining higher standards across products, services, and decision-making. DATs enhance quality (2/3.3%) by automating tasks, minimizing human errors, enabling data-driven decisions, and accelerating creativity [73]. The combined category of quality and productivity (2/3.3%) emphasizes the dual impact of DATs in enhancing service reliability while optimizing operational efficiency [74]. Chen et al. (2023) confirm that AI chatbots improve service quality, thereby strengthening customer loyalty and driving business success [75].

Creativity and learning. The creativity and learning cluster (4/6.6%) highlights the expanding role of DATs in knowledge work, idea generation, and creativity enhancement. These technologies lower entry barriers to learning and support continuous professional development [76], [77], [78]. Creativity and efficiency (2/3.3%) enabled by DATs enhance ideation, design, and content creation while streamlining workflows and reducing cognitive load, driving faster innovation and execution [65], [79]. Cheng et al. (2023) positively evaluated the use of AI to increase efficiency in abstract writing for preclinical research [80]. Similarly, learning (2/3.3%) leverages AI-powered training to support skill development, personalized learning, and adaptive programs, improving feedback and knowledge retention by, for example, reducing language barriers and enhancing content accessibility [53], [81].

TABLE IV

KEY FOCUS AREAS OF DAT APPLICATION

Productivity & efficiency (32/52.5%)	**Business growth & development (14/23%)**
Efficiency/quality (13/21.3%)	Focus on value added tasks (7/11.5%)
Productivity (9/14.8%)	Decision making support/efficiency/quality (3/4.9%)
Efficiency (5/8.2%)	Business expansion/development (2/3.3%)
Efficiency/productivity (2/3.3%)	Revenue growth/cost reduction (2/3.3%)
Efficiency/effectiveness (1/1.6%)	**Quality (4/6.6%)**
Efficiency/quality/productivity (1/1.6%)	Quality (2/3.3%)
Productivity/health protection (1/1.6%)	Quality/productivity (2/3.3%)
Cost & resource optimization (7/11.5%)	**Creativity & learning (4/6.6%)**
Capacity increase (3/4.9%)	Creativity/efficiency (2/3.3%)
Cost reduction (3/4.9%)	Learning (2/3.3%)
(cost-)effectiveness/process automation (1/1.6%)	

VIII. RESEARCH AGENDA

Based on the five main focus areas identified through our SLR, we propose a research agenda that highlights key gaps and formulates future research questions within the most relevant and emerging research streams (see Fig. 5). Our analysis shows that the adoption of GAI-driven DATs hold the potential to reshape productivity, efficiency, business models, and strategy [55], [82], [83]. However, a critical gap remains: the lack of empirical, longitudinal evidence on the impact of DATs. While many articles focus on DAT implementation and task- or user-level benefits, few explore how organizations leverage the capacity freed by these technologies or how such changes influence strategic and operational adaptation—an area warranting deeper investigation [84], [85], [86]. Despite the widespread adoption of Microsoft 365 and Google Workspace, these platforms are notably absent from the current research sample. This is also true for articles addressing post-adoption dynamics, largely due to the novelty of GAI-enabled DATs [87]. Although our findings suggest positive effects, comprehensive empirical validation—particularly within large enterprise contexts and across industries—remains limited. To address these gaps, we suggest leveraging established IS theories [88]. The TAM can support research on individual adoption and sustained use. The technology–organization–environment (TOE) framework offers a holistic lens for analyzing organizational adoption conditions. For

post-adoption dynamics, dynamic capabilities theory can explain how firms reconfigure resources to generate long-term value, while adaptive structuration theory provides insight into how DATs shape and are shaped by evolving work practices. We recommend longitudinal and case-based methods to trace adoption trajectories and long-term impact, complemented by controlled experiments, expert interviews, and large-scale surveys [42], [47], [59]. Beyond performance metrics, articles should examine structural shifts—such as evolving roles, skills, and workforce distribution—and their implications for strategy and competitiveness. Fig. 5 outlines potential research questions across five key research streams, organized into two rows: the first focuses on motivations and considerations for adopting GAI-enabled DATs in proprietary productivity software; the second highlights organizational shifts resulting from their usage and adoption.

IX. Conclusion and Limitations

This SLR analyzed 61 articles on organizational support via DATs and their strategic implications. In response to RQ1, the review identified five primary areas of impact: productivity and efficiency enhancement, strategic business growth, resource optimization, quality improvement, and the promotion of creativity and learning. Regarding RQ2, the findings highlight a lack of research on long-term strategic adaptations, largely due to the novelty of GAI-enabled DATs. Addressing this gap is essential to advance scholarly understanding of effective DAT implementation.

Limitations of the paper at hand include exclusive reliance on IS journals indexed in Scopus, potentially overlooking interdisciplinary insights. Additionally, the frequency-based analytical methods and VOSviewer utilized in this study may introduce algorithmic biases and do not permit the establish-

	Research stream				
	Productivity & efficiency	**Business growth & development**	**Cost & resource optimization**	**Quality**	**Creativity & learning**
Key focus areas for organizations using GAI-enabled DATs in proprietary productivity software. (referred to as DAT for readability in this table)	• To what extent do DAT impact productivity and operational performance over a X-year period? • By how much do DAT reduce mental health sick leaves by alleviating high workloads? • How do proprietary DATs transform workplace productivity and efficiency? • What is the critical mass of adoption within companies required to realize productivity gains from Copilot? • What key tasks see the greatest efficiency improvements through DAT integration?	• How can DATs impact organizational expansion? • What measurable impact do DATs have on business cycle time reduction? • By how much does market capitalization increase through DAT implementation? • How do DATs qualitatively reshape strategic business growth and value creation processes? • What business areas demonstrate significant growth potential due to DAT deployment? • What new strategic directions can organizations pursue based on DAT capabilities?	• What are the cost savings achievable through DAT? • How does the implementation of DAT impact the average number of work hours saved per employee per week? • What specific business functions achieve the highest cost savings through DAT implementation? • What specific business functions achieve the highest cost savings through DAT implementation? • What resource allocation strategies are most effective when incorporating DAT-driven automation?	• How do DAT influence work/output quality (e.g. NPS, customer satisfaction)? • To what extent do DAT help reduce cull across different work processes? • What changes in quality and service delivery occur through DAT integration? • What quality assurance processes are most significantly improved through DAT integration? • What quality-related KPIs show measurable improvement due to DAT usage?	• How many additional non-native speakers are employed through DAT application and support? • What alterations in creative processes and learning methods are driven by DATs? • What creative processes/job benefit most from DAT support? • What competencies/skills are essential for maximizing the effectiveness of DAT usage? • What learning outcomes are most significantly enhanced by DAT integration?
Organizational shifts driven by GAI-enabled DATs in proprietary productivity software.	• How does overall business performance change after reallocated capacity is freed up? • What changes in workforce roles and team dynamics are driven by DATs? • What is the critical mass of adoption within companies required to realize productivity gains from Copilot? • How does DAT implementation affect workforce dynamics, job roles, and team performance?	• What impact can DAT have on organizational strategy, overall performance, and market perception? • What proportion of business growth can be attributed to improved operational efficiency from GAI-enabled DATs • What changes in organizational structure and governance are attributed to DAT implementation? • How do DAT reshape organizational strategies and business processes? • What new strategic directions can organizations pursue based on DAT capabilities?	• How effectively do organizations measure the financial impact of DAT-driven resource reallocation? • How do companies perceive and manage cost savings and resource reallocations? • What impacts occur from resource redistribution post-DAT implementation? • What resource allocation strategies are most effective when incorporating DAT-driven automation?	• How do organizations track quality improvements from DAT integration? • What adaptations occur in governance to maintain quality post-DAT implementation? • What quality assurance processes are most significantly improved through DAT integration? • What KPIs show measurable improvement due to DAT usage?	• How can organizations assess DAT-driven innovation outcomes? • How much additional innovation is driven by DAT implementation? • How do DATs influence creativity, learning processes, and innovation within organizations? • What creative processes benefit most from DAT support? • What learning outcomes are most significantly enhanced by DAT integration?

Fig. 5. Indicative potential research questions in DAT research

ment of causal relationships [41], [89]. Finally, the narrow search strategy could have resulted in biases toward specific terminologies or perspectives, thereby limiting the broader applicability of findings due to terminological ambiguities, commonly referred to as jingle-jangle fallacies, surrounding DAT [1]. Future articles should broaden their methodological scope to comprehensively capture interdisciplinary insights and mitigate these limitations.

APPENDIX

Additional data in form of an online appendix is provided here: https://doi.org/10.6084/m9.figshare.28776362.v1

REFERENCES

[1] N. Preiß and M. Westner, 'Unraveling jingle-jangle fallacies in digital assistant technologies: A comprehensive systematic review and research agenda', in *ICICT 2025: 10th International Congress on Information and Communication Technology*, London, 2025.

[2] P. Amiri and E. Karahanna, 'Chatbot use cases in the Covid-19 public health response', *Journal of the American Medical Informatics Association*, vol. 29, no. 5, pp. 1000–1010, 2022, doi: 10.1093/jamia/ocac014.

[3] M. McKillop, B. R. South, A. Preininger, M. Mason, and G. P. Jackson, 'Leveraging conversational technology to answer common COVID-19 questions', *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 850–855, 2021, doi: 10.1093/jamia/ocaa316.

[4] H. Al Naqbi, Z. Bahroun, and V. Ahmed, 'Enhancing work productivity through generative artificial intelligence: A comprehensive literature review', *Sustainability*, vol. 16, no. 3, Art. no. 3, Jan. 2024, doi: 10.3390/su16031166.

[5] D. Czarnitzki, G. P. Fernández, and C. Rammer, 'Artificial intelligence and firm-level productivity', *Journal of Economic Behavior & Organization*, vol. 211, pp. 188–205, 2023, doi: 10.1016/j.jebo.2023.05.008.

[6] Google, 'Introducing Gemini: our largest and most capable AI model', Google. Accessed: Aug. 21, 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/

[7] Microsoft Corporation, 'Introducing Microsoft 365 Copilot: your copilot for work'. Accessed: Sep. 07, 2024. [Online]. Available: https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work

[8] L. S. Vailshery, 'Office productivity software global market share 2025', Statista. Accessed: Apr. 24, 2025. [Online]. Available: https://www.statista.com/statistics/983299/worldwide-market-share-of-office-productivity-software/

[9] J. D'Souza, 'Microsoft 365 Statistics By Downloads, Subscribers, Revenue and Facts', Electro IQ. Accessed: Apr. 24, 2025. [Online]. Available: https://electroiq.com/stats/microsoft-365-statistics/

[10] M. McDade, 'Microsoft 365 Usage and Security Statistics for 2025', Expert Insights. Accessed: Apr. 24, 2025. [Online]. Available: https://expertinsights.com/email-security/microsoft-365-usage-and-security-statistics-for-2024

[11] D. Darwish, 'Chatbots vs. AI chatbots vs. virtual assistants':, in *Advances in Computational Intelligence and Robotics*, D. Darwish, Ed., IGI Global, 2024, pp. 26–50. doi: 10.4018/979-8-3693-1830-0.ch002.

[12] P. Preiß and M. Westner, 'Towards a taxonomy for digital assistant technologies: Addressing the jingle-jangle fallacies', in *Proceedings of the 38th Bled eConference*, 2025.

[13] J. Uszkoreit, 'Transformer: A novel neural network architecture for language understanding', 2017, [Online]. Available: https://blog.research.google/2017/08/transformer-novel-neural-network.html

[14] E. Han, D. Yin, and H. Zhang, 'Bots with feelings: Should AI agents express positive emotion in customer service?', *Information Systems Research*, vol. 34, no. 3, pp. 1296–1311, 2023, doi: 10.1287/ISRE.2022.1179.

[15] T. Li, X. Zhang, Y. Wang, Q. Zhou, Y. Wang, and F. Dong, 'Machine learning for requirements engineering (ML4RE): A systematic literature review complemented by practitioners' voices from Stack Overflow', *Information and Software Technology*, vol. 172, 2024, doi: 10.1016/j.infsof.2024.107477.

[16] E. Brynjolfsson, D. Li, and L. Raymond, 'Generative AI at work', *The Quarterly Journal of Economics*, no. 140, pp. 889–942, 2023, doi: 10.3386/w31161.

[17] S. Hazmoune and F. Bougamouza, 'Using transformers for multimodal emotion recognition: Taxonomies and state of the art review', *Engineering Applications of Artificial Intelligence*, vol. 133, no. 6, pp. 107–113, Jul. 2024, doi: 10.1016/j.engappai.2024.108339.

[18] OpenAI, 'Introducing ChatGPT'. Accessed: Aug. 07, 2024. [Online]. Available: https://openai.com/index/chatgpt/

[19] T. Vaikunta Pai, P. S. Nethravathi, R. Birau, V. Popescu, B. H. Karthik Pai, and P. V. Naik, 'Multimodal ChatGPT: Extending ChatGPT to enable rich multimodal conversations using deep neural network', *Journal of Intelligent & Fuzzy Systems*, vol. Preprint, no. Preprint, pp. 1–17, 2024, doi: 10.3233/JIFS-239465.

[20] L. Gkinko and A. Elbanna, 'Chatbots at work: A taxonomy of the use of chatbots in the workplace', in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, and M. Mäntymäki, Eds., Cham: Springer International Publishing, 2021, pp. 29–39. doi: 10.1007/978-3-030-85447-8_3.

[21] Y. Gao, H. Rui, and S. Sun, 'The power of identity cues in text-based customer service: Evidence from Twitter', *MIS Quarterly*, vol. 47, no. 3, pp. 983–1014, 2023, doi: 10.25300/MISQ/2022/17366.

[22] L. R. Anderson and S. Conyers, 'Freeware versus commercial office productivity software', Acquisition Research Program, Technical Report, Dec. 2016. Accessed: Jan. 11, 2025. [Online]. Available: https://dair.nps.edu/handle/123456789/2191

[23] M. Keith, R. Santanam, and R. Sinha, 'Switching costs, satisfaction, loyalty, and willingness to pay for office productivity software', in *2010 43rd Hawaii International Conference on System Sciences*, Jan. 2010, pp. 1–9. doi: 10.1109/HICSS.2010.359.

[24] N. Gandal, S. Markovich, and M. H. Riordan, 'Ain't it "suite"? Bundling in the PC office software market', *Strategic Management Journal*, vol. 39, no. 8, pp. 2120–2151, Aug. 2018, doi: 10.1002/smj.2797.

[25] D. Tapscott, 'Office efficiency, effectiveness, and productivity', in *Office Automation: A User-Driven Method*, D. Tapscott, Ed., Boston, MA: Springer US, 1982, pp. 71–79. doi: 10.1007/978-1-4613-2489-8_5.

[26] K. Koenigsbauer, 'Microsoft 365 Copilot Chat', techcommunity.microsoft.com. Accessed: Mar. 15, 2025. [Online]. Available: https://techcommunity.microsoft.com/blog/microsoft365copilotblog/microsoft-365-copilot-chat-%e2%80%93-copilot-for-all-your-employees/4392002

[27] H. Snyder, 'Literature review as a research methodology: An overview and guidelines', *Journal of Business Research*, vol. 104, pp. 333–339, Nov. 2019, doi: 10.1016/j.jbusres.2019.07.039.

[28] D. Tranfield, D. Denyer, and P. Smart, 'Towards a methodology for developing evidence-informed management knowledge by means of systematic review', *British Journal of Management*, vol. 14, no. 3, pp. 207–222, 2003, doi: 10.1111/1467-8551.00375.

[29] R. F. Baumeister and M. R. Leary, 'Writing narrative literature reviews', *Review of General Psychology*, vol. 3, no. 1, pp. 311–320, 1997, doi: 10.1037/1089-2680.1.3.311.

[30] B. Kitchenham and S. Charters, 'Guidelines for performing systematic literature reviews in software engineering', vol. 2, Jan. 2007, doi: https://doi.org/10.53832/edtechhub.1003.

[31] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement', *International Journal of Surgery*, vol. 8, no. 5, pp. 336–341, Jan. 2010, doi: 10.1016/j.ijsu.2010.02.007.

[32] B. M. Shrestha, 'Systematic reviews and meta-analysis: Principles and practice', *JNMA J Nepal Med Assoc*, vol. 57, no. 215, pp. 1–2, Feb. 2019, doi: 10.31729/jnma.3986.

[33] F. Tingelhoff, M. Brugger, and J. M. Leimeister, 'A guide for structured literature reviews in business research: The state-of-the-

art and how to integrate generative artificial intelligence', *Journal of Information Technology*, vol. 40, no. 1, pp. 77–99, Mar. 2025, doi: 10.1177/02683962241304105.
[34] J. Wolfswinkel, E. Furtmueller, and C. Wilderom, 'Using grounded theory as a method for rigorously reviewing literature', *European Journal of Information Systems*, vol. 22, pp. 1–11, Nov. 2013, doi: 10.1057/ejis.2011.51.
[35] C. Hart, *Doing a Literature Review: Releasing the Research Imagination*, 2nd ed. Sage, 2018.
[36] O. Al-Tabbaa, S. Ankrah, and N. Zahoor, *Systematic literature review in management and business studies: A case study on university-industry collaboration*. 2019. doi: 10.4135/9781526467263.
[37] J. Webster and R. T. Watson, 'Analyzing the past to prepare for the future: Writing a literature review', *MIS Quarterly*, vol. 26, no. 2, pp. 8–13, 2002.
[38] B. González-Albo and M. Bordons, 'Articles vs. proceedings papers: Do they differ in research relevance and impact? A case study in the Library and Information Science field', *Journal of Informetrics*, vol. 5, no. 3, pp. 369–381, Jul. 2011, doi: 10.1016/j.joi.2011.01.011.
[39] S. Boell and B. Wang, 'www.litbaskets.io, an it artifact supporting exploratory literature searches for information systems research', presented at the ACIS 2019 Proceedings. 71, Dec. 2019. [Online]. Available: https://aisel.aisnet.org/acis2019/71
[40] Scopus, 'Scopus | The largest database of peer-reviewed literature | Elsevier'. Accessed: Aug. 05, 2024. [Online]. Available: https://elsevier.international/en-in/solutions/scopus.html
[41] R. D. Jong and D. Bus, 'VOSviewer: putting research into context', *Research Software Community Leiden*, Mar. 2023, doi: 10.21428/a1847950.acdc99d6.
[42] R. Kumar, *Research methodology : A step-by-step guide for beginners*, 5th ed., vol. 5. in 5, vol. 5. Los Angeles : Sage, 2014.
[43] D. Rooein, D. Bianchini, F. Leotta, M. Mecella, P. Paolini, and B. Pernici, 'aCHAT-WF: Generating conversational agents for teaching business process models', *Software and Systems Modeling*, vol. 21, no. 3, pp. 891–914, 2022, doi: 10.1007/s10270-021-00925-7.
[44] D. Rooein, D. Bianchini, F. Leotta, M. Mecella, P. Paolini, and B. Pernici, 'Chatting about processes in digital factories: A model-based approach', presented at the Lecture Notes in Business Information Processing, Nurcan S., Reinhartz-Berger I., Soffer P., and Zdravkovic J., Eds., Springer, 2020, pp. 70–84. doi: 10.1007/978-3-030-49418-6_5.
[45] F. De Luzi, M. Macrì, M. Mecella, and T. Mencattini, 'Cicero: An AI-Based Writing Assistant for Legal Users', presented at the Lecture Notes in Business Information Processing, Cabanillas C. and Pérez F., Eds., Springer Science and Business Media Deutschland GmbH, 2023, pp. 103–111. doi: 10.1007/978-3-031-34674-3_13.
[46] H. W. Marsh *et al.*, 'The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies.', *Journal of Educational Psychology*, vol. 111, no. 2, pp. 331–353, Feb. 2019, doi: 10.1037/edu0000281.
[47] J. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*, vol. 4th. SAGE Publications, 2014. Accessed: Aug. 20, 2024. [Online]. Available: http://archive.org/details/methodology-alobatnic-libraries-creswell
[48] A. Rindfleisch, A. J. Malter, S. Ganesan, and C. Moorman, 'Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines', *Journal of Marketing Research*, vol. 45, no. 3, pp. 261–279, Jun. 2008, doi: 10.1509/jmkr.45.3.261.
[49] M. E. Koltko-Rivera, 'The Psychology of Worldviews', *Review of General Psychology*, vol. 8, no. 1, pp. 3–58, 2004, doi: 10.1037/1089-2680.8.1.3.
[50] P. Johnson and J. Duberley, *Understanding management research: An introduction to epistemology*, 1st ed., vol. 1. Sage Publications Ltd, 2000. doi: 10.4135/9780857020185.
[51] C. A. Saliya, 'Research philosophy: paradigms, world views, perspectives, and theories', 2023, pp. 35–51. doi: 10.4018/978-1-6684-6859-3.ch004.
[52] A. VanScoy, H. Julien, A. Buckley, and J. Goodell, 'Theory usage in empirical research in ISIC conference papers (1996-2020)', Sep. 2022, doi: 10.18452/25229.
[53] O. Boubker, 'From chatting to self-educating: Can AI tools boost student learning outcomes?', *Expert Systems with Applications*, vol. 238, 2024, doi: 10.1016/j.eswa.2023.121820.
[54] F. Brachten, T. Kissmer, and S. Stieglitz, 'The acceptance of chatbots in an enterprise context – A survey study', *International Journal of Information Management*, vol. 60, 2021, doi: 10.1016/j.ijinfomgt.2021.102375.
[55] J. Zhang, X. Lu, W. Zheng, and X. Wang, 'It's better than nothing: The influence of service failures on user reusage intention in AI chatbot', *Electronic Commerce Research and Applications*, vol. 67, 2024, doi: 10.1016/j.elerap.2024.101421.
[56] D. Kudryavtsev, 'AI-driven digital business design assistant: A prototype demo', presented at the CEUR Workshop Proceedings, CEUR-WS, 2024.
[57] F. Rowe, 'What literature review is not: diversity, boundaries and recommendations', *Eur J Inf Syst*, vol. 23, no. 3, pp. 241–255, May 2014, doi: 10.1057/ejis.2014.7.
[58] S. Haag and A. Eckhardt, 'Organizational cloud service adoption: a scientometric and content-based literature analysis', *J Bus Econ*, vol. 84, no. 3, pp. 407–440, Apr. 2014, doi: 10.1007/s11573-014-0716-6.
[59] W. L. Neuman, *Social research methods: Qualitative and quantitative approaches*, 7th ed. Pearson Education, 2002.
[60] D. Dennehy, 'The use of open, axial and selective coding echniques: A literature analysis of IS research', in *Academy for Information Systems Conference Proceedings 2023*, 2023. [Online]. Available: https://aisel.aisnet.org/ukais2023/20
[61] T. I. Wilhelm, J. Roos, and R. Kaczmarczyk, 'Large language models for therapy recommendations across 3 clinical specialties: Comparative study', *Journal of Medical Internet Research*, vol. 25, 2023, doi: 10.2196/49324.
[62] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento, 'Bots with feelings: Should AI agents express positive emotion in customer service', *International Journal of Human Computer Studies*, vol. 150, pp. 1–24, 2021, doi: 10.1016/j.ijhcs.2021.102607.
[63] N. Belhaj, A. Hamdane, N. El Houda Chaoui, H. Chaoui, and M. El Bekkali, 'Engaging students to fill surveys using chatbots: University case study', *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 473–483, 2021, doi: 10.11591/ijeecs.v24.i1.pp473-483.
[64] T. Lewandowski, E. Kučević, S. Leible, M. Poser, and T. Böhmann, 'Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement', *Electronic Markets*, vol. 33, no. 1, pp. 1–20, 2023, doi: 10.1007/s12525-023-00662-3.
[65] F. Almeida, 'Implementation of a chatbot in a unified communication channel', *Journal of Systems and Information Technology*, vol. 27, no. 1, pp. 94–115, 2025, doi: 10.1108/JSIT-08-2023-0160.
[66] S. Jackson and N. Panteli, 'AI-based digital assistants in the workplace: An idiomatic analysis', *Communications of the Association for Information Systems*, vol. 55, pp. 627–653, 2024.
[67] E. Brynjolfsson and A. McAfee, 'Winning the race with ever-smarter machines', *MIT SMR*, no. 53, pp. 53–60, Dec. 2011.
[68] B. Zhong, W. He, Z. Huang, P. E. D. Love, J. Tang, and H. Luo, 'A building regulation question answering system: A deep learning methodology', *Advanced Engineering Informatics*, vol. 46, pp. 1–11, 2020, doi: 10.1016/j.aei.2020.101195.
[69] M. Nadeem, L. Javed, S. S. Sohail, E. Cambria, and A. Hussain, 'Are foundation models the next-generation social media content moderators?', *IEEE Intelligent Systems*, vol. 39, no. 6, pp. 70–80, 2024, doi: 10.1109/MIS.2024.3477109.
[70] J. Choi and D. L. Nazareth, 'Repairing trust in an e-commerce and security context: An agent-based modeling approach', *Information Management and Computer Security*, vol. 22, no. 5, pp. 490–512, 2014, doi: 10.1108/IMCS-09-2013-0069.
[71] G. Graham, T. M. Nisar, G. Prabhakar, R. Meriton, and S. Malik, 'Chatbots in customer service within banking and finance: Do chatbots herald the start of an AI revolution in the corporate world?', *Computers in Human Behavior*, vol. 165, pp. 1–15, 2025, doi: 10.1016/j.chb.2025.108570.
[72] J. J. Bird and A. Lotfi, 'Customer service chatbot enhancement with attention-based transfer learning', *Knowledge-Based Systems*, vol. 301, pp. 1–12, 2024, doi: 10.1016/j.knosys.2024.112293.

[73] Y. Zhang, C. Liang, and X. Li, 'Understanding virtual agents' service quality in the context of customer service: A fit-viability perspective', Electronic Commerce Research and Applications, vol. 65, pp. 1–12, 2024, doi: 10.1016/j.elerap.2024.101380.

[74] N. Köster, C. Mundt, and H. Lödding, 'Planning and control of maritime commissioning: planning concept', presented at the IFIP Advances in Information and Communication Technology, Alfnes E., Romsdal A., Strandhagen J.O., von Cieminski G., and Romero D., Eds., 2023, pp. 735–749. doi: 10.1007/978-3-031-43670-3_51.

[75] Q. Chen, Y. Lu, Y. Gong, and J. Xiong, 'Can AI chatbots help retain customers? Impact of AI service quality on customer loyalty', Internet Research, vol. 33, no. 6, pp. 2205–2243, 2023, doi: 10.1108/INTR-09-2021-0686.

[76] S. W. Chae, Y. W. Seo, and K. C. Lee, 'Task difficulty and team diversity on team creativity: Multi-agent simulation approach', Computers in Human Behavior, vol. 42, pp. 83–92, 2015, doi: 10.1016/j.chb.2014.03.032.

[77] P. Mikalef and M. Gupta, 'Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance', Inf. Manage., vol. 58, no. 3, pp. 1–20, Apr. 2021, doi: 10.1016/j.im.2021.103434.

[78] R. Abdelghani, P.-Y. Oudeyer, E. Law, C. de Vulpillières, and H. Sauzéon, 'Conversational agents for fostering curiosity-driven learning in children', International Journal of Human Computer Studies, vol. 167, pp. 1–21, 2022, doi: 10.1016/j.ijhcs.2022.102887.

[79] X. Lin, X. Wang, B. Shao, and J. Taylor, 'How chatbots augment human intelligence in customer services: A mixed-methods study', Journal of Management Information Systems, vol. 41, no. 4, pp. 1016–1041, 2024, doi: 10.1080/07421222.2024.2415773.

[80] S.-L. Cheng et al., 'Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: Cross-sectional study', Journal of Medical Internet Research, vol. 25, no. 1, pp. 1–9, 2023, doi: 10.2196/51229.

[81] L. Willcocks, M. Lacity, and A. Craig, 'Robotic process automation: Strategic transformation lever for global business services?', Journal of Information Technology Teaching Cases, vol. 7, no. 1, pp. 17–28, May 2017, doi: 10.1057/s41266-016-0016-9.

[82] R. Figliè, T. Turchi, G. Baldi, and D. Mazzei, 'Towards an LLM-based intelligent assistant for industry 5.0', presented at the CEUR Workshop Proceedings, Dix A., Roach M., Turchi T., Malizia A., and Wilson B., Eds., CEUR-WS, 2024.

[83] Z. Eberhart, A. Bansal, and C. McMillan, 'A Wizard of Oz Study Simulating API Usage Dialogues With a Virtual Assistant', IEEE Transactions on Software Engineering, vol. 48, no. 6, pp. 1883–1904, 2022, doi: 10.1109/TSE.2020.3040935.

[84] N. Ben-Shabat et al., 'Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study', International Journal of Medical Informatics, vol. 168, 2022, doi: 10.1016/j.ijmedinf.2022.104897.

[85] A. Siddig and A. Hines, 'A psychologist chatbot developing experience', presented at the CEUR Workshop Proceedings, Curry E., Keane M., Ojo A., and Salwala D., Eds., CEUR-WS, 2019, pp. 200–211.

[86] C. Wang et al., 'Application of large language models in medical training evaluation—Using ChatGPT as a standardized patient: multimetric assessment', Journal of Medical Internet Research, vol. 27, pp. 1–15, 2025, doi: 10.2196/59435.

[87] S. Gregor, 'The nature of theory in information systems', MIS Quarterly, vol. 30, no. 3, pp. 611–642, 2006, doi: 10.2307/25148742.

[88] K. Larsen, N. Saini, and R. Mueller, 'Theories used in IS research wiki', Information Systems Theories. Accessed: Mar. 31, 2025. [Online]. Available: https://is.theorizeit.org/wiki/Main_Page

[89] J. Reis, M. Amorim, N. Melão, Y. Cohen, and M. Rodrigues, 'Digitalization: A literature review and research agenda', in Proceedings on 25th International Joint Conference on Industrial Engineering and Operations Management - IJCIEOM 2019, Z. Anisic, B. Lalic, and D. Gracanin, Eds., Cham: Springer International Publishing, 2020, pp. 443–456. doi: https://doi.org/10.1007/978-3-030-43616-2_47.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 577–586
DOI: 10.15439/2025F7738
ISSN 2300-5963 ACSIS, Vol. 43

Vehicle Routing under Complex Access-to-Energy Constraints

Alain Quilliot and Hélène Toussaint
LIMOS Lab.
UCA, CNRS and EMSE
Clermont-Ferrand, France
Email: alain.quilliot@uca.fr

***Abstract*—Photovoltaic platforms enable a single agent to simultaneously act as both a producer and a consumer of power, facilitating *self-consumption* strategies. This trend aligns with the goal of reducing CO2 emissions and is poised to significantly transform the structure of energy markets. It also introduces specific challenges—both tactical (e.g., pricing) and operational (e.g., routing, scheduling)—related to synchronizing energy production with consumption. In this work, we address the problem of efficiently routing a fleet of electric autonomous vehicles (EAVs), using energy that is either produced by a photovoltaic platform or purchased from the general power grid. We propose an exact Mixed-Integer Linear Programming (MILP) formulation of the problem, along with a heuristic approach that approximates the power production component of the model using surrogate representations.**

I. Introduction

RENEWABLE energy sources (e.g., photovoltaic, wind, hydrogen) are driving the emergence of local, in situ producers who simultaneously consume energy—such as factories and farms. In this context, the energy production-consumption process becomes partially endogenous, forming a closed-loop system under the *self-consumption* paradigm [15]. This paradigm is expected to have a significant impact on energy economics [11], [14]. It raises various challenges, ranging from operational issues—related to the scheduling and synchronization of production and consumption—to tactical and strategic decisions, such as pricing, storage, and interaction with the central grid. In scenarios where the decentralized producer-consumer entity is a consortium of independent agents, each with its own schedule and shared access to the production platform, fairly distributing costs and benefits raises the *cooperative* issue.

We consider here the problem of routing a fleet of electric autonomous vehicles (EAVs), powered either by photovoltaic-generated electricity—available at time-dependent rates—or by electricity purchased from the general grid at time-dependent prices. This *Vehicle Routing under Energy Production Costs* (**VR_EPC**) problem requires fleet managers to synchronize vehicle activity with energy production and procurement, while taking into account the limited storage capacities of both the platform and the vehicles.

A. State of the Art

Numerous studies have explored the routing of electric vehicles with the aim of minimizing energy costs or addressing environmental concerns, such as in the *Green Vehicle Routing Problem* (Green VRP), the *Pollution Routing Problem*, and the *Hybrid Vehicle Routing Problem*: [5], [16], [21],[19]). They most often proposed models involving refueling transactions subject to time windows or shared access constraints. For example:

- Erdoğan and al. [6] introduced the *Green VRP*, which minimizes both total travel distance and the number of refueling transactions.
- Franceschetti and al. [8] proposed a *Pollution-Routing* model, incorporating time-dependent costs and an objective function accounting for driver wages and fuel costs.
- Koç and al. [12] focused on determining where and how much vehicles should be recharged, considering access-to-energy constraints.
- In [21], authors addressed customer sequencing, while explicitly accounting for the recharge capacity of the charging stations.

Some authors also dealt with CO2 emitting vehicles, with the purpose of controlling related CO2 emission:

- Raylan and al. [17] minimized emissions resulting from both routing and load-related factors.
- Sachenbacher and al. [18] accounted for time-dependent factors and refueling schemes, adapting shortest path algorithms accordingly.
- Kuo [13] explored speed control to balance energy consumption, distance, and time.
- Schneider and al. [19] developed heuristics for a Vehicle Routing problem with time windows and safety-specific refueling constraints.
- Lajunen [14] performed simulation-based comparisons of energy savings across urban shuttle and bus configurations.

Despite these contributions, few studies have jointly managed both energy production and consumption. Addressing this challenge requires integrating heterogeneous routing and scheduling processes while accounting for storage constraints (see [2], [3], [1]). Some formulations borrow from the *Lot*

Sizing framework or its multi-level scheduling variants (see [4], [10], [20]), but these models involve highly heterogeneous variables and constraints, often leading to poor performance of linear relaxations. Moreover, they implicitly rely on the existence of a single decider provided with full control and full information with respect to the system, inducing an assumption which does not fit most real life contexts. Alternatives such as constraint programming and bi-level formulations may help coordinate interactions across decision levels. Nonetheless, whatever be the point of view (centralized or collaborative), designing efficient solution algorithms remains a major challenge. A recent trend—facilitated by machine learning—is to approximate complex sub-models using surrogate constraints or cost functions.

B. Main Achievements

Since our **VR_EPC** problem is likely to involve in practice collaborative features making it difficult to suppose the existence of a global decider, we adopt here the point of view of the EAV. Also, we skip the uncertainty issue, while supposing that the behavior of our system, and more specifically of the photovoltaic platform, is deterministic. Then we propose two solution approaches for the **VR_EPC** problem.

- The first one is an exact MILP formulation that considers the *vehicle* variables as master variables and includes specific *Recharge Decomposition* constraints requiring the application of a specific separation procedure, which works in polynomial time. This MILP model is solved using a branch-and-cut method. However, this approach becomes computationally inefficient for large instances.
- So we also design 2 heuristic algorithms, which makes the EAV decide under a partial knowledge of the behavior of the photovoltaic platform, while using an approximation of the energy production model. The resulting algorithms pave the way for more efficient handling of uncertainty and collaboration challenges, particularly in scenarios where the photovoltaic (PV) platform and the vehicles are run by distinct players, who do not fully share neither the same goals nor the access to information.

The paper is organized as follows. Section II provides a detailed description of the *Vehicle Routing under Energy Production Costs* problem (**VR_EPC**). In Section III, we present the MILP formulation and its resolution using a branch-and-cut approach. Section IV offers a statistical analysis that provides the basis for the development and testing of 2 local search heuristic. We conclude briefly in Section V.

II. The Vehicle Routing under Energy Production Costs Problem

We consider a photovoltaic (PV) micro-plant, referred to as **PVP**, along with a fleet of electric autonomous vehicles (EAV) in charge of visiting a set of stations. These two components interact through recharge transactions when the vehicles return to the micro-plant to recharge (see Fig. 1). For the sake of simplicity, we limit our study to a single vehicle that must complete a *Traveling Salesman*-type tour, and we assume the system operates in a deterministic manner. Thus, the main components of our target system are:

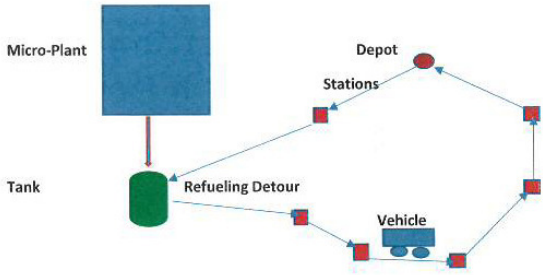


Fig. 1. The PV-Micro-Plant Interacting with a Vehicle Route

- **A Photo-Voltaic Micro-Plant PVP**
  The time horizon of **PVP** is divided into N periods $i = 0, \ldots, N-1$, all with a same duration p. Thus the starting timestamp of period i is equal to $p \cdot i$ and its ending timestamp equal to $p \cdot (i+1)$. During a period i, **PVP** is expected to produce R_i energy units. It may also buy an additional amount y_i of power, that cannot exceed a *charge* capacity denoted by C^{Ch}. The cost of buying such an additional energy amount y_i depends on both i and y_i and may be written $\Phi_i(y)$, where Φ_i is a piecewise linear increasing convex function. The convexity of this cost function Φ_i expresses the fact that marginal power purchase prices are usually increasing. **PVP** is provided with a macro-battery, with storage capacity $C^{PVP} \geq C^{Ch}$ and initial load H_0^{PVP}. It must manage its purchase operations in a way which makes it able to meet vehicle's demand without exceeding storage and charge capacities, while ending the process with a load at least equal to H_0^{PVP} and minimizing related purchase costs.
- **An Electric Autonomous Vehicle EVeh and a Set of Stations J**
  This vehicle, denoted by **EVeh**, is initially located at a specific station *Depot*. It must visit and service, within the time horizon $[0, N.p]$, a set $\mathbf{J} = \{1, \ldots, M\}$ of stations according to a TSP (*Traveling Salesman*) route Γ, before coming back to *Depot*. Moving from a station j to a station k requires $E_{j,k}$ energy units and $T_{j,k}$ time units, service times being included into the $T_{j,k}$ values. We suppose that vectors T, θ, θ^* $(E, \epsilon, \epsilon^*)$ define a distance on the set defined by $Depot$, **PVP** and **J**. **EVeh** is provided with a battery, with storage capacity C^{Veh} and initial load H_0^{Veh}. It must end with a load at least equal to H_0^{Veh} while minimizing the time when it returns to $Depot$. $Depot$ is considered as a station, identified as $Depot = 0$ at the beginning of the process and $Depot = M + 1$ at the end. By the same way, **PVP** is also considered as a station, identified as **PVP** $= -1$

- **Recharge Transactions**
  Because of its storage capacity C^{Veh} and the constraint about its load at the end of the process, **EVeh** must periodically move to **PVP** in order to recharge. Such a *recharge transaction* can be achieved in a single period i and takes place between two stations j, k consecutive according to the route Γ: **EVeh** moves from j to **PVP** in order to arrive before time $p \cdot i$, receives some amount $m \leq Inf(C^{Ch}, C^{Veh})$ of energy, and starts again from **PVP** at time $p\cdot(i+1)$ in order to reach k. Since the vehicle is a kind of robot, this recharge transaction, denoted by $\omega = (i, j, k, m)$, involves some human resource and so it induces a cost Ψ_i. This human resource cost depends on i and is independent on m. For the sake of safety, purchasing power is forbidden during the period i when the recharge transaction ω takes place. Moving from j to **PVP** requires ϵ_j energy units and θ_j additional time, while moving from **PVP** to k requires ϵ_k^* energy units and θ_k^* additional time. A recharge transaction may impose **EVeh** to wait at **PVP** until the beginning of period i in case it arrives before time $p\cdot i$. If we denote by τ_j the time when **EVeh** arrives at j, by V_j^{Veh} its energy load and by V_i^{PVP} the energy load of the PV-plant at the beginning of period i, then we must have:
  - $p \cdot i \geq \tau_j + \theta_j$; $\tau_k = p \cdot (i+1) + \theta_k^*$;
  - $V_{i+1}^{MP} = V_i^P - m \geq 0$;
  - The load of **EVeh** at the end of period i must be equal to $V_j^{Veh} - \epsilon_j + m \leq C^{Veh}$;
  - The load of **EVeh** at the beginning of period i must be equal to $V_j^{Veh} - \epsilon_j \geq 0$.

Since the system described this way involves two players, each provided with its own performance criterion, resulting problem might be set according to the bi-objective format. Yet, our focus here is on the algorithms. Therefore, we introduce a *time versus money* coefficient α, and formulate **VR_EPC** as a mono-objective problem. By the same way, since we tend here to adopt the point of view of the vehicle manager **EVeh**, we set related model as a bi-level one.

VR_EPC: ***Vehicle Routing with Energy Production Costs*** : {Compute the route Γ followed by **EVeh** together with the recharge transactions that link **PVP** and **EVeh** in such a way that:

- All stations are visited once within the time horizon $[0, p \cdot N]$.
- Vehicle storage capacity constraints and energy requirements are satisfied.
- Some extended cost $PCost+\alpha.Veh_Time$ is minimized, where $PCost$ means the optimal cost value induced by the **PVP** sub-problem consisting in deciding the purchase vector $y = (y_i, i = 0, \ldots, N-1)$ in such a way that:
  - It meets **PVP** storage and charge capacities;
  - It meets the needs related to the recharge transactions while allowing **PVP** to end with at least as much energy as when it started;
  - $PCost = \sum_i (\Psi_i \cdot \delta_i + \Phi_i(y_i)).\}$

TABLE I
PRODUCTION RATES, UNIT PURCHASE PRICES AND RECHARGE COSTS

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13
R_i	2	2	0	0	4	4	4	0	0	0	0	2	2	0
P_i	2	3	5	5	1	1	1	5	5	5	5	2	2	5
Ψ_i	1	1	1	1	2	2	1	1	1	1	1	1	2	2

An Example: Let us suppose that:

- M = 5, $N = 14$, $p = 2$; $\alpha = 1$;
- $C^{Veh} = 16, H^{Veh} = 9, C^{PVP} = 20, H^{PVP} = 5, C^{Ch} = 15$;
- Expected productions $R_i, i = 0, \ldots, N-1$ come as in Table I;
- For any period i, function Φ_i comes as: $\Phi_i(y) = P_i \cdot y$, y denoting the additional energy bought at period i and prices P_i come according to table I.

Let us consider a route $\Gamma = \{0, 1, 2, 3, 4, 5, 6\}$ together with time and energy requirements given according to Figure 2:

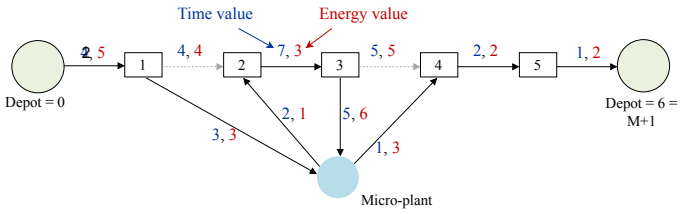


Fig. 2. The Route Γ, together with the time and energy requirements

Then we get a feasible **VR_EPC** solution involving Γ by specifying the recharge transactions as follows:

- A first recharge transaction occurs at period 3, while **EVeh** is moving from station 1 to station 2. Related energy amount m is equal to 10.
- Another recharge transaction occurs at period 11, while **EVeh** is moving from station 3 to station 4. Related energy amount m is equal to 15.
- **PVP** buys 3 energy units at period 0 and 3 energy units at period 1. Those operations induce a purchase cost equal to $3 \cdot 3 + 2 \cdot 3 = 15$.
- The 2 recharge transactions induce a cost equal to 2, which makes the $PCost$ value equal to $15+2 = 17$, while the value Veh_Time is equal to $14 \cdot 2 = 28$, inducing a generalized cost value $\alpha \cdot PCost + \beta \cdot Veh_Time$ equal to $1 \cdot 17 + 1 \cdot 28 = 45$.

III. A MILP MODEL

Though **VR_EPC** is too complex for an exact handling of large size instances, casting it into the MILP format is important, not only because resulting **MILP_VR_EPC** model will provide us with benchmark results in the case of small instances, but also because it will contain a structural analysis of the problem.

A. The Variables of the ***MILP_VR_EPC*** model

Getting a MILP formulation first requires identifying the variables. Since we adopt here the point of view of the vehicle manager, our master variables are going to be 2 $\{0,1\}$-valued vectors $Z = (Z_{j,k}, k \neq j \in \{-1, 0, \ldots, M+1\})$ and $X = (X_{j,k}, k \neq j \in \{0, \ldots, M+1\})$, that respectively describe the routes followed by the vehicle with and without the detours induced by the recharging transactions. Those main *vehicle* variables are going to be completed by auxiliary variables related to the time values when the vehicle visits the stations and to its energy load at this time. We also need *PV-Plant* variables that allow us to describe the evolution along the periods of the power load stored inside the **PVP** macro-battery. Finally, we need *Recharge Transaction* variables, linking the respective trajectories of the vehicle **EVeh** and the PV-Plant **PVP**. This leads us to introduce the following variables:

- **Vehicle** variables:
  - $\{0,1\}$-valued variables $Z_{j,k}, k \neq j \in \{-1, 0, \ldots, M+1\}$: $Z_{j,k} = 1$ iff the vehicle moves from j to k: -1 denotes here the micro-plant **PVP**.
  - $\{0,1\}$-valued variables $X_{j,k}, k \neq j \in \{0, \ldots, M+1\}$: $X_{j,k} = 1$ iff the vehicle moves either from j to k or from j to **PVP**, and next from **PVP** to k.
  - Non negative variables $L_j^{Veh}, j = 0, \ldots, M$: L_j^{MP} means the energy transferred to **EVeh** just after j, in case $Z_{j,-1} = 1$;
  - Non negative variables$V_j^{Veh}, j = 0, \ldots, M+1$: V_j^{Veh} means the energy stored by **EVeh** when it arrives at j;
  - Non negative variables $\tau_j, j = 0, \ldots, M+1$: τ_j means the time when **EVeh** arrives at j;
  - Non negative variables $\tau_j^*, j = 0, \ldots, M+1$: If $Z_{j,-1} = 1$ then τ_j^* means the time when **EVeh** starts recharging after station j in case $Z_{j,-1} = 1$.
- **PV-Plant** variables:
  - Non negative variables $y_i, i = 0, \ldots, N-1$: y_i means the energy amount bought by the PV-Plant's during period i;
  - $\{0,1\}$-valued variables $\delta_i, i = 0, \ldots, N-1$: $\delta_i = 1$ means that some recharge transaction takes place at i;
  - Non negative variables $V_i^{PVP}, i = 0, \ldots, N-1, N$: V_i^{PVP} means the energy stored by **PVP** at the beginning of i;
  - Non negative variables $L_i^{PVP}, i = 0, \ldots, N-1$: L_i^{PVP} means the energy transferred from **PVP** period i in case in case $\delta_i = 1$.
- **Recharge Transaction** variables:
  - $\{0,1\}$-valued variables $U_{i,j}, i = 0, \ldots, N-1, j = 0, \ldots, M$: $U_{i,j} = 1$ means that some recharge transactions occurs at period i that involves station j and its sucessor in Γ;
  - Non negative variables $m_{i,j}, i = 0, \ldots, N-1, j = 0, \ldots, M$: $m_{i,j}$ means related amount of energy.

B. Structural Recharge Decomposition Vehicle Constraints

Z and X describe the full route followed by **EVeh**. They must clearly meet the following standard *vehicle routing* constraints:

- $Z_{M+1,0} = 1; \quad \forall j : Z_{j,j} = 0;$ (VR1)
- $\forall j = 0, \ldots, M+1$: $\sum_{k=-1,\ldots,M+1} Z_{j,k} = 1 = \sum_{k=-1,\ldots,M+1} Z_{k,j}$; (VR2)
- $\sum_{j=0,\ldots,M+1} Z_{j,-1} = \sum_{j=0,\ldots,M+1} Z_{-1,j} \geq 1;$ (VR3)
- $\forall j, k \in \{0, \ldots, M+1\}: \quad X_{j,k} \geq Z_{j,k}$; (VR4)
- $\forall j \in \{0, \ldots, M+1\}$: $\sum_{k=0,1,\ldots,M+1} X_{j,k} = 1 = \sum_{k=0,1,\ldots,M+1} X_{k,j}$; (VR5)

Yet we see that (VR1, ..., VR5) are not enough to ensure that we may interpret Z, X as a route followed by **EVeh**, feasible with respect to the energy requirements. In fact, they are not enough to prevent us from the existence of standard *Traveling Salesman sub-tours*. So we must reinforce them. We do it by noticing that if **EVeh** spends W energy while moving inside or at the boarder of some station subset A which does contain the micro-plant, then it must move at least $\lceil \frac{W}{C^{Veh}} \rceil$ times towards **PVP** in order to refuel. In order to formalize this, we must introduce some additional notations:

- For any such a subset A of $\{-1, 0, \ldots, M, M+1\}$:
  - $Cl(A) = \{(j,k)$ s.t at least j or k is in $A\}$;
  - $\delta(A) = \{(j,k)$, s.t $j \notin A$ and $k \in A\}$.
- For any (j,k):
  - $\Pi_{j,k} = E_0$ if $(j,k) = (M+1, 0)$ and $\Pi_{j,k} = C^{Veh}$ else.
  - $\Pi^*_{j,k} = C^{Veh} - E_0$ if $(j,k) = (M+1, 0)$ and $\Pi^*_{j,k} = C^{Veh}$ else.

Then we become able to derive the following *Recharge Decomposition* constraints that reinforce (VR1, ..., VR5):

- For any $A \subseteq \{0, \ldots, M+1\}$, $\sum_{(j,k)\in\delta(A)} \Pi_{j,k} \cdot Z_{j,k} \geq \sum_{(j,k)\in Cl(A)} E_{j,k} \cdot Z_{j,k}$ (VR6)
- For any $A \subseteq \{0, \ldots, M+1\}$, $\sum_{(j,k)\in\delta(J-A)} \Pi^*_{j,k} \cdot Z_{j,k} \geq \sum_{(j,k)\in Cl(A)} E_{j,k} \cdot Z_{j,k}$ (VR6-Bis)

The following Lemma 1 provides us with the structural meaning of these constraints: It tells us that adding them to (VR1, ..., VR5) ensures us that (Z, X) are going to describe a vehicle route that excludes *sub-tours* and can be decomposed into a sequence of sub-routes ($Depot \rightarrow$ **PVP**), (**PVP** $\rightarrow$ **PVP**) and (**PVP** $\rightarrow Depot$), all feasible with respect to the energy requirements.

Lemma 1: *The VRP constraints (VR1, ..., VR6-Bis) hold if and only if the arcs (j,k) such that $Z_{j,k} = 1$ define a collection γ of sub-tours $\gamma_s, s = 0, \ldots, S$ such that:*

- *γ_0 starts from $Depot = 0$, ends into **PVP** $= -1$, and spends less than H_0^{Veh} energy.* (SUB1)
- *γ_S starts from **PVP** $= -1$, ends into $Depot = 0$, and spends less than $C^{Veh} - H_0^{Veh}$ energy.* (SUB2)
- *For any $s = 1, \ldots, S-1$, γ_s starts from **PVP** $= -1$, ends into -1 and does not require more than C^{Veh} energy.* (SUB3)
- *Every station $j = 1, \ldots, M$ is visited once.* (SUB4)

Proof. We only need to simultaneously follow the TSP route defined by X and its extension defined by Z. **End-Proof**

Remark 1: Constraints (VR6, VR6-Bis) neither tells in which order the sub-routes $\gamma_s, s = 0, \ldots, S$ must be performed, nor it ensures the feasibility of resulting route with respect to the time horizon $p \cdot N$.

Separating the Recharge Decomposition Constraints: As usually when it comes to the management of constraints on the subsets of a given set, we must cope with the separability issue. Given 2 possibly non integral vectors (Z, X), separating the constraints (VR6, VR6-Bis) means checking that all those constraints are satisfied by (Z, X) and, in case they are not, computing a subset $A \subseteq \{0, \ldots, M+1\}$ such that related VR6 or VR6-Bis constraint is violated by (Z, X). Theorem 1 below not only tells us that separating (VR6, VR6-Bis) can be done in polynomial time, but related proof provides us with a polynomial time algorithm that will perform this separation process.

Theorem 1: *The Recharge Decomposition constraints can be separated in polynomial time, by application of a min cost flow algorithm.*

Sketch of the Proof: Let us restrict ourselves to the case of the constraints VR6 (the case of the constraints VR6-Bis is quite similar). Given Z, possibly non integral. Separating VR6 means searching for $A \subseteq \{0, \ldots, M+1\}$, which does not contain -1 and is such that:

$$\sum_{(j,k) \in \delta(A)} \Pi_{j,k} \cdot Z_{j,k} < \sum_{(j,k) \in Cl(A)} E_{j,k} \cdot Z_{j,k},$$

or equivalently for $B = \{-1, 0, \ldots, M+1\} \setminus A$, which contains -1 and is such that: (*)

$$\sum_{(j,k)\ s.t. j \in B, k \notin B} \Pi_{j,k} \cdot Z_{j,k} + \sum_{(j,k)\ s.t. j \in B, k \in B} E_{j,k} \cdot Z_{j,k} < \sum_{(j,k)} E_{j,k} \cdot Z_{j,k}.$$

Let us set $\Delta = \sum_{(j,k)} E_{j,k} \cdot Z_{j,k}$, and let us construct an auxiliary multi-graph $G = (X, A)$ as follows:

- $X = \{Source = -1, 0, \ldots, M+1, M+2 = Sink\}$;
- A is defined as the set of all *simple-arc* $(j,k), j \neq k \in \{-1, 0, \ldots, M+1\}$, augmented, for every j, with *copy-arcs* $(j, M+1)^k, k \neq j \in \{-1, 0, \ldots, M+1\}$, which connect j to $Sink = (M+2)$ and are provided with label k. Every *copy-arc* $a = (j, M+1)^k$ is provided with a weight w_a equal to $E_{j,k} \cdot Z_{j,k}$. Every *simple-arc* $a = (j,k)$ is provided with a weight w_a equal to $\Pi_{j,k} \cdot Z_{j,k} - E_{j,k} \cdot Z_{j,k}$.

Then we easily check that computing B such that (*) holds means computing a cut B' which separates $Source = -1$ from $Sink = M+2$ in G and is such that $\sum_{a\ s.t.\ origin(a) \in B', destination(a) \notin B'} w_a < \Delta$. We know that this can be done in polynomial time through a simple Max-Flow algorithm. **End-Proof**.

C. The *VR_EPC_MILP* MILP formulation

We are now able to extend the structural *vehicle routing* constraints VR1, ..., VR6-Bis into a **VR_EPC_MILP** setting of **VR_EPC**. We do it while distinguishing 3 main groups of constraints:

- The **PV-Plant Constraints**: They involve the variables related to the purchase of power and express the evolution along the periods of the load V_i^{PVP} of the **PVP** battery, the fact that this load cannot exceed the capacity C^{PVP}, and the fact that the final load V_N^{PVP} must be no smaller than the initial load H_0^{PVP}.
- The **Vehicle Routing Constraints**: They contain VR1, ..., VR6-Bis, together with constraints related to the time values when the vehicle arrives to the stations or to **PVP** for the recharging transactions. Those temporal constraints fix the order according to which the sub-routes involved in Lemma 1 are visited. The *Vehicle Routing* constraints also contain constraints which express the evolution along the route defined by (Z, X) of the load V_j^{Veh} of the **EVeh** battery, the fact that this load cannot exceed the capacity C^{Veh}, and the fact that the final load V_{M+1}^{Veh} must be no smaller than the initial load H_0^{Veh}.
- The **Synchronization Constraints**: They link together the **PVP** periods and the time horizon $[0, p \cdot N]$ of the vehicle, and synchronize the energy received by the vehicle and the energy delivered by the PV-Plant during the recharge transactions.

Those constraints may be formalized as follows (for the sake of simplicity, we replace the "Big M" formulations by implications):

VR_EPC Constraints and Objective Function:

- **Objective**: Minimize $\sum_i (\Psi_i.\delta_i + \Phi_i(y_i) + \alpha.\tau_{M+1}$
- **PV-Plant Constraints**
  - $\forall i = 1, \ldots, N-1$: $y_i \leq C^{Ch} \cdot (1 - \delta_i)$; (PC1)
  - $\forall i = 0, \ldots, N$: $V_i^{PVP} \leq C^{PVP}$; (PC2)
  - $V_0^{PVP} = H_0^{PVP}$; $V_N^{PVP} \geq H_0^{PVP}$; (PC3)
  - $\forall i = 1, \ldots, N$: $V_i^{PVP} = V_{i-1}^{PVP} + y_i - L_i^{PVP}$; (PC4)
- **Vehicle Routing Constraints**
  - (VR1, ..., VR6-Bis) involved in Lemma 1;
  - $V_0^{Veh} = H_0^{Veh}$; $V_{M+1}^{Veh} \geq H_0^{Veh}$; (VR7)
  - $\forall j = 0, \ldots, M+1$: $E_{j,-1} \leq V_j^{Veh} \leq C^{Veh}$; (VR8)
  - $\forall j, k = 0, \ldots, M+1$: $X_{j,k} \rightarrow (V_k^{Veh} + E_{j,k} + (Z_{j,k} - 1) \cdot (\epsilon_j + \epsilon_k - E_{j,k})) \leq (V_j^{Veh} + L_j^{Veh})$; (VR9)
  - $\forall j = 0, \ldots, M$: $Z_{j,-1} \rightarrow (V_j^{Veh} \geq \epsilon_j)$; (VR10)
  - $\forall j = 0, \ldots, M$: $Z_{j,-1} \rightarrow (V_j^{Veh} + L_j^{Veh} \leq \epsilon_j + C^{Veh})$; (VR10-Bis)
  - $\tau_0 = 0$; $\tau_{M+1} \leq p \cdot N$; (VR11)
  - $\forall j, k = 0, \ldots, M+1$: $Z_{j,k} \rightarrow (\tau_j + T_{j,k} \leq \tau_k)$; (VR12)
  - $\forall j, k = 0, \ldots, M+1$: $(X_{j,k} - Z_{j,k} = 1) \rightarrow (\tau_j^* + p + T_{-1,k} \leq \tau_k)$; (VR13)
  - $\forall j = 0, \ldots, M+1$: $(Z_{j,-1}) \rightarrow (\tau_j + T_{j,-1} \leq -\tau_j^*)$; (VR13-Bis)
- **Synchronization Constraints**
  - $\forall j = 0, \ldots, M$: $\sum_{i=0,\ldots,N-1} U_{i,j} = Z_{j,-1}$; (SY1)
  - $\forall j = 0, \ldots, M$: $\sum_{i=0,\ldots,N-1} m_{i,j} = L_j^{Veh}$; (SY1-Bis)
  - $\forall i = 0, \ldots, N-1$: $\sum_{j=0,\ldots,M} U_{i,j} = \delta_i$; (SY2)

- $\forall i = 0, \ldots, N-1$: $\sum_{j=0,\ldots,M} m_{i,j} = L_i^{PVP}$; (SY2-Bis)
- $\forall j = 0, \ldots, M$: $\sum_{i=0,\ldots,N-1} p \cdot i \cdot U_{i,j} = \tau_j^*$; (SY3)
- $\forall i = 0, \ldots, N-1, j = 0, \ldots, M$: $m_{i,j} \leq Inf(C^{PVP}, C^{Veh}) \cdot U_{i,j}$; (SY4)

Theorem 2: *Above **VR_EPC_MILP** MILP model solves **VR_EPC** in an exact way.*

Proof: Lemma 1 tells us that X, Z define a feasible vehicle route Γ. Constraints (SY1, SY2) link the variables U, m with the route Z and the energy received by the vehicle, with the PV-Plant variables δ and the energy transferred by **PVP**. (SY4) makes those energy amounts be consistent with the storage capacities. (SY3) links the time values of **PVP** and **EVeh**. (PC4) describes the evolution throughout the time of the energy stored by **PVP**. Vehicle constraints (VR7, ..., VR10) describe the evolution throughout the time of the energy stored by **EVeh**. They keep this energy from becoming negative or larger than the capacity C^{Veh}. (VR11, ..., VR13-Bis) describe the evolution of the time while the vehicle follows the route described by X and Z. **End-Proof**

Practical Handling of the VR_EPC_MILP MILP model. We handle **VR_EPC_MILP** through Branch and Cut, while applying the auxiliary graph and the Max-Flow algorithm involved in Theorem 1 in order to separate the *Recharge Decomposition* constraints.

D. Numerical Experiments

Purpose: Evaluating the **VR_EPC_MILP** MILP model.

Technical Context: The experiments are performed on a computer with AMD EPYC 7H12 64-Core processor, and are running under Gnu/linux Ubuntu 20.04.2. The MILP library CPLEX 12.10 is used in a single-thread mode.

1) Instance generation: The main parameters of every instance are the station number M, the period number N, and the length p of the periods. M varies from 5 to 30; N varies from 40 to 320; p takes values in $\{1, 2, 4\}$. We derive from the instances with $p = 1$ the instances with $p = 2$ and $p = 4$ by merging the periods in a natural way, the length $TMax = p \cdot N$ of the time horizon remaining the same.

Vehicle coefficients: The M stations, together with $Depot$ and **HMP** are generated as integral points of a square in the 2D plane. Vectors T, θ, θ^* corresponds to a rounding of the Euclidean distance, and vectors E, ϵ, ϵ^* correspond to the Manhattan distance.

PV-Plant coefficients: We cluster production periods into 5 super-periods of same length, each provided with symbolic mean production and cost values $\overline{R_{cl}}$, $\overline{Cost_{cl}}$ in $\{Low, Medium, High\}$. Then, integral time dependent production coefficients R_i, together with time dependent cost convex piecewise (2 pieces) linear functions $\Phi_i, i = 0, \ldots, N-1$, are generated accordingly in such a way that resulting instance is feasible and that this feasibility requires the purchase of additional power.

Storage capacities: We control the number of recharge transactions by maintaining $\frac{C^{PVP}}{C^{Veh}}$ between 0.5 and 3.

TABLE II
BEHAVIOR OF **VR_EPC_MILP**

Id	(N, M, p)	LBG_LP	UBG_LP
1	(160, 10, 1)	56,7	173
2	(160, 10, 1)	76,6	131
1-2	(160, 10, 2)	58,5	175
2-2	(160, 10, 2)	71,7	135
3	(160, 10, 1)	51,6	171
4	(160, 10, 1)	64,0	125
3-2	(160, 10, 2)	51,7	174
4-2	(160, 10, 2)	63,4	134
5	(240, 20, 1)	43,8	148
6	(240, 20, 1)	59,5	176
7	(240, 20, 1)	83,2	208
5-2	(240, 20, 2)	43,9	148
6-2	(240, 20, 2)	59,1	201
7-2	(240, 20, 2)	82,8	255
8	(320, 30, 1)	58,0	151
9	(320, 30, 1)	51,7	117
10	(320, 30, 1)	63,2	150
8-2	(320, 30, 2)	58,0	153
9-2	(320, 30, 2)	51,2	127
10-2	(320, 30, 2)	63,08	163

Scaling coefficients α: We do in such a way that the weights of respectively $PCost$ and $\alpha.Veh_Time$ remain integral and comparable.

For every instance Id (Tables II), we provide input values values N, M, p. We tested 242 instances. Yet, we restrict ourselves here to 20 instances, with $p = 1$, referred to as Id, and $p = 2$ referred to as $Id - 2$.

2) Results: For every instance, Table II displays The lower bound LBG_LP and the upper bound UBG_UB computed through Branch and Cut by the CPLEX library in 2 CPU hours (7200 seconds). We boost the computation of upper bounds UBG_UB with the CPLEX parameter $MIP_emphasis_switch$.

Comments: Even when allowed to run during 2 hours, the model never reaches optimality, though it often allows the computation of a good feasible solution. That is why we are now going to derive from some statistical analysis some efficient heuristics for our problem.

IV. HEURISTIC HANDLING OF VR_EPC

The bi-level structure of **VR_EPC** and the fact that this bi-level structure will correspond in most cases to some collaborative decision context, suggests to manage **VR_EPC** while acting on the master route Γ through local search operators and devices that anticipate the behavior of resulting sub-problem **VR_EPC**(Γ) through the use of simple devices. There exists many ways to proceed this way. The approach which we are going to propose here relies on a statistical analysis of the correlation which may exist between the optimal value of **VR_EPC** and some key features of the route Γ.

In order to keep on, we need some additional notations:

- We denote by **VR_EPC**(Γ) the **VR_EPC** instance induced by fixing Γ (that means, referring to the **VR_EPC_MILP** MILP model, if we fix X in such a

TABLE III
CORRELATION ANALYSIS

It	1	2	3	4	5	6	7	8	9
$W(\Gamma)$	564	504	500	462	464	435	388	390	378
$L^T(\Gamma)$	197	183	182	165	162	152	135	132	126
NRT	11	10	9	9	9	9	7	7	8

TABLE IV
CORRELATION ANALYSIS

It	10	11	12	13	14	15	16	17	18
$W(\Gamma)$	564	504	500	462	319	315	314	326	324
$L^T(\Gamma)$	197	183	182	165	105	99	101	100	99
NRT	11	10	9	9	6	6	6	7	7

way it meets (VR3, VR4) and does not induce any sub-tour). We denote by $W(\Gamma)$ its optimal value;
- We denote by $L^T(\Gamma), L^E(\Gamma)$ the lengths of Γ in the sense of respectively vectors T and E.

Let us also recall the way the standard *Traveling Salesman 2_Opt* and *Reloc* operators act on any current route Γ through 2 parameters j_1, j_2 in $\mathbf{J} + \{Depot\}$:

- *2_Opt*(Γ, j_1, j_2) replaces the moves from j_1 to its successor $\bar{j}_1$ and from j_2 to its successor $\bar{j}_2$ by moves from j_1 to j_2 and from $\bar{j}_1$ to $\bar{j}_2$, and reverses the orientation of Γ between j_2 and $\bar{j}_1$;
- *Reloc*(Γ, j_1, j_2) relocates j_1 between j_2 and $\bar{j}_2$.

A. Experimentally Linking $W(\Gamma)$, $L^T(\Gamma)$ and $L^E(\Gamma)$

Tables III, ..., VI show that $W(\Gamma)$, $L^T(\Gamma)$, $L^E(\Gamma)$ and the number NRT of recharge transactions are strongly correlated along the 36 iterations $It = 1, \ldots, 36$ of a descent loop involving *2_Opt* and *Reloc* and performed on 20 stations. We observe the same kind of correlation when performing the same experimental process on all the instances involved in table II.

B. A Simple PDYN_EPC Algorithm for the Computation of an Upper Bound $W_Aux(\Gamma)$ of $W(\Gamma)$

Though fixing Γ deeply simplifies the **VR_EPC_MILP** MILP model, it does make it suitable for repeated applications inside a local search process. So we briefly describe here a simple dynamic programming algorithm *PDYN_EPC* which computes an upper bound $W_Aux(\Gamma)$ of $W(\Gamma)$ under running times that allow its insertion into such a local search process. We design *PDYN_EPC* by noticing that once Γ is fixed, a full solution is determined by the sequence of recharge transactions linking **EVeh** and **PVP** together, augmented with the amounts of power bought during the periods separating 2 consecutive recharge transactions. More precisely, *PDYN_EPC* searches for a path (a sequence of transitions) in a *state* graph whose nodes (the states) and arcs (decisions and transitions) are defined as follows:

- A *state* in the sense of *PDYN_EPC* is a 4-uple $S = (j, i, V_j^{Veh}, V_i^{PVP})$, where j is a station, i a period, V_j^{Veh} is the power load of **EVeh** when it leaves j and V_i^{PVP} the power load of **PVP** at the end of i, with the implicit meaning that a recharge transaction involving i and j has just been performed. Such a state S comes with the smallest cost value $\Pi(S)$ of the cost of a path connecting initial state S_0 to S according to the *PDYN_EPC* process.
- A decision consists, in a natural way, in a 4-uple $(j_1, i_1, m_1, \bar{y})$, where (j_1, i_1, m_1) means the next recharge transaction, and $\bar{y}$ means the power which is going to be bought by **PVP** during the periods $i + 1, \ldots, i_1 - 1$. Resulting state and feasibility constraints derive in a natural way from the **VR_EPC_MILP** setting.
- The cost of related transition is $p \cdot (i_1 - i)$ augmented with the purchase cost of $\bar{y}$. This purchase cost corresponds to the optimal value of some convex optimization program. Since it does not depend on the vehicle inputs, it is computed as part of a pre-process and stored in a table Pr_Cost which, with any pair of periods i, i_1 and any purchase value $\bar{y}$ in a target discrete set, associates the optimal cost of purchasing $\bar{y}$ energy during the periods $i + 1, \ldots, i_1 - 1$.
- Initial state S_0 is a 4-uple $(Depot, -1, H_0^{EVeh}, H_0^{PVP})$. A final state is any 4-uple $(Depot, N - 1, V_j^{Veh} \geq H_0^{EVeh}, V_j^{PVP} \geq H_0^{PVP})$, taking into account that this final state is not related to any recharge transaction and that the cost of related transition must be adapted in order to refer to the time when **EVeh** is back to $Depot$.

We do not detail the description of *PDYN_EPC*, which by many ways works as the standard path search algorithm *A**, and restrict ourselves to say that:

- In order to make this algorithm run fast, we fix an upper bound $NDec$ (in our case, we shall set $NDec = 10$) on the number of feasible decisions that may be tried from a given state S. A consequence is that *PDYN_EPC*$(NDec)$ is only a heuristic algorithm, that provides us with an upper bound for the optimal value $W(\Gamma)$ of **VR_EPC**(Γ).
- We introduce a lower bound computation device, which, to any state $S = (j, i, V_j^{Veh}, V_i^{PVP})$, makes correspond some lower bound $VAL(S)$ of the cost of a path in the state graph that would connect S to some final state. This lower bound device allows us to discard states S which appear to be poorly promising according to related values

TABLE V
CORRELATION ANALYSIS

It	19	20	21	22	23	24	25	26	27
$W(\Gamma)$	205	202	211	200	199	178	178	185	179
$L^T(\Gamma)$	72	70	67	64	64	62	62	61	61
NRT	3	3	4	3	3	4	4	3	3

TABLE VI
CORRELATION ANALYSIS

It	28	29	30	31	32	33	34	35	36
$W(\Gamma)$	173	297	259	267	263	276	275	247	205
$L^T(\Gamma)$	57	97	89	89	87	87	87	76	75
NRT	3	4	3	3	3	4	3	3	4

$(\Pi(S), VAL(S))$.

C. Two Simple Heuristic Algorithms

As told at the beginning of this section, above experiments suggest handling Γ while partially short-cutting the computation of $W(\Gamma)$ and relying on $L^T(\Gamma)$ and $L^E(\Gamma)$ in order to drive Γ towards good solutions. We do it here in two ways:

- A GRASP Algorithm *GRASP_VR_EPC*(Q), Q being a number of replications, that perform Q descent processes on Q randomly generated linear combinations of $L^T(\Gamma)$ and $L^E(\Gamma)$;
- An Algorithm *Descent_VR_EPC* that pipelines *GRASP_VR_EPC*(Q) and some pseudo-descent process parametrized by 2 flexibility parameters δ^{Time} and δ^{Ener}, that allow modifying the route Γ even when neither δ^{Time} nor δ^{Ener} decreases.

The GRASP Algorithm *GRASP_VR_EPC*.
It considers a replication parameter Q and works as follows:

1) At any replication q, it randomly generates two non negative scaling parameters ω^{Time} and ω^{Ener} together with an initial route $\Gamma(q)$;
2) Then it applies *2_Opt* and *Reloc* until $\Gamma(q)$ becomes a local optimum with respect to the surrogate objective function $\omega^{Time} \cdot L^T(\Gamma) + \omega^{Ener} \cdot L^E(\Gamma)$;
3) Finally it computes $W_Aux(\Gamma(q))$ while applying the *PDYN_EPC* algorithm and updates the best current route $\Gamma(q_{Best})$. When the whole process is over, it computes the exact value $W(\Gamma(q_{Best}))$ while applying the restriction of **VR_EPC_MILP** to Γ.

This is summarized by Algorithm 1:

Algorithm 1: GRASP_VR_EPC(Replication Parameter Q).

input : The inputs of the **VR_EPC Problem**, that we suppose to be feasible
output: A best route Γ^{Best} and related value W^{Best}

$\Gamma^{Best} \leftarrow Undefined$; $W_Aux^{Best} \leftarrow +\infty$;
for $q = 0, \ldots, Q$ **do**
  Randomly generate coefficients ω^{Time} and ω^{Ener} ;
  Randomly generate $\Gamma(q)$; `Stop` $\leftarrow$ `False` ;
  while not `Stop` **do**
    By applying *2_Opt* or *Reloc*, decrease $\omega^{Time} \cdot L^T(\Gamma(q)) + \omega^{Ener} \cdot L^E(\Gamma(q))$;
    if *Fail(decrease)* **then**
      `Stop` $\leftarrow$ `True`
    else
      Modify $\Gamma(q)$ accordingly
    end
  end
  Compute $W_Aux(\Gamma(q))$ by applying *PDYN_EPC*;
  if $W_Aux^{Best} > W_Aux(\Gamma(q))$ **then**
    $\Gamma^{Best} \leftarrow \Gamma(q)$; $W_Aux^{Best} \leftarrow W_Aux(\Gamma(q))$;
end
Compute the exact value W^{Best} (by application of the restriction of **VR_EPC_MILP** to Γ^{Best};
return Γ^{Best} and W^{Best}.

The Pseudo-Descent Algorithm *Descent_VR_EPC*: Clearly, it may occur that current route Γ may be improved by operators *2_Opt* and *Reloc* while neither $L^{Time}(\Gamma)$ nor $L^{Ener}(\Gamma)$ may be improved. This suggests allowing the *2_Opt* and *Reloc* operators to slightly deteriorate the values $L^{Time}(\Gamma)$ and $L^{Ener}(\Gamma)$ and deciding about their application after computing $W_Aux(\Gamma)$ while relying on the algorithm *PDYN_EPC*. More precisely, let us consider 2 parameters $\delta^{Time} > 0, \delta^{Ener} > 0$ and 2 routes Γ_1, Γ_2. We say that Γ_2 deteriorates Γ_1 by no more than $(\delta^{Time}, \delta^{Ener})$ if at least one of the 2 following inequalities holds:

1) $L^{Time}(\Gamma_2) - L^{Time}(\Gamma_1) \leq \delta_1$;
2) $L^{Ener}(\Gamma_2) - L^{Ener}(\Gamma_1) \leq \delta_2$.

Then the parametric *Descent_VR_EPC*($\delta^{Time}, \delta^{Ener}$) algorithm works according to the following descent loop:

1) It initializes Γ as a *good quality* route (for instance by applying *GRASP_SVR_EP*(1) (1 replication));
2) At any iteration of the descent process, it generates all the *2_Opt* and *Reloc* parameters (j_1, j_2) such that resulting route deteriorates Γ by no more than $(\delta^{Time}, \delta^{Ener})$;
3) For any parameter (j_1, j_2) selected this way and any resulting route Γ_1 it tests its impact on the value $W_Aux(\Gamma_1)$ obtained through application of the *PDYN_EPC* Algorithm;
4) If improving Γ this way is possible, then it does it according to a *Best Descent* strategy else it stops. In case it stops then it compute the exact current value $W(\Gamma)$.

This is summarized by Algorithm 2:

Remark 2: Both *GRASP_VR_EPC* and *Descent_VR_EPC* might be easily extended along more sophisticated metaheuristic scheme, more specifically genetic and ant colony based schemes (see [9] and [7]).

D. Numerical Experiments

Table VII involves the same instances as Table II. It displays:

- The value $GRASP_50$ obtained by *GRASP_VR_EPC*(50) (50 replications) and related CPU time CPU_Grasp_50.
- The value $GRASP_150$ obtained by *GRASP_VR_EPC*(150) (150 replications).

Table VIII considers the same instances as Table III and displays The value $Desc_1 + LS$ obtained by the application of the following pipeline: *GRASP_VR_EPC*(1) (1 replication) $\rightarrow$ *Descent_VR_EPC*(4, 4), together with related CPU time $CPU_Desc + LS$.

Comments: *GRASP_VR_EPC* with 150 replications most often reaches quasi-optimality. It outperforms the upper bound computed by CPLEX in 4 CPU hours in 8 among the 20 instances and obtains the same value as CPLEX in 6 other instances. This trend is reinforced when $M = 30$.

Algorithm 2: Descent_VR_EPC(Parameters $\delta^{Time} \geq 0, \delta^{Ener} \geq 0$).

input : The inputs of the **VR_EPC Problem**, that we suppose to be feasible
output: A (good quality) route Γ

Initialize Γ and compute an approximation W of $W(\Gamma)$ by applying the *PDYN_EPC* Algorithm (value $W_Aux(\Gamma)$);
`Stop` $\leftarrow$ `False` ;
while not `Stop` **do**
  Generate all the parameters j_1, j_2 of the *2_Opt* operator such that applying *2_Opt* or *Reloc* to Γ deteriorates Γ by no more than $(\delta^{Time}, \delta^{Ener})$;
  for *Any 3-uple* (Op, j_1, j_2) *(Op meaning either Reloc or 2_Opt) generated this way* **do**
    Apply the *PDYN_EPC* Algorithm to the route resulting from application of (Op, j_1, j_2) to Γ;
  end
  if *A decrease of* $W_Aux(\Gamma)$ *with respect to* W *is obtained at least once* **then**
    Replace Γ by $Op(\Gamma, j_1, j_2)$ and W by related W_Aux value (*PDYN_EPC* Algorithm), with (Op, j_1, j_2) related to the best improvement
  else
    `Stop` $\leftarrow$ `False`;
  end
end
Compute the exact value $W(\Gamma)$ (by application of the restriction of **VR_EPC_MILP** to Γ;
return Γ and $W(\Gamma)$.

Applying the pipe-line *GRASP_VR_EPC*(1) (1 replication) $\rightarrow$ *Descent_VR_EPC*(4, 4) improves the value obtained by *GRASP_VR_EP5*(50) in 8 among the 20 instances, and the value obtained by *GRASP_VR_EPC*(150) in 5 among the 20 instances. The running times induced by the calls to *Descent_VR_EPC*(4, 4) are rather high, due to the fact that we did not optimize the transformation of a route Γ into a *PDYN_EPC* input. Also, we may notice that in some cases, *Descent_VR_EPC*(4, 4) yields poor results (if we refer to *GRASP_VR_EP5*(50)) due to a poor initial route computed by a single *GRASP_VR_EPC*(1) run.

V. Conclusion

We dealt here with a complex model which make interact a routing process with an energy production process. W adopted a centralized point of view, which skips uncertainties, and tried both an exact MILP approach and a heuristic approach based upon a statistical correlation analysis. In the future, it will of course be interesting to try other methods. Also we plan dealing with *collaborative* contexts, when several *consumers* interact, each provided with its own agenda and incomplete information, and with the issue related to the uncertainty induced by solar energy production.

TABLE VII
Results for *GRASP_VR_EPC*

Id	$GRASP_50$	$GRASP_150$	CPU_Grasp_50
1	171	171	52,9
2	131	131	95,1
1-2	175	175	10,4
2-2	135	135	14,8
3	173	173	58,4
4	129	126	71,4
3-2	176	176	19,8
4-2	138	134	20,8
5	156	156	158,7
6	175	175	333,5
7	214	214	296,2
5-2	158	158	58,5
6-2	181	181	106,3
7-2	226	226	74,6
8	121	121	227,5
9	123	119	225,5
10	159	159	317,6
8-1	135	132	91,7
9-1	128	123	78,7
10-1	163	153	96,7

TABLE VIII
Results for *GRASP_VR_EPC*

Id	$GRASP_150$	$Desc_1 + LS$	$CPU_Desc + LS$
1	171	171	22,7
2	131	134	16,5
1-2	175	175	15,6
2-2	135	140	5,4
3	173	165	197,0
4	126	144	111,0
3-2	176	171	36,7
4-2	134	161	41,1
5	156	168	356,2
6	175	176	565,7
7	214	207	254,4
5-2	158	173	109,8
6-2	181	181	203,6
7-2	226	220	109,0
8	121	121	1998,6
9	119	122	2351,0
10	159	151	876,2
8-1	132	134	773,0
9-1	123	130	629,1
10-1	153	159	633,5

References

[1] C. Artigues, E. Hébrard, A. Quilliot, H. Toussaint, *Models and algorithms for natural disaster evacuation problems*, Proc. 2019 FEDCSIS WCO Conf., p 143-146, 2019. doi.org/10.15439/2019F90

[2] Bendali F., Mole Kamga E., Mailfert J., Quilliot A., Toussaint H.: *Synchronizing Energy Production and Vehicle Routing*. RAIRO-O.R, 55 (4),pp. 2141-2163. (2021). https://doi.org/10.1051/ro/2021093.

[3] F. Bendali, J. Mailfert, E. Mole-Kamga, A. Quilliot, H. Toussaint, *Pipelining dynamic programming process in order to synchronize energy production and consumption*, Proc. 2020 FEDCSIS WCO Conf., p 303-306, 2020. doi.org/10.15439/978-83-955416-7-4.

[4] Drexl M.: Synchronization in vehicle routing–a survey; *Transportation Science* 46, pp. 297-316, (2012). https://dx.doi.org/10.1287/trsc.1110.0400

[5] Erdelic T., Caric T., Lalla-Ruiz E.: A Survey on the Electric Vehicle Routing Problem.*Journal of Advanced Transportation* (2019). https://doi.org/10.1155/2019/5075671.

[6] Erdogan S., Miller-Hooks E.. A green vehicle routing prob-

lem. *Transportation Research Part E* , 109, p 100–114, (2012). https://doi.org/10.1016/j.tre.2011.08.001

[7] S. Fidanova, O. Roeva, M. Ganzha, *Ant colony optimization algorithm for fuzzy transport modelling*, Proc. 2020 FEDCSIS WCO Conference, p 237-240, 2020. doi.org/10.15439/978-83-955416-7-4

[8] Franceschetti A., Demir E., Honhon D., Van Woensel T., Laporte G., and Stobbe M.: A metaheuristic for the time dependent pollution-routing problem. *EJOR*, 259 (3) :972 – 991, (2017). https://doi.org/10.1016/j.ejor.2016.11.026

[9] K. Gmyrek, M. Antkiewicz, P. Myszkowski: Genetic Algorithm for Planning and Scheduling Problem – StarCraft II Build Order case study; *Proceedings of the 18th Conference on Computer Science and Intelligence Systems* ACSIS, Vol. 35, pages 131–140 (2023). DOI: http://dx.doi.org/10.15439/2023F6015.

[10] Hung Y. F., Chien K. L.: A multi-class multi-level capacitated lot sizing model;*Journal of the Operational Research Society*, Palgrave Macmillan;The OR Society, vol. 51(11), pages 1309-1318, (2000). DOI: 10.1057/palgrave.jors.2601026.

[11] Irani S, Pruhs K.: Algorithmic problems in power management. *ACM SIGACT News* 36-2, p 63-76 (2005). DOI:10.1145/1067309.1067324.

[12] Koç C., Jabali O., Mendoza J., Laporte G.. The electric vehicle routing problem with shared charging stations. *ITOR*, 26, (2018). DOI:10.1111/itor.12620

[13] Kuo Y.. Simulated annealing to minimize fuel consumption for the time-dependent VRP. *Comp. Indust. Eng.*, 59 :157–165, (2010). https://doi.org/10.1016/j.cie.2010.03.012

[14] Lajunen A.: Energy consumption and cost analysis of electric city buses. *Transportation Research Part C*, 38 :1–15, (2014). https://doi.org/10.1016/j.trc.2013.10.008

[15] Luthander R., Widen J., Nilsson D., Palm J.: Photovoltaic self-consumption in buildings. *Applied Energy* 142, p 80-94, (2015). https://doi.org/10.1016/j.apenergy.2014.12.028

[16] Macrina G., Di Puglia L., Guerriero F.: The Green-Vehicle Routing Problem: A Survey. *Model. and Optim. in Logistics* p 1-26 (2020). DOI: https://doi.org/10.1007/978-3-030-45308-4_1.

[17] Raylan M., Matos S., Frota Y., and Satoru Ochi L.: Green vehicle routing and scheduling problem with split delivery. *Electronic Notes in Discrete Mathematics*, 69 :13 – 20, 2018. https://doi.org/10.1016/j.endm.2018.07.003

[18] Sachenbacher M., Leucker M., Artmeier A., Haselmayr J.: Efficient energy routing for electric vehicles. Proc *AAAI 2011 Int. Conf*, (2011). https://doi.org/10.1609/aaai.v25i1.7803

[19] Schneider M., Stenger A., Goeke D.: The electric vehicle-routing problem with time windows and recharging stations. *Transportation Science* 48, p 500-520, (2014). DOI:10.1287/trsc.2013.0490

[20] K. Stoilova, T. Stoilov, *Bi-level optimization application for urban traffic management*, Proc. 2020 FEDCSIS WCO Conf., p 327-336, 2020. doi.org/10.15439/978-83-949419-5-6.

[21] Verma A.: Electric vehicle routing problem with time windows, recharging stations and battery swapping stations. *Euro Journal of Transportation Logistics* 7, p 415-451, (2018).

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 587–594
DOI: 10.15439/2025F1136
ISSN 2300-5963 ACSIS, Vol. 43

An Application of Tensors in the Stochastic Reaction Diffusion Master Equation

Md Mustafijur Rahman
0000-0002-6874-0801
Texas A&M University - San Antonio,
Department of Computational, Engineering,
and Mathematical Sciences (CEMS),
San Antonio, TX 78224, USA
Email: mrahman2@tamusa.edu

Roger B. Sidje
0000-0001-5353-3642
The University of Alabama,
Department of Mathematics,
Tuscaloosa, AL 35487, USA
Email: roger.b.sidje@ua.edu

***Abstract*—Various scientific fields are seeking to exploit tensors, which are multi-dimensional arrays that expand on the concept of matrices. Here, we consider an application arising from complex biological systems where the data can be processed and analyzed using tensors. Our modeling framework is the reaction-diffusion master equation (RDME) used to describe the dynamics of biological systems involving both reaction and diffusion processes. It is already notoriously hard to solve the more familiar chemical master equation (CME) that only involves reaction processes. Solving the RDME is even harder because its state space is considerably larger compared to that of the CME, and this further motivates the utilization of tensors. Our study is an illustrative example of how tensor techniques can be used to make predictions on the dynamics of a metapopulation model based on its RDME formulation.**

I. Introduction

TENSORS, fundamentally, are extensions of vectors, and matrices, with the ability to represent and manipulate data in multiple dimensions. Classical linear algebra approaches can circumvent tensors by flattening the data through mechanisms known as matricization or vectorization, although doing so can increase the computational overhead, especially as dimensionality grows. By contrast, specialized tensor methods avoid unfolding the tensor into a matrix format, with the trade-off that the preserved multi-dimensional format drastically complicates the design and implementation of algorithms, not to mention their subsequent understanding and maintenance by others than the original software architects.

Tensor methods encompass the efficient utilization of tensors for the purpose of problem-solving, data analysis, and the extraction of significant patterns from high-dimensional datasets. Tensors find utility not only in the biological application considered here, but also in many other areas, including popular image processing applications (e.g., facial recognition, analysis of musical scores, and identification of cliques within social networks [1], [2]). Together with earlier works on tensors and their decompositions that go back to 1927 [3], we note that more recent literature overviews of tensor techniques for large-scale numerical computations are given in [4], [5], [6], geared towards a scientific computing audience. Of course, there are way more applications of tensors than we can reasonably cite in a communication such as this one.

In the field of biochemistry, complex networks are shown to emerge as a result of the interactions among many cellular components, including DNA and RNA molecules. Determining the time-varying probability distribution of their states involves solving the chemical master equation (CME) [7]. The stochastic simulation algorithm (SSA) [8] is useful for analyzing the CME. An alternative method for addressing the CME is the finite state projection algorithm (FSP) [9], which amounts to solving differential equations that characterize the evolution of the probability distribution of the underlying Markov chain of the CME over time.

Spatial heterogeneity is a frequent property of many cellular habitats, which are not homogeneous in space. A number of distinct but similar mathematical techniques have been used to model stochastic reaction-diffusion systems in biological cells [10], [11]. For example, Smoluchowski proposed a modeling approach named the spatially continuous diffusion-limited reaction (SCDLR) [12]. Unlike these approaches, the reaction-diffusion master equation (RDME) [13] is obtained when we treat the diffusion at the molecular level as a unique set of reactions in the CME.

The space is partitioned into compartments in the RDME. Every compartment is thought to be thoroughly mixed, and the reactions found within a specific compartment are thought to be consistent with the homogeneous scenario. Additionally, molecules can diffuse (jump) in between neighboring compartments, and here we will model the jump processes as reactions. Similar to the CME but with a much larger dimensionality, the RDME explains the time-dependent probability distribution function (PDF) of the system state. The works in [14], [15], [16] provide instances of the SSA being used in reaction-diffusion systems with the space being one dimensional. The next reaction method (NRM) [17] is an efficient version of the SSA. A software called MesoRD [18] offers a specialized implementation of the NRM, which is designed for diffusive systems and known as the next subvolume method (NSM) [19].

In this contribution rooted in [20], our aim is to illustrate

how a tensor-based approach, specifically a quantized tensor-train decomposition, can be effectively used to manage the potentially high-dimensional state space arising in the RDME of a metapopulation model. Although the RDME shares similarities with the CME, it necessitates greater computational resources due to its considerably larger state space. We employ the tensor technique to derive the marginal distribution of the relevant species. By comparing the tensor technique with the classical FSP, this work aims to highlight the capability of tensors to reduce computational overhead while maintaining accuracy and generalizability across diverse applications.

II. Preliminaries

A. Tensors

We will soon recap how the CME and RDME lead to formulations involving tensors. We begin with some background to define the notation and situate our context. More formally, tensors are multi-dimensional arrays in $\mathbb{K}^{I_1 \times \cdots \times I_N}$, where N is the order, i.e., the number of dimensions, which are called modes or ways, and $I_1, \ldots, I_N$, are index ranges in each dimension. Tensors extend the notion of vectors and matrices, referred to as 1-way and 2-way tensors, respectively, whereas a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, depicted in Fig. 1, is a cuboid (or cube-like) with elements $x_{ijk} \equiv \mathcal{X}(i, j, k)$, $1 \leq i \leq I_1, 1 \leq j \leq I_2, 1 \leq k \leq I_3$. The field $\mathbb{K}$ will be $\mathbb{R}$ in our application but it can be $\mathbb{C}$ in general.

Index fixing can be used to build subarrays. A typical scenario is when all indices are fixed except one, in which case the resulting subarrays are referred to as fibers, c.f. Fig. 2. They can alternatively be thought of as mode-n fibers that are produced when all indices are fixed except the nth one. Another common scenario is when all indices are fixed except two, yielding the examples of slices depicted on Fig. 3 for a 3-way tensor. These examples make clear that there are many different possible orderings, meaning that software developers must agree on common conventions for consistency in their codes. If we stack the resulting fibers or slices, we are performing what is termed vectorization or matricization. In general, tensor unfolding denotes the mechanism by which to flatten a tensor to systematically arrange its entries into a matrix, thereby providing a bridge to circumvent tensors and revert to classical linear algebra operations.

On tensors of equal sizes, simple operations like addition and subtraction are performed component-by-component. However, beyond ordinary matrix-like definitions, there are other tensor-specific operations and decompositions that are much more sophisticated by far.

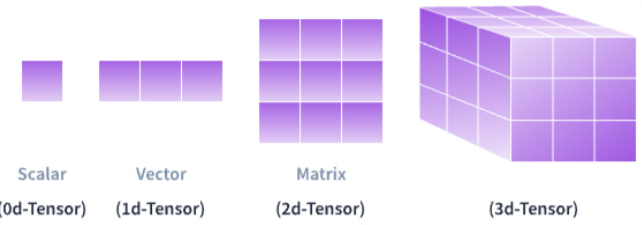


Fig. 1. Tensors of order 0, 1, 2, and 3.

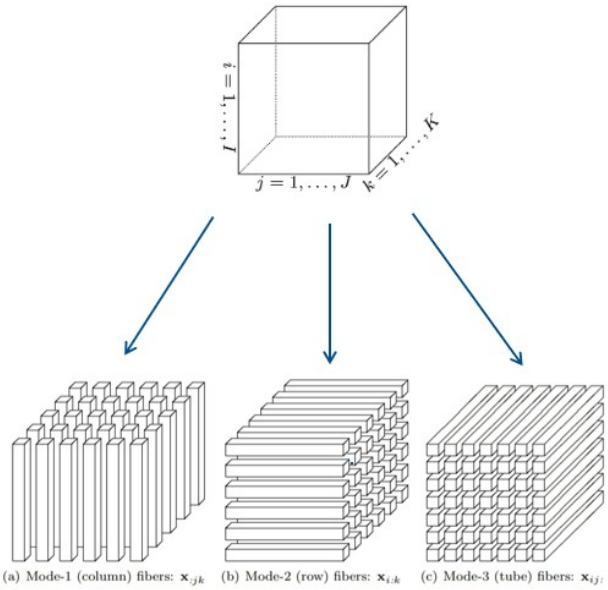


Fig. 2. Different fibers of a 3-mode tensor.

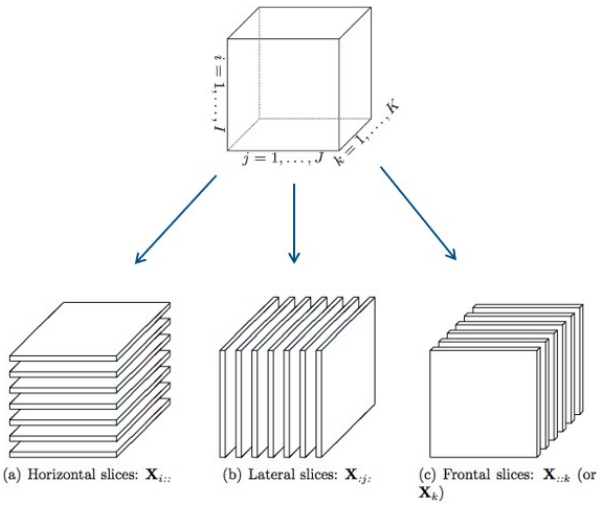


Fig. 3. Different slices of a 3-mode tensor.

B. Higher Order Singular Value Decomposition (HOSVD)

Consider the Singular Value Decomposition (SVD), which is a fundamental matrix factorization that represents any matrix as the product of two orthogonal matrices together with a diagonal matrix of singular values, encapsulating the inherent geometric characteristics of the data. Higher Order Singular Value Decomposition (HOSVD) generalizes this concept to multi-dimensional tensors, by decomposing them into a core tensor and a collection of orthogonal matrices corresponding to each mode. Below we reproduce a characterization of the HOSVD [21], mainly to illustrate how concepts that are familiar with matrices can quickly become drastically harder to grasp when reformulated through tensors.

Theorem II-B.1: Every tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ can be written as

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_N \mathbf{A}^{(N)},$$

where

1) $\mathbf{A}^{(n)} = \begin{pmatrix} A_1^{(n)} & A_2^{(n)} & \cdots & A_{I_n}^{(n)} \end{pmatrix}$ is an orthogonal matrix of order $I_n \times I_n$,
2) $\mathcal{G} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ of which the subtensors $\mathcal{G}_{i_n=\alpha}$, attained by fixing the nth index to α, have the properties of
   (i) ordering:
   $$\|\mathcal{G}_{i_n=1}\|_F \geq \|\mathcal{G}_{i_n=2}\|_F \geq \cdots \geq \|\mathcal{G}_{i_n=I_n}\|_F \geq 0,$$
   for all possible values of n,
   (ii) all-orthogonality: for all possible values of n, α and β, two subtensors $\mathcal{G}_{i_n=\alpha}$ and $\mathcal{G}_{i_n=\beta}$ are orthogonal subject to $\alpha \neq \beta$:
   $$\langle \mathcal{G}_{i_n=\alpha}, \mathcal{G}_{i_n=\beta} \rangle = 0 \text{ for } \alpha \neq \beta.$$
3) $\mathcal{G} = \mathcal{X} \times_1 \mathbf{A}^{(1)T} \times_2 \mathbf{A}^{(2)T} \cdots \times_N \mathbf{A}^{(N)T}$.

The vector $A_i^{(n)}$ is an n-mode singular vector and the Frobenius-norms $\|\mathcal{G}_{i_n=i}\|_F$ are n-mode singular values of $\mathcal{X}$. N denotes the total number of modes of the tensor, and $n \in \{1, \ldots, N\}$ refers to an arbitrary mode index. The symbol $\times_n$ denotes the mode-n product, which is an operation where each mode-n fiber of a tensor is multiplied by a matrix, affecting only the nth dimension while keeping others fixed.

C. CANDECOMP/PARAFAC (CP) Decomposition

A popular decomposition in practice is the CP decomposition, which for $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, is given by [22],

$$\mathcal{X} = \mathcal{I} \times_1 \mathbf{A}^{(1)} \cdots \times_N \mathbf{A}^{(N)} \equiv \left[\!\!\left[\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)} \right]\!\!\right] \quad (1)$$

where $\mathcal{I}$ is the identity tensor having ones along the super-diagonal and zeros elsewhere, and each $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ for $n = 1, \ldots, N$ is a factor matrix. The notation $[\![\cdot]\!]$ denotes the CP decomposition as a sum of R rank-one tensors formed by outer products of the columns of the factor matrices. The CP rank $R \in \mathbb{N}$ is left as a parameter that tunes the accuracy of the decomposition. The fact that CP boils down to resorting to matrices explains its appeal and popularity.

According to [23], (1) can be written element-wise as a sum of products of the entries of those matrices

$$\mathcal{X}(i_1, \ldots, i_N) = \sum_{r=1}^{R} \mathbf{A}^{(1)}(i_1, r) \cdots \mathbf{A}^{(N)}(i_N, r). \quad (2)$$

Equation (3) below is another compact notation for writing (1) or (2) via the outer product

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(N)}. \quad (3)$$

by referencing columns of $\mathbf{A}^{(n)} = \begin{pmatrix} \mathbf{a}_1^{(n)} & \ldots & \mathbf{a}_R^{(n)} \end{pmatrix} \in \mathbb{R}^{I_n \times R}, n = 1, \ldots, N$.

D. Tensor Train (TT) Decomposition

Our present numerical implementation is based on another decomposition, namely the tensor train (TT) decomposition. For any $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, it can be shown that its entries can be written in the form [24],

$$\mathcal{X}(i_1, \ldots, i_N) = \mathcal{G}_1(i_1) \cdot \ldots \cdot \mathcal{G}_N(i_N). \quad (4)$$

Here, $\mathcal{G}_n(i_n) \in \mathbb{R}^{r_{n-1} \times r_n}$ are matrices (slices) of the cuboid (3-way) tensors $\mathcal{G}_n \in \mathbb{R}^{r_{n-1} \times r_n \times r_n}$, which are called TT cores. Moreover, "boundary conditions" $r_0 = r_N = 1$ are imposed so that the chain product (4) always yields a scalar. TT ranks are defined as $(r_1, \cdots, r_{N-1})$.

An explicit storage of $\mathcal{X}$ would cost $\prod_{n=1}^{N} I_n = O(I^N)$, where $I = \max_{1 \leq n \leq N} I_n$, so that if $N = 10$ in a CME application and $I \approx 10$, its corresponding matrix (7) would be of immense size $10^{10} \times 10^{10}$ and even much more for the RDME. By contrast, the TT format only stores the TT cores at a total cost of $O(r_{TT}^2 \cdot N \cdot I)$, where $r_{TT} = \max_{1 \leq n \leq N} r_n$. Thus, provided that the maximum rank r_{TT} is not large, the storage of the TT format scales linearly with N, leading to dramatic savings from the explicit format where the storage scales exponentially with N. Our actual code builds on another variant, specifically the quantized tensor train (QTT) format that brings the storage requirements further down to $O(r_{QTT}^2 \cdot N \cdot \log_2 I)$. Although r_{QTT} is not necessarily r_{TT}, this reduction remarkably means a linear cost in the number of modes and logarithmic in the mode sizes.

Our implementation did not have to invent everything from scratch. Rather, we leveraged past research, including [25], [26], and the TT toolbox software [24]. Our contribution here is sharing the findings of our hands-on experience on these topics and their application to the RDME context that we describe next.

E. CME and FSP in tensor format

Consider a chemical reacting system in a spatial domain Ω involving N species $\{S_1, \ldots, S_N\}$, represented by the state vector $\mathbf{X}(t) = [X_1(t), \ldots, X_N(t)]^T$, where each $X_i(t)$ is a non-negative integer denoting the population of species S_i at time t. Suppose there are M reaction channels, denoted

by $\{R_1, \ldots, R_M\}$, and assume that the system is well mixed and in thermal equilibrium. The dynamics of the jth reaction channel R_j is characterized by

(i) the propensity function $a_j(\mathbf{x})$, which indicates how likely it is that R_j will occur given the system's current state $\mathbf{x}$, and

(ii) the stoichiometric (or state change) vector $\boldsymbol{\nu}_j = [\nu_{1j}, \ldots, \nu_{Nj}]^T$, which specifies the change in the population count of each species that results from one occurrence of R_j.

Thus, $a_j(\mathbf{x})dt$ gives the probability that, given $\mathbf{X}(t) = \mathbf{x}$, one R_j reaction will occur in the next infinitesimal time interval $[t, t + dt]$, and ν_{ij} gives the change in X_i induced by one R_j reaction. This system can be modeled as a Markov chain, with its behavior determined by the Chemical Master Equation (CME) [7]

$$\frac{dP(\mathbf{x},t)}{dt} = \sum_{j=1}^{M} \Big[a_j(\mathbf{x} - \boldsymbol{\nu}_j)P(\mathbf{x} - \boldsymbol{\nu}_j, t) - a_j(\mathbf{x})P(\mathbf{x},t)\Big], \tag{5}$$

where the function $P(\mathbf{x}, t)$ denotes the probability that $\mathbf{X}(t)$ will be $\mathbf{x}$. Equation (5) may be written in an equivalent matrix–vector form by enumerating all the states. If there are n possible states, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the CME takes the form of a system of linear ordinary differential equations (ODEs),

$$\dot{\mathbf{P}}(t) = \boldsymbol{\mathcal{M}} \cdot \mathbf{P}(t), \tag{6}$$

where the probability vector $\mathbf{P} = [p_1, \ldots, p_n]^T$ is such that each component $p_i \equiv P(\mathbf{x}_i, t) \equiv \text{Prob}\{\mathbf{X}(t) = \mathbf{x}_i\}$, the probability of being at state $\mathbf{x}_i$ at time t, for $i = 1, \ldots, n$.

The matrix $\boldsymbol{\mathcal{M}}$ is a sparse $n \times n$ matrix that is populated by the propensities and it represents the transition rate matrix of the Markov chain underlying the CME. Its entries are constructed using

$$\boldsymbol{\mathcal{M}}(i,j) = \begin{cases} a_k(\mathbf{x}_j) \text{ if } \mathbf{x}_i = \mathbf{x}_j + \boldsymbol{\nu}_k \\ -\sum_{k=1}^{M} a_k(\mathbf{x}_j) \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}. \tag{7}$$

The FSP method [9] and improvements such as [27] directly compute an approximation to the solution of the CME. The FSP method solves the CME and estimates the probability vector (PV) of the populations in a chemical reaction system using a truncated state space. For a truncated transition matrix $\boldsymbol{\mathcal{M}}_T$ and initial truncated PV denoted by $\mathbf{P}_T(0)$, the FSP finds the PV at any time $\mathbf{P}_T(t)$ by the following,

$$\dot{\mathbf{P}}_T(t) = \boldsymbol{\mathcal{M}}_T \cdot \mathbf{P}_T(t). \tag{8}$$

Equation (8) is a system of linear ODEs and the solution is given by

$$\mathbf{P}_T(t) = \exp(\boldsymbol{\mathcal{M}}_T t)\, \mathbf{P}_T(0). \tag{9}$$

The CME matrix $\boldsymbol{\mathcal{M}}$ remains sparse under any order relation imposed on the enumeration of the states, and it is natural to use the lexicographic order on the state space as a subset of $\mathbb{N}$, the set of nonnegative natural numbers. If we impose the lexicographic order on a state space $\mathbb{X}$ in the form of a window of lower bounds $(lb^{(1)}, \ldots, lb^{(N)})$ and upper bounds $(ub^{(1)}, \ldots, ub^{(N)})$, i.e., $\mathbb{X}$ consists of states $\mathbf{x} = (x_1, \cdots, x_N)^T$ such that $lb^{(s)} \le x_s \le ub^{(s)}, \forall s = 1, \ldots, N$, then the transition rate matrix can be decomposed into a sum of Kronecker products [28], [29]

$$\boldsymbol{\mathcal{M}} = \sum_{k=1}^{M} \left(\otimes_{s=1}^{N} \mathbf{S}_k^{(s)} - \mathbf{I}\right) \mathbf{M}_k. \tag{10}$$

Each term in the sum (10) corresponds to a reaction: $\mathbf{S}_k^{(s)}$ is a "shift-diagonal" matrix corresponding to the change in species s when reaction k happens; $\mathbf{I}$ is the identity matrix; and $\mathbf{M}_k$ is the diagonal matrix that stores the values of the propensity function a_k. Precisely, let $I_s = ub^{(s)} - lb^{(s)} + 1$ be the window size for each species, $n = I_1 \cdot \ldots \cdot I_N$ is the total number of states in $\mathbb{X}$, and denote $\boldsymbol{\nu}_k = \left(\nu_k^{(1)}, \cdots, \nu_k^{(N)}\right)^T$. The identity matrix $\mathbf{I}$ is then of order n, $\mathbf{S}_k^{(s)} \in \mathbb{R}^{I_s \times I_s}$ is given by

$$\mathbf{S}_k^{(s)} \equiv \begin{cases} \begin{pmatrix} 0 & \cdots & 1 & & \\ & \ddots & & \ddots & \\ & & \ddots & & 1 \\ & & & \ddots & \vdots \\ & & & & 0 \end{pmatrix} & \text{if } \nu_k^{(s)} < 0, \text{shifted up } \left|\nu_k^{(s)}\right| \text{ rows} \\ \begin{pmatrix} 0 & & & & \\ \vdots & \ddots & & & \\ 1 & & \ddots & & \\ & \ddots & & \ddots & \\ & & 1 & \cdots & 0 \end{pmatrix} & \text{if } \nu_k^{(s)} \ge 0, \text{shifted down } \nu_k^{(s)} \text{ rows} \end{cases}$$

and

$$\mathbf{M}_k \equiv \text{diag}\left(a_k(\mathbf{x}_1), \cdots, a_k(\mathbf{x}_n)\right), \tag{11}$$

where $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are states of $\mathbb{X}$ in the increasing lexicographic order. Furthermore, by assuming that the propensity functions are separable, i.e.,

$$a_k(\mathbf{x}) = a_k^{(1)}(x_1) \cdot \ldots \cdot a_k^{(N)}(x_N), \quad k = 1, \cdots, M,$$

where $a_k^{(\cdot)} \ge 0$ are scalar functions, (11) can be rewritten as

$$\mathbf{M}_k = \otimes_{s=1}^{N} \text{diag}\ a_k^{(s)},$$

where, for $s = 1, \cdots, N$,

$$\text{diag}\, a_k^{(s)} \equiv \text{diag}\left(a_k^{(s)}(lb^{(s)}), \cdots, a_k^{(s)}(ub^{(s)})\right).$$

Under these assumptions, the transition matrix (10) can be represented by simple matrices of small sizes,

$$\boldsymbol{\mathcal{M}} = \sum_{k=1}^{M} \left(\otimes_{s=1}^{N} \mathbf{S}_k^{(s)} - \mathbf{I}\right) \left(\otimes_{s=1}^{N} \text{diag}\ a_k^{(s)}\right). \tag{12}$$

In [26], the lower and upper bounds are in the form $lb^{(s)} = 0$ and $ub^{(s)} = I_s - 1$, yielding

$$\operatorname{diag} a_k^{(s)} = \operatorname{diag}\left(a_k^{(s)}(0), \cdots, a_k^{(s)}(I_s - 1)\right).$$

The reason for holding the lower bounds anchored at 0 was to help attenuate some of the complexity of dealing with tensor train decompositions when their sizes changed adaptively. Subsequent improvements in [25] allowed both the lower and upper bounds to slide dynamically. A central aspect in our body of research [25], [20], [26] has been that by keeping the state space in the form of a hyper-rectangle,

$$\mathbf{x} \in \left[lb^{(1)} : ub^{(1)}\right] \times \cdots \times \left[lb^{(N)} : ub^{(N)}\right],$$

both the transition operator $\boldsymbol{\mathcal{M}}$ and the probability distribution $\mathbf{P}$ can be represented as multi-dimensional arrays that allow us to use tensor decompositions as pioneered in [28], [29] to resolve storage issues.

F. RDME

Assume the spatial domain Ω is partitioned into compartments (voxels), denoted V_k, $k = 1, \ldots, K$. Within each compartment, species have the ability to react with one another as well as diffuse (jump) over the boundaries to reach adjacent compartments. Let $X_{i,k}(t)$ be the population count of species S_i in compartment V_k at time t. Then each species in the domain is given by the subvector $\mathbf{X}_i(t) = [X_{i,1}(t), \ldots, X_{i,K}(t)]$, $i = 1, \ldots, N$. Hence, $\mathbf{X}_1$ is the subvector which represents the species 1 in all compartments. Likewise, $\mathbf{X}_2$ represents species 2 in all compartments, and so on. Thus, the state vector of the system is $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_N]$. The diffusion propensity function $d_{i,j,k}$ and the state change vector $\boldsymbol{\mu}_{k,j}$ characterize the dynamics of the diffusion of species S_i from compartment V_k to V_j. The vector $\boldsymbol{\mu}_{k,j}$ has a length of K with -1 in the kth position, 1 in the jth position, and 0 elsewhere. Given $\mathbf{X}(t) = \mathbf{x}$, the diffusion master equation (DME) can be written as

$$\begin{aligned}\frac{dP(\mathbf{x},t)}{dt} = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{K}\Big[d_{i,j,k}(\mathbf{x}_i - \boldsymbol{\mu}_{k,j})P(\mathbf{x}_1, \ldots, \\ \mathbf{x}_i - \boldsymbol{\mu}_{k,j}, \ldots, \mathbf{x}_N, t) - d_{i,j,k}(\mathbf{x}_i)P(\mathbf{x},t)\Big].\end{aligned} \tag{13}$$

We take the diffusion propensity function $d_{i,j,k}(\mathbf{x}_i) = D/l^2$, for $k = j \pm 1$, where D is the diffusion rate and l is the characteristic length of the compartment, meaning that our setting ultimately assumes that $d_{i,j,k} \equiv d$ is constant. Some other settings could make different choices. If $\boldsymbol{\mathcal{D}}$ is the transition matrix describing the diffusion (or jump) of species between compartments, the equivalent matrix–vector form of (13) can be written as

$$\dot{\mathbf{P}}(t) = \boldsymbol{\mathcal{D}} \cdot \mathbf{P}(t). \tag{14}$$

Combining equations (6) and (14) we get the matrix-vector form of the RDME,

$$\dot{\mathbf{P}}(t) = \boldsymbol{\mathcal{M}} \cdot \mathbf{P}(t) + \boldsymbol{\mathcal{D}} \cdot \mathbf{P}(t). \tag{15}$$

Equation (15) is a system of linear constant coefficient ODEs and gives us more possible states than the CME, and so its corresponding transition matrix is substantially extended to represent species in compartments. However, it should be remembered that it is ultimately a tensor representation of the form (12) that is used in the calculations.

III. Study Findings

A. Metapopulation Model

Populations that are segmented into smaller groups linked by migration are studied using metapopulation models. Metapopulation models often employ reaction-diffusion principles to describe spatial dynamics [30]. In a reaction-diffusion metapopulation model, the reaction term depicts the occurences of births and deaths within each subpopulation, whereas the diffusion term captures the mobility of an individual between nearby subpopulations. By taking into account both diffusion and reaction terms, the model can better capture complex interactions between local population dynamics and spatial dispersal. Using the RDME in a metapopulation model, each subpopulation's population is modeled as a stochastic process with birth, death, and migration events occurring randomly. We examine a simple reaction scheme in this study that conserves the population count in order to empirically provide a proof-of-concept

$$\beta \xrightarrow{c_1} \alpha, \tag{16}$$

$$\alpha + \beta \xrightarrow{c_2} 2\beta. \tag{17}$$

In the context of our metapopulation model, α species represent normal species (healthy individuals) and β species represent active species (infected individuals) intermingling from area to area. This scheme is also known as the SIS model (Susceptible-Infectious-Susceptible) [31]. Recall that when we analyze this model without considering compartments and diffusion, the stoichiometric matrix is

$$S = \begin{array}{c} \\ \alpha \\ \beta \end{array} \begin{array}{c} \begin{array}{cc|cc} \alpha & \beta & \alpha & \beta \end{array} \\ \left(\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{array}\right) \end{array}$$

in which the first and second row of the matrix stand for reactions (16) and (17) respectively. The two left columns contain the coefficients of the reactants while the other two right columns contain the coefficients of the products. Meaning that from $S = (s_{ij})$, the components of the state change vector associated to a reaction R_k, $k = 1, \ldots, M$, are effectively

$$(s_{k,N+1:2N} - s_{k,1:N}) \tag{18}$$

where in this example $N = 2$ and $M = 2$. It is less error prone to set S and programmatically proceed as shown above.

We now consider two neighboring areas (i.e., compartments) where both the healthy and infected individuals can move back and forth. We label the areas as Area 1 and Area 2. Consequently, the scheme can be represented by a reaction-diffusion process where reactions (16) and (17) can take place inside both areas separately and the individuals can move back

and forth between Area 1 and Area 2. An illustration is shown in Fig. 4. (Imagine Area 1 as indoor and Area 2 as outdoor.)

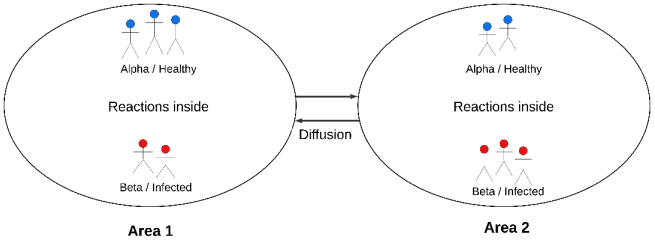


Fig. 4. Illustration of the metapopulation example.

The process is a random process and can be modeled by the RDME consisting of two reactions with two compartments. By our construction, species change forms without creating or destroying themselves, hence the overall population count in the process stays constant.

To make the details clear, we take a small example where we assume that initially, Area 1 has $\alpha = 2$ healthy individuals and $\beta = 1$ infected individual. Area 2 has $\alpha = 1$ healthy individual and $\beta = 2$ infected individuals. With the convention of section II-F, the initial state vector is therefore

$$[\overbrace{\underbrace{2}_{\text{Area 1}},\underbrace{1}_{\text{Area 2}}}^{\alpha},\overbrace{\underbrace{1}_{\text{Area 1}},\underbrace{2}_{\text{Area 2}}}^{\beta}] \tag{19}$$

with diacritical annotations in (19) meant to ease the interpretation of the indexing of the components. We set the reaction parameters $c_1 = .30$ and $c_2 = 1.0$, and the diffusion parameters $D = 1.0$ and $l = 10$, yielding $d = D/l^2 = 10^{-2}$. The RDME gives us all the possible states we can get when those individuals react and diffuse randomly, and recall that along a similar spirit to [32, Eq. (4)], diffusion in our context is modeled via stochastic jump processes through additional reactions,

$$(\text{Area 1})\ \alpha \underset{d}{\overset{d}{\rightleftharpoons}} \alpha\ (\text{Area 2}), \tag{20}$$

$$(\text{Area 1})\ \beta \underset{d}{\overset{d}{\rightleftharpoons}} \beta\ (\text{Area 2}). \tag{21}$$

Now that we are considering the system in two areas where diffusion also occurs, the stoichiometric matrix grows much larger to take the form below.

$$S = \begin{array}{c} \\ \textbf{Area\#} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad}^{\textbf{Reactants}} \qquad \overbrace{\qquad\qquad\qquad\qquad}^{\textbf{Products}} \\ \overbrace{\quad}^{\alpha}\ \overbrace{\quad}^{\beta}\ \overbrace{\quad}^{\alpha}\ \overbrace{\quad}^{\beta} \\ \textbf{1}\ \ \textbf{2}\ \ \textbf{1}\ \ \textbf{2}\ \ \textbf{1}\ \ \textbf{2}\ \ \textbf{1}\ \ \textbf{2} \\ \hline \\ \\ \\ \\ \hline \\ \\ \\ \\ \end{array} \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 2 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{array}{l} \\ \\ \\ \\ \\ \textbf{Reactions} \\ \\ \\ \\ \textbf{Diffusions} \end{array}$$

The annotations here are also aimed at easing the connection of how to map the location of entries with the convention of section II-F. In the matrix, the first two rows indicate reactions (16) and (17) inside Area 1. The third and fourth rows represent the same reactions respectively in Area 2. Diffusions (20) and (21) are characterized by the last four rows. The four left columns contain the coefficients of the reactants and the other four right columns contain the coefficients of the products. Using (18) with $N = 4$ and $M = 8$ gives the state change vectors.

Even starting with a simple initial state as $[2, 1, 1, 2]$ explained in (19), the system explodes to a total of 84 possible states (and quickly much more if we set larger starting values as we alerted before). We solve the associated RDME using the classical FSP equation (9) with a state space truncated to 70 to continue our empirical illustration of the computational building blocks. Taking $t = 10$, we obtain an approximation of the marginal probability distribution of the healthy and infected individuals in the two areas. Results are shown in Fig. 5.

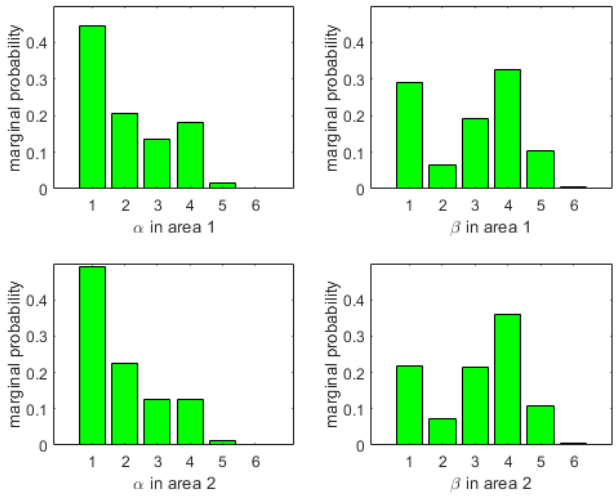


Fig. 5. Marginal probability distribution of the number of healthy and infected individuals in the two areas using the classical FSP at time $t = 10$.

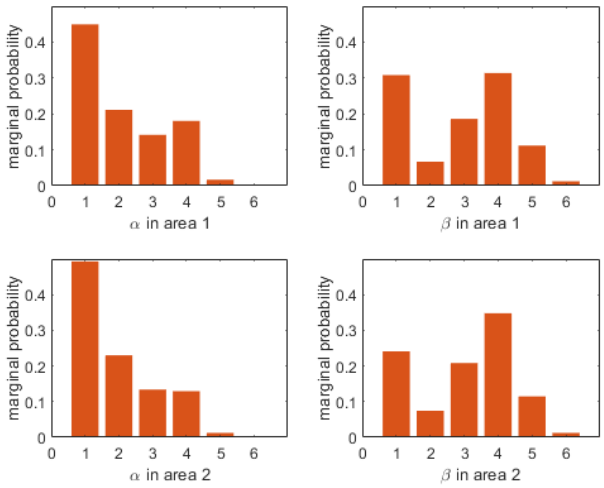


Fig. 6. Marginal probability distribution of the number of healthy and infected individuals in the two areas using tensors at time $t = 10$.

Having obtained a reference solution via the classical FSP, we now attempt to reproduce results using tensors on the same metapopulation model. As explained in the text, the main concept is to express a high-dimensional tensor as an ordered product of smaller tensors that can scale better as problems become more complex than the simple example shown here for illustration and ease of interpretation. For our purposes, the marginal probability distribution of the individuals are found by using the adaptive tensor code in [25], which has already been demonstrated to be effective on much larger problems albeit in the CME context. Fig. 6 shows the computed marginal probability distribution of the healthy and infected individuals in the two areas using the tensor technique in our RDME prototype. Fig. 7 illustrates the accuracy of the implementation of tensors by showing the near alignment of the marginal distributions obtained using the two alternative approaches.

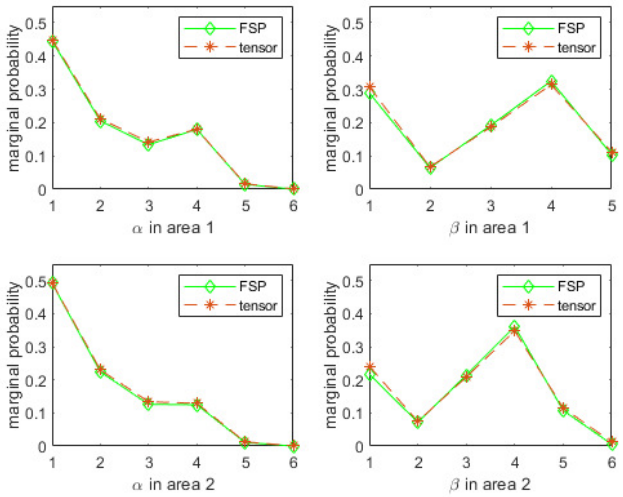


Fig. 7. Comparison of the marginal probability distributions computed using the classical FSP vs tensors.

IV. Conclusions

We have applied the RDME formulation in a metapopulation model that includes both reaction and diffusion processes. Although our initialization was kept simple to detail the algorithmic steps, it is perhaps worth mentioning that the broader significance of the metapopulation model with reaction-diffusion is to represent the interplay between local population dynamics and migration. It is a way of understanding how populations behave when they are distributed across discrete, interconnected areas and subject to both local environmental factors (birth/death rates) and movement between these areas (diffusion). This model is of value in the literature because it provides insights into how populations persist, grow, or face extinction in fragmented landscapes. The model is essential for addressing real-world issues such as conservation, habitat fragmentation, disease spread, and the management of invasive species, where the dynamics of movement and local interactions play key roles in shaping long-term outcomes.

Therefore, the fact that we have an RDME formulation opens avenues for further computerized simulations. Here, we have utilized the FSP and tensors to approximate the marginal probability distribution. Rather than taking into account all possible states, the FSP operates with a reduced state space. On top of that, the compact representation of the transition matrix via tensors goes hand in hand with tensor train decompositions. Our body of research has shown that tensors scale very well to tackle multi-dimensional CME problems, suggesting promising prospects for their application in the RDME as we illustrated in this case study. Although the tensor approach introduces a significant complexity in the derivation of algorithms, it offers distinct scalability advantages. It excels in

managing high-dimensional problems by efficiently representing extensive state spaces with reduced memory requirements, making it particularly effective in scenarios where the classical matrix-based FSP becomes computationally expensive or infeasible.

REFERENCES

[1] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*. Springer, 2002, pp. 447–460, doi: 10.1007/3-540-47969-4_30.

[2] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, "Parcube: Sparse parallelizable tensor decompositions," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I 23*. Springer, 2012, pp. 521–536, doi: 10.1007/978-3-642-33460-3_39.

[3] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927. doi: 10.1002/sapm192761164

[4] L. Grasedyck, D. Kressner, and C. Tobler, "A literature survey of low-rank tensor approximation techniques," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013. doi: 10.1002/gamm.201310004

[5] W. Hackbusch, *Tensor spaces and numerical tensor calculus*. Springer, 2012, vol. 42, doi: 10.1007/978-3-030-35554-8.

[6] J. Grabek and B. Cyganek, "An impact of tensor-based data compression methods on deep neural network accuracy," in *Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 26. PTI, 2021. doi: 10.15439/2021F127 p. 3–11. [Online]. Available: http://dx.doi.org/10.15439/2021F127

[7] D. T. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A: Statistical Mechanics and its Applications*, vol. 188, no. 1-3, pp. 404–425, 1992. doi: 10.1016/0378-4371(92)90283-V

[8] ——, "Exact stochastic simulation of coupled chemical reactions," *The journal of physical chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977. doi: 10.1021/j100540a008

[9] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *The Journal of Chemical Physics*, vol. 124, no. 4, p. 044104, 2006. doi: 10.1063/1.2145882

[10] J. S. van Zon, M. J. Morelli, S. Tănase-Nicola, and P. R. ten Wolde, "Diffusion of transcription factors can drastically enhance the noise in gene expression," *Biophysical Journal*, vol. 91, no. 12, pp. 4350–4367, 2006. doi: 10.1529/biophysj.106.086157

[11] N. Pavin, H. Č. Paljetak, and V. Krstić, "Min-protein oscillations in Escherichia coli with spontaneous formation of two-stranded filaments in a three-dimensional stochastic reaction-diffusion model," *Physical Review E*, vol. 73, no. 2, p. 021904, 2006. doi: 10.1103/PhysRevE.73.021904

[12] M. Smoluchowski, "Mathematical theory of the kinetics of the coagulation of colloidal solutions," *Z. Phys. Chem.*, vol. 92, pp. 129–168, 1917.

[13] C. Gardiner, K. McNeil, D. Walls, and I. Matheson, "Correlations in stochastic theories of chemical reactions," *Journal of Statistical Physics*, vol. 14, pp. 307–331, 1976. doi: 10.1007/BF01030197

[14] A. B. Stundzia and C. J. Lumsden, "Stochastic simulation of coupled reaction–diffusion processes," *Journal of Computational Physics*, vol. 127, no. 1, pp. 196–207, 1996. doi: 10.1006/jcph.1996.0168

[15] M. M. Rahman and R. B. Sidje, "A study of a metapopulation model using the stochastic reaction diffusion master equation," in *Southwest Data Science Conference*. Springer, 2023, pp. 242–253, doi: 10.1007/978-3-031-61816-1_17.

[16] M. M. Rahman, K. Cook, and R. B. Sidje, "Parallel implementation of the stochastic reaction diffusion master equation," in *Proceedings of the 17th International Conference on Bioinformatics and Computational Biology*, T. Aldwairi, H. Al-Mubaid, and O. Eulenstein, Eds. Cham: Springer, 2025. ISBN 978-3-031-94039-2 Doi: 10.1007/978-3-031-94039-2_10.

[17] M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," *The journal of physical chemistry A*, vol. 104, no. 9, pp. 1876–1889, 2000. doi: 10.1021/jp993732q

[18] J. Hattne, D. Fange, and J. Elf, "Stochastic reaction-diffusion simulation with MesoRD," *Bioinformatics*, vol. 21, no. 12, pp. 2923–2924, 2005. doi: 10.1093/bioinformatics/bti431

[19] J. Elf and M. Ehrenberg, "Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases," *Systems biology*, vol. 1, no. 2, pp. 230–236, 2004. doi: 10.1049/sb:20045021

[20] M. M. Rahman, "A numerical study of the stochastic reaction-diffusion master equation using tensors and parallelism with application to biological models," Ph.D. dissertation, The University of Alabama, 2025.

[21] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000. doi: 10.1137/S089547989630569

[22] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009. doi: 10.1137/07070111X

[23] T. G. Kolda, "Multilinear operators for higher-order decompositions," Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA (United States), Tech. Rep., 2006, doi: 10.2172/923081.

[24] I. V. Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011. doi: 10.1137/090752286

[25] T. Dinh and R. B. Sidje, "An adaptive solution to the chemical master equation using quantized tensor trains with sliding windows," *Physical Biology*, vol. 17, no. 6, p. 065014, 2020. doi: 10.1088/1478-3975/aba1d2

[26] H. D. Vo and R. B. Sidje, "An adaptive solution to the chemical master equation using tensors," *The Journal of Chemical Physics*, vol. 147, no. 4, 2017. doi: 10.1063/1.4994917

[27] K. Burrage, M. Hegland, S. Macnamara, and R. B. Sidje, "A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems," in *Proc. of The AA Markov 150th Anniversary Meeting*, no. 21-37, 2006.

[28] M. Hegland and J. Garcke, "On the numerical solution of the chemical master equation with sums of rank one tensors," *ANZIAM Journal*, vol. 52, pp. C628–C643, 2010. doi: 10.21914/anziamj.v52i0.3895

[29] V. Kazeev, M. Khammash, M. Nip, and C. Schwab, "Direct solution of the chemical master equation using quantized tensor trains," *PLoS computational biology*, vol. 10, no. 3, p. e1003359, 2014. doi: 10.1371/journal.pcbi.1003359

[30] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics*, vol. 3, no. 4, pp. 276–282, 2007. doi: 10.1038/nphys560

[31] R. Anderson, R. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992, doi: 10.1093/oso/9780198545996.001.0001.

[32] S. Smith and R. Grima, "Breakdown of the reaction-diffusion master equation with nonelementary rates," *Phys. Rev. E*, vol. 93, p. 052135, May 2016. doi: 10.1103/PhysRevE.93.052135. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.93.052135

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 595–603
DOI: 10.15439/2025F9798
ISSN 2300-5963 ACSIS, Vol. 43

Leveraging Business-IT Collaboration: Effective Design and Challenges of the IT Business Partner Role

Constanze Riedinger, Christopher Rentrop
0009-0003-0226-4114, 0000-0003-3629-3821
kips, Konstanz University of Applied Sciences,
78467 Konstanz, Germany
Email: {constanze.riedinger, rentrop}@htwg-konstanz.de

Melanie Huber
0000-0001-8020-9055
BITCO³ GmbH, 78467 Konstanz, Germany
Email: melanie.huber@bitco3.com

Stephan Zimmermann
0009-0000-1977-6659
Technical University of Applied Sciences Augsburg,
86161 Augsburg, Germany
Email: stephan.zimmermann@tha.de

***Abstract*—The IT Business Partner (ITBP) role is increasingly adopted to bridge the gap between business and IT and foster strategic alignment. Even though organizations still struggle with implementing an effective collaboration between business and IT, the design of this role and the challenges it faces remain underexplored. This study is the first scientific paper to examine the ITBP. Through a case study and expert interviews, it derives eleven key tasks, with relationship management and advice at its core, and 14 required skills for effectiveness. It identifies challenges of the ITBP role, such as organizational integration, skill gaps and a lack of strategic focus. This study contributes to the field by offering a comprehensive understanding of the ITBP's design and role for business IT alignment. It provides organizations with guidance on the implementation and skill development of the ITBP role. Future research should validate findings and further explore the role's effectiveness.**

***Index Terms*— IT business partner, business relationship manager, IT role design, tasks, skills, challenges, business-IT alignment.**

I. Introduction

IN TIMES of digital transformation, it becomes a competitive advantage to work in an integrated way and align the business and information technology (IT) mindsets across the organization [1]. However, a survey reveals that only 30% of the 114 responding managers perceive their collaboration between business and IT units as adequate [2]. The majority struggles with IT optimization decisions and highlights a poor and complex collaboration process. To improve collaboration between business and IT, companies implement interface roles like IT Business Partner (ITBP), Business Analyst, and similar positions. In 2018, Gartner Inc. introduces such a role highlighting its importance for alignment and value creation [3].

Literature recognizes the importance of these interface functions for business-IT alignment (BITA) and to establish relational mechanisms [4], [5]. IT management frameworks such as Information Technology Infrastructure Library (ITIL) and Technology Business Management (TBM) mention a Business Relationship Manager (BRM) as a relevant role to build a strong relationship between the IT organization and the business units [6], [7]. Even though the role becomes increasingly important, and organizations still struggle to implement it effectively, to our knowledge, there is no academic, peer-reviewed in-depth study that examines this role at the interface between business and IT and its contribution to bringing business and IT closer together. The objective of this paper is, therefore, to provide a detailed understanding of the ITBP role, including its role design, key tasks, required skills, and challenges. To address this, we pose two research questions (RQ): (1) *How can organizations design the interface role of the IT Business Partner for effective business-IT collaboration?* and (2) *What practical challenges does the IT Business Partner face in organizations?* By answering these questions, we expand the theoretical knowledge about the interface role and its impact on business-IT alignment. Additionally, we provide practical support in designing the role and selecting individuals with the relevant skills to fulfil it.

The article is structured as follows: the second section examines the theoretical background on collaboration between business and IT and the ITBP role. We then outline our methodological approach triangulating a case study and expert interviews for an in-depth, practical study. Section four presents the findings, followed by a discussion of the results in the context of successful implementation within organizations. Finally, we conclude with the study's limitations and offer an outlook for future research.

II. Background And Related Work

Effective collaboration between the business and the IT department is the basis for achieving added value through IT [4]. Thus, we first introduce research on this collaboration. Then, we examine roles at the interface and the ITBP role with its "mission to increase business value" [3].

A. Business-IT Collaboration

The collaboration between business and IT has been a central focus of research for decades, as organizations strive to bridge the gap between strategic goals and technological implementation by aligning the two disciplines [1], [8]. A foundational framework for understanding this relationship is the Strategic Alignment Model (SAM) introduced by [9]: it describes alignment as the dynamic interplay between strategic and operational levels and the functional integration of business and IT domains. Notably, alignment is not a static state but an ongoing process requiring continuous communication, mutual understanding, and shared objectives [10]. A higher degree of maturity in this alignment directly contributes to improved organizational performance, innovative ability, and competitive advantage [8]. According to Luftman's maturity model [11], strong Business-IT partnerships and effective communication are enablers of alignment. Fostering such a partnership requires, among other things, building trust through a shared understanding of the value of IT and shared goals and risks. The *COBIT 2019* framework also highlights that one IT Governance and Management objective for organizations are managed relationships (APO08) in order to create improved outcomes, mutual trust and effective use of resources [12]. While a former study shows that in traditional organizations roles from business and IT at different hierarchical levels (including top-, middle-, lower-level management and non-management level) are involved in strategic alignment [13], COBIT mentions the necessity to implement a role that manages these business-IT relations. An interface role, such as the ITBP, may foster this partnership acting as a bridge between business and IT departments, ensuring that both sides understand each other's needs and priorities [5].

B. Roles at the interface between Business and IT

The critical role of fostering collaboration between business and IT, driving innovation, and achieving business goals through technology solutions may be referred to as an "IT business partner" or "IT business relationship manager" [14], [15], [16]. Reference [17] call this role a "boundary spanner": the role acts as a translator, mediator and aligner between business and IT, as well as a motivator for both sides and a representative of IT to business and vice versa. Reference [18] describes how organizations structure this role in IT business partner organizations to foster business-IT collaboration and communication. Reference [16] emphasizes the overlap with other roles and gives an overview about three main tasks of the ITBP: close collaboration, network maintenance and continuous integration of IT into the business units While, to our knowledge, except for two German articles [16], [18], academic peer-reviewed literature does not outline the role of the ITBP, several IT management frameworks mention its necessity for organizations: TBM introduces a BRM acting like an account manager responsible for the business, understanding their plans and proposing solutions. They should be able to communicate the business value of IT and therefore, require knowledge about the portfolio, service costs, and business consumption. To that end, they work closely together with other IT roles such as service owners or EAs [7]. ITIL mentions, besides fostering collaboration, main tasks of an ITBP such as demand management, stakeholder management and support in service management when it comes to service level agreements (SLAs) and the strategic optimization of IT services [6]. The *Skill Framework for the Information Age* (SFIA) states a business analysis role with a more operative focus responsible for investigating operational issues and finding new business solutions [19]. Other frameworks such as COBIT 2019, the scaled agile framework (SAFe) and The Open Group Architecture Framework (TOGAF) do not explicitly mention the ITBP role. Yet, they stress the importance of business-IT alignment, stakeholder engagement and value delivery [20] as well as relationship management, coordination, and communication with the business [12]. While there are various terms for describing the interface role, we use ITBP in this paper while encompassing relevant literature for both.

The design of interface roles between business and IT varies from company to company. ITBPs are typically positioned at the middle management level but can also hold higher positions such as directors [14]. Existing literature shows that interface roles can be part of the IT department [21] or they can also be located in dedicated digitalization or innovation departments or directly in the business units [22]. The role of the ITBP can be performed as a primary or secondary occupation. If it is the primary occupation, the employees are often organizationally grouped into an IT Business Partner Organization (ITBPO) [18], [23]. If the function of an ITBP is performed as a secondary occupation, ITBP tasks are embedded into other roles. For example, architects or demand managers often act as ITBPs. Architects are considered a critical function for successful BITA, as they shape an organization's IT landscape [24]. Thus, [24] mentions business or domain architects, who primarily represent individual functions or processes in this landscape. Demand managers can also improve the relationship between business and IT by implementing strategic planning and portfolio management [25], thus overlapping with classic ITBP roles. Related to the collaboration between business and IT the SFIA names stakeholder relationship management as a relevant competency in the information age. It describes the skill to "systematically analyze, manage and influence stakeholder relationships to achieve mutually beneficial outcomes through structured engagement" [19]. However, the mentioned literature neither give insights into the design and implementation of the ITBP role, nor about the challenges ITBPs face in their daily business.

III. Methodology

As the design of roles in organizations is influenced by the context, earlier research on interface roles between business and IT applied qualitative methods such as expert interviews in an exploratory study [17] or case study research [26]. We answer our research questions by applying a triangulation of these methods and data from various sources, which enriches

insights and the validity of findings [27]. We triangulate a parallel approach of one case study and expert interviews. The case study enables research of a phenomenon in its real-world context [28] such as to explore ITBP implementation in-depth in a specific organizational context. Expert interviews lead to a comprehensive picture of the phenomenon in practice [29].

A. Case Study

We follow case study research guidelines to design our study and to select our case [28], [30]. Our case was an international consumer goods company with a size of about 30.000 employees, a central IT department and a dedicated IT governance department. The central IT department provides services for the twelve international business divisions (i.e. legal entities differentiated by product line and region) and other central functions (e.g., HR, procurement or finance). We chose this single-case design for two reasons: First, our company reflects a common case [28], as it is representative of large, globally active production companies with central IT structures and therefore has a high transferability to similar contexts. Second, through an existing and trusting relationship between the researchers and the company, the case study could benefit from access to internal information as well as open and honest discussions with the participants.

For the alignment between business and IT, the company established an ITBPO within the IT department. In order to study the ITBP role in depth and enhance the insights, we focused on multiple units of analysis [28], as the ITBPO consists of one leading ITBP and five ITBPs at different locations. Each of the five ITBPs is responsible for partnering one or more business divisions. Furthermore, the organization employs two EAs to strategically plan and align the enterprise target landscape with the business. For the delimitation of the ITBP role with its tasks and required skills, we included a review of the EA role in our study. We could accompany the case for 1.5 years from June 2023 to December 2024.

For the data collection, different sources of evidence, mainly documents, interviews, and focus groups are used [28] (see Fig 1). We reviewed the overall organizational structure and IT operating model to set up the case. The role descriptions served as the basis for the analysis of the ITBP role. We observed meetings between ITBPs and their business counterparts, to understand current interaction and derive responsibilities and skills. We further conducted note-based semi-structured interviews with the ITBPs and EAs to understand their perception of their tasks based on their specific challenges within their individual location. The workshop with the stakeholders aligned different views, the understanding of their role, responsibilities and skills and challenges throughout the whole organization.

B. Expert Interviews

We carried out seven expert interviews to broaden the insights on the ITBP role in different organizational contexts. The interviews were semi-structured to have the opportunity to ask more detailed follow-up questions during the conversation [29], [31]. We chose interviewees considering the following criteria: First, the interviewees needed to have specific knowledge about the ITBP role due to performing it, being responsible for an ITBPO or ITBPs in an organization or having experienced it from an external perspective. Second, they needed to be able to consider the overall perspective of the role in terms of practical implementation within organizations. Therefore, we chose ITBPs with several years of experience and managers that were leading ITBPs. Third, we wanted to cover a broad spectrum of organizational contexts in the interviews to get as complete a perspective on the ITBP role as possible. Therefore, the interviewees have various backgrounds coming from different industries, organizational sizes, and digital maturities. All interviewees were already part of the network of the research team. Table I shows an overview of all interviewees (I) (chronological listing based on the date of the interview). The guiding questions that were used during the interviews are:

1. Why do you implement an interface role between business and IT?
2. How do you integrate the role in the company?
3. What are the main tasks of the ITBP? Are other tasks performed additionally?
4. What are the skills required to be an ITBP?
5. What challenges do you observe for an ITBP?

TABLE I. OVERVIEW INTERVIEWEES AND GUIDING QUESTIONS

ID	Function	Industry	# of Employees (EMP)
I1	CIO	Equipment Manufacturing	~3.500 EMP
I2	CIO	Transportation and Logistics	~6.500 EMP
I3	IT Business Partner	Consumer Goods	~1.500 EMP
I4	CEO	IT Consultancy	50 EMP
I5	TBM Consultant	Free Lancer in TBM Consulting	-
I6	IT Business Partner	Medical Technology	~9.400 EMP
I7	IT Business Partner	Retail	~99.000 EMP

The interviews were performed between October 2024 and February 2025 by the first author via Microsoft Teams in German or English. For data collection, they were either recorded and transcribed during the sessions using Microsoft Copilot or notes were taken. After the interview phase, we proofread and anonymized the automatic transcripts of all interviews and summarized notes.

C. Data Analysis

The triangulation of qualitative methods with parallel set up and data collection in case study and expert interview was then followed by a data analysis. This overall methodology is displayed in Fig 1. For our overall data analysis, we conducted qualitative coding [32]. This entailed the assignment of respective codes to the data in three iterations, in accordance with the guidelines for open and selective coding. The first coding was conducted by the first author. In two further iterations, the research team, comprising four researchers, undertook a review and adjustment of the codes to derive tasks

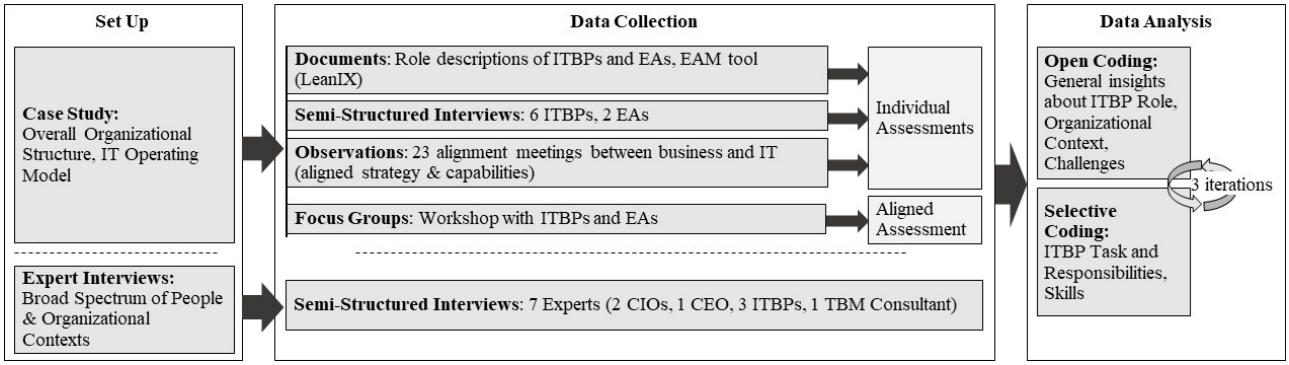


Fig 1. Methodology for Set Up, Data Collection and Data Analysis

and responsibilities as well as required skills for the role. A digital whiteboard tool was used to visualize and structure the coding process. This enabled collaborative and iterative coding of the interview data, whereby codes, categories and correlations were visually recorded, grouped and further developed. Furthermore, we identified and analyzed the challenges of the ITBP role. Table II indicates examples of the qualitative coding results.

TABLE II. EXAMPLES OF QUALITATIVE CODING

Source	Citation or Notes	Qual. Analysis (Category > Codes >> Subcodes)
I4	"Understanding depends on personal factors. People often talking about one another rather than with one another"	Challenge > Human factor >> lack of open communication >> personal factor Skills > Networking & Interpersonal Skills >> Build relationship on personal level
I6	"I give input and ideas from IT side to see how we can help and support for a given problem and how to approach the problem."	Task > Advice and Guidance Skill > Requirements Analysis, Analytical thinking
I7	"We have too many tasks for the IT business partners. And if you have too many roles at once, you no longer have the capacity."	Challenge > role's definition >> lack of task delimitation

IV. FINDINGS

The qualitative study shows that the purpose of an ITBP is always related to the alignment of business and IT. The following section presents eleven tasks with associated responsibilities and 14 skills divided into soft and hard skills to describe the role of the ITBP. Then, we outline five challenges ITBPs are facing.

A. The role of the IT Business Partner

The design of the interface role of the ITBP is "almost like a belief – you can put it in business, you can put it in IT, or you can even put it somewhere in between as a separate function, so there are different models" (I1). Our study shows different characteristics of the ITBP role in different organizational contexts. In Table III we indicate their title, their organizational integration, and their purpose as basis for answering our first RQ: (1) *How can organizations design the interface role of the IT Business Partner for effective business-IT collaboration?*

TABLE III. CHARACTERISTICS OF IT BUSINESS PARTNERS IN DIFFERENT ORGANIZATIONAL SETTINGS

Source	Name	Organizational Integration	Purpose of the role
Case Study	IT Business Partner Manager	6 full time ITBP in the IT unit divided by business divisions with a regional focus	"Develop and maintain collaborative relationships between IT and key business stakeholders to successfully establish and maintain strategic & operational alignment"
I1	Business Engager	Planned ITBP Role in the IT unit divided by business functions	Engage with counterparts ("PMO Leads") in business focusing on alignment and bringing in the IT knowledge and process perspective
I2	Business Systems Function	3 full time ITBP in a separate digital unit in the business divided by business functions	Work closely with business counterparts within the business functions to steer demands, develop requirements and execute IT projects in collaboration with IT experts
I3	Business Analyst/ Partner	3 full time ITBP in the IT unit divided by business functions	Currently, Business Analyst with focus on demands and representation of business domain. Future goal to become a strategic partner with decision rights to support shaping the domain and consulting them strategically
I4, I5	Business Relationship Manager	*I4 and I5 share experience from working in consultancy in different settings.*	"In general, ITBPs are integrators between business and IT" (I4) with a main purpose to "conduct value conversations to drive partnership and IT business value" (I5)
I6	IT Business Partner	10 part time ITBPs in the IT unit divided by business functions and regions	First contact point for business in IT aiming to understand business issues, and propose and coordinate solutions to drive value of IT
I7	IT Business Partner	10 full time ITBPs in the IT unit divided by business functions	"The two main tasks are representing both the business and IT while identifying challenges and finding solutions."

Table IV presents the tasks of ITBPs derived from our triangulation, structured by tasks and the mentioned responsibilities ordered by the number of mentions. The execution of tasks and the individual responsibilities vary depending on their specific job description. However, some tasks are always

mentioned as core activities of the ITBPs, which are **relationship management** and **advice and guidance**. In addition, most ITBPs also take on (strategic) demand management. Further responsibilities were partially identified, either as part of the formal role description or as responsibilities that were overtaken by the ITBP due to a lack of clear delineation from other areas. For instance, I6 with a dual role oversees IT demand, IT project, and IT portfolio management, besides the role as ITBP. The ITBP I7 has a full-time role as ITBP but due to special expertise also supports innovation and organizational change management. The ITBPs within the case study reported engagement in IT project and IT service management activities, in addition to providing support in enterprise architecture management (EAM).

TABLE IV. TASKS AND RESPONSIBILITIES OF IT BUSINESS PARTNERS

Task	Responsibilities	CS	I1	I2	I3	I4	I5	I6	I7
Relationship Management	• Primary business contact and relationship development • Trust building, communication and continuous relationship improvement • Stakeholder engagement, feedback and personalized involvement • Strategic collaboration, integration and alignment of business and IT	x	x	x	x	x	x	x	x
Advice and Guidance	• IT input, ideation, and innovation advisory • Proactive consultancy and orchestration of IT ideas, demands and projects • Value conversations and strategic guidance	x	x	x	x	x	x	x	x
(Strategic) Demand Management	• Idea-to-demand support and derivation, synergies identification • Demand preparation and management • (Initial) Requirements analysis	x	x	x	x			x	x
IT Project Management	• Facilitation of technical solutions, support project implementations • Proof of concept support for technical solutions • Strategic Project Management and Implementation Coordination	x		x	x			x	x
IT Portfolio Management	• Strategic domain planning, resource identification • IT project portfolio accountability	x		x	x			x	x
Enterprise Architecture Management	• Solution architecture coordination • IT landscape modeling and target planning support	x			x			x	x
IT Cost Management	• Cost savings identification and efficiency improvement • Budget management	x				x	x		
IT Service Management	• Major incident management • SLA definition and monitoring for IT services	x					x		
IT Governance, Risk and Compliance	• IT governance and audit compliance assurance • Education and training on IT policies and procedures	x							x
Org. Change Management	• Organizational change management support								x
Innovation Management	• Ideation and business innovation enablement								x

To perform these tasks, ITBPs require skills that are crucial for success. Table V describes these skills, however, "above all, IT Business Partners must be generalists. They should have knowledge across various domains but do not need to master them in detail" (I7).

TABLE V. SKILL MAP OF AN IT BUSINESS PARTNER

	Skill	Ability to…
Soft Skills	Discussion Skills	…conduct eye-level discussions with business leaders (CS,I3,I7), advise and consult on IT solutions and business processes (CS,I1,I6,I7), and use mediation skills (CS,I7).
	Leadership Skills	…think strategically and conceptually (CS,I7), support and make informed decisions (I3), and lead people effectively (CS,I6,I7).
	Communication Skills	…communicate, present, and interact with stakeholders on all hierarchical levels and different languages (CS,I6,I5), speak a "common language" with the business (I4,I7), and foster business-IT talkability (I1,I2,I3).
	Team Understanding	…understand team structures and individual roles within teams (I7,CS), identify the relevant stakeholders for strategic conversations (CS), comprehend team dynamics (I7), and establish understanding for each other (cultivate an exchange culture) (I2).
	Networking & Interpersonal Skills	…build relationships (also on a personal level) (CS,I7), become a trusted partner (CS), and bring people together (CS,I6).
	Analytical Thinking	…collect and analyze the relevant data (I2,I3), develop informed solutions to solve problems (CS), and pay attention to details to prevent errors and inconsistencies (CS).
	Flexibility	…adapt to change (CS), react resiliently (CS,I7), and operate creatively within dynamic environments (I3,I7).
Hard Skills	Business Acumen	…understand the business model and strategy (CS), understand business capabilities (CS,I3,I6,I7), and comprehend business processes in detail as well as in their broader context (I3,I6,I7).
	IT & Data Literacy	…understand and maintain consistent data and data relationships (I3,I2), understand technologies and digital solutions (CS,I1,I6,I7), and use IT tools (for project management, architecture, etc.) (CS,I7).
	Requirements Analysis	…understand and translate requirements of business and IT stakeholders (I3), propose adequate IT solutions (I7), and leverage technology and its manyfold use (I3).
	Project Management	…plan and manage projects (CS,I7), identify risks and manage major incidents (CS), and track and report on progress (CS,I6).
	Strategic Management	…comprehend strategic perspectives and demands (CS, I7), develop long-term goals and objectives (CS), and align projects with overarching strategic directions to develop a portfolio (CS,I6).
	Governance Expertise	… understand organizational decision structures and service level agreements (CS), comprehend IT

		security, legal compliance and risk regulations (including data protection) (I7), continuously improve through critical success factors and key performance indicators (CS,I4,I5), and support business value realization (CS,I4,I5).
	Enterprise Architecture Knowledge	…understand (IT) organization and (IT) processes (I1,I7), know current IT applications and the IT landscape (I7,CS), and comprehend the "big picture" and overarching correlations (I6, I7).

B. Challenges of an IT Business Partner

We identified several challenges the ITBP faces that impact its effectiveness and integration within organizations. We thereby answer our RQ *(2) What practical challenges does the IT Business Partner face in organizations?* One key issue is the **human factor**, which is reflected in the lack of open, two-way communication across hierarchical levels. Instead of a meaningful dialogue, interactions are often characterized by "talking about one another rather than with one another" (I4). Furthermore, understanding the role often depends on individual perspectives and personal factors, leading to inconsistencies. While in some cases the ITBP facilitates excellent collaboration, in others he struggles to achieve alignment or discussions on eye-level (CS, I3, I6, I7).

Another significant challenge is the **role's definition and integration within the organization**. Despite the need for a clear understanding of the ITBP's responsibilities, its boundaries with other roles, such as EA or domain architects, project or service managers, are often blurred (CS, I2, I3, I7). This lack of clarity complicates collaboration and role alignment with business stakeholders and IT colleagues. But even if "the role is described clearly on paper, this does not lead to successful integration and acceptance automatically" (I7). Organizations face a challenge to introduce the role in a structured approach, with clear responsibilities and to distinguish it from similar roles like the product owner (I6) or demand manager (I7). Thus, the role lacks acceptance or even face competitiveness with IT colleagues (I7).

The **lack of methodological approach and organizational structure** is mentioned as another challenge for the ITBP role. The ITBPs work closely with the relevant business units, often making them feel more part of the business than part of IT or of an ITBPO, especially as explicit ITBPOs are often underrepresented or not managed at all. As a result, there is a lack of identity as an ITBP team and a lack of opportunities to share ideas with each other. But even when such an organizational structure is in place and a clear job description exists, it is a challenge to compare results and identify synergies, because "every ITBP approached it in their own way, resulting in ten different approaches" (I7).

A further difficulty lies in the **limited time for focused work and absence of strategic alignment**. ITBPs are frequently consumed by operational tasks and "firefighting" (CS), leaving little room to focus on long-term strategic initiatives (CS, I3). Urgent demands, combined with tight collaboration timeframes, often divert attention from value-driven work (I5, I6). Organizationally, the demand for IT support often exceeds the available capacity, forcing the ITBP to act as a mediator between IT and the business instead of a strategic enabler (CS, I6). The study also showed that discussions between ITBP and the business often focus on projects and progress rather than on strategy or deriving necessary future measures from this.

Finally, **skill gaps of the ITBP** present a major obstacle. There are significant differences in maturity levels and knowledge on both sides, with business often lacking sufficient understanding of IT processes and IT stakeholders failing to adopt lean and strategic approaches to support business needs (CS, I1, I4, I5, I6). Communication skills and the ability of business-IT talkability are, therefore, critical for ITBPs where deficiencies are observed (CS, I2, I3, I4, I5, I6). In addition, ITBPs may lack leadership skills and decision-making competence to strategically support and drive decisions (CS, I3, I7). As a result, the ITBP role is not always effectively implemented, limiting its ability to drive value and organizational alignment.

V. DISCUSSION

The findings do not only show a variety of tasks and required skills of an ITBP but also challenges they face in their daily business. In the following we discuss how the design of the role may support overcoming the identified challenges. We focus on the delimitation of the role and required capacities, the opportunities for organizations to find skilled people to act as ITBP and discuss its contribution to strategic alignment.

A. Determining the Organizational Integration of an IT Business Partner

The in-depth study shows various designs of the ITBP role. One decision that a company needs to take is the organizational embedment of the ITBPs. Depending on the existing structure, an organization needs to determine whether the ITBPs support the business divisionally, functionally, or regionally. I6 stated that there were both functionally assigned ITBPs and those assigned to specific regions. In the case study, ITBPs were assigned to business divisions. Divisional ITBPs are required to know the division-specific product and the complete end-to-end operation. In a regional setting, the location-specific requirements and the cultural aspect may be addressed better. A functional allocation has the advantage of allowing the ITBP to develop domain-specific skills, enabling better business communication and leveraging synergy effects.

Another aspect is to design the ITBP role as a primary or secondary responsibility [14]. The study showed that a primary ITBP role allows for greater focus on core activities but may limit networking opportunities, requiring proactive information gathering with other IT colleagues. When an ITBP performs as a secondary role, it is often assigned to individuals in senior positions with strong networks and close business interactions through their primary task in project or portfolio management (I6). While this facilitates relationship-building, it limits the focus on partnering and strategical guidance unless specific time and capacity are allocated. However, even

full-time ITBPs face challenges in balancing their workload and ensuring a standardized approach of all ITBPs (I7). Establishing an ITBPO, aligning the way of work of all ITBPs, and following a structured methodological approach for communication between business and IT may support a clear positioning of the ITBP and the standardized sharing of information towards business and IT [33], regardless of whether the employment is full-time or part-time.

B. Overcoming the Skill Gap in IT Business Partner Positions

The study highlights that, in addition to hard skills, soft skills are particularly relevant for the relationship-intensive role of ITBP. One of the skills emphasized is communication. This is not only essential for establishing a common understanding but is also a proven prerequisite for Business-IT Alignment [11]. Communication involves both clarity and language proficiency—often English in an international context. However, it is equally important to establish a common ground in communication, frequently referred to as a "common language" [5]. This means that an ITBP must be able to communicate fluently in both business and IT terminology, ensuring a seamless flow of information linking business stakeholders and IT colleagues. Language training might therefore be mandatory for ITBPs, and job rotation could ease the establishment of a common language to the business counterparts.

Beyond communication, other soft skills, such as understanding team dynamics, were particularly evident in the case study. It is crucial to be familiar with the corporate culture, actively engage in it, and navigate hierarchies without being discouraged by them [5], [13]. In its role as a boundary spanner [17], an ITBP must act as a connector, bringing people together and serving as a central point of coordination (I6). This skill is difficult to teach, making it essential to consider interpersonal skills during recruitment or to fill the role internally with individuals who are already well-connected and established within the organization [13].

To address the skill gap between business and IT, job descriptions for ITBPs should clearly outline the necessary competencies. When developing job descriptions, identifying key skills through profile challenges can be beneficial [19]. The skill map outlined in Table V can support organizations in designing an effective ITBP job description. Additionally, companies can use this transparency to assess and develop the skills of current ITBPs, enabling organizations to upskill existing employees in this role. This approach allows ITBPs to "plan their professional development and career progression" [19]. It also contributes to human resource development, career planning and the provision of learning paths, which, as seen in SFIA, can be structured into different levels.

C. Focusing on Core Tasks and Strategic Business-IT Alignment

The findings illustrate that an issue for ITBPs is having to perform a variety of tasks without a clear focus. For an ITBP to fulfil their purpose to relate business and IT and strategically align both perspectives, the study shows that relationship management and advice and guidance should be the core tasks (see Fig 2). As “already for relationship management and advisory together with demand management, an ITBP needs business and IT knowledge and many interactions, there is no time for more without making mistakes” (I7). Therefore, the ITBP requires a clear mandate and capacity for these core activities, to achieve their alignment goals and being accepted by business and IT colleagues.

While demand management is often also covered by the ITBP, organizations should clearly define the transition to project, service, portfolio, or EAM. According to ITIL [6], ITBPs work closely with service owners and service level managers, which emphasizes the importance of defining the interface. TBM highlights further relations to the EAs [7]. Reference [24] mentions in this context domain architects, who support a functional business domain and align strategically with them. During the case study, the distinction in tasks and assigned business stakeholders between ITBPs and EAs was defined: The EAs take on the domain-specific alignment with business functions, whereas the ITBPs focus on alignment with business divisions. This helped to distinguish between the roles. The close interaction of the ITBP with other stakeholders emphasizes the need for internal alignment with IT colleagues to share knowledge and a clear distinction to leverage effective collaboration [33]. Other IT disciplines such as IT cost management, IT governance, organizational

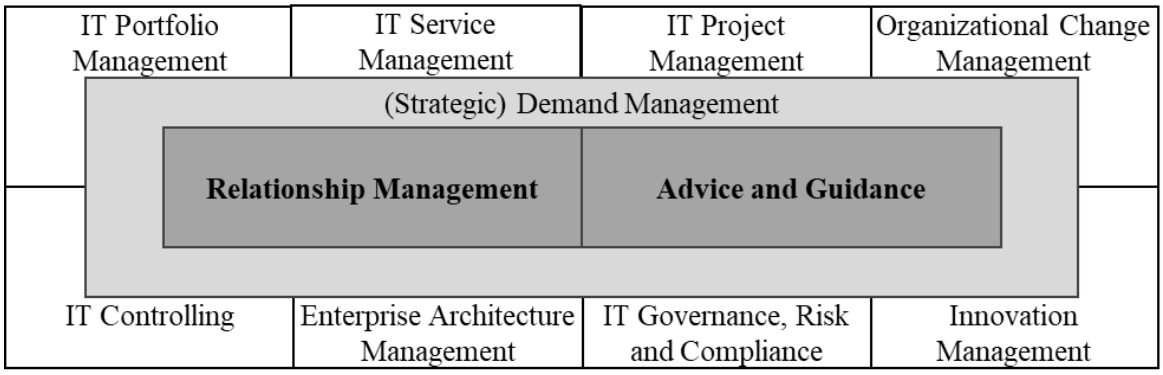


Fig 2. Core Tasks and Further Responsibilities of IT Business Partners

change, and innovation management may be used by the ITBP to gather and feedback information, but decision makers should be careful in delegating active responsibility of the ITBP in these tasks.

Besides the lack of focus, the ITBP, being primarily driven by operational tasks, often tends to neglect strategic work. However, it is essential that the ITBP has sufficient capacity "to make sure that its work contributes to the strategic alignment" (I7). The Strategic Alignment Model (SAM) by [9] provides a suitable framework for illustrating how the ITBP's activities contribute to strategic alignment. According to the SAM, strategic alignment is based on a complex, multidimensional relationship between business and IT. The ITBP plays a crucial role in this process by building and maintaining **relationships**, **advising** on IT matters by proactively consulting stakeholders, and driving value-oriented discussions. Through these core activities, the ITBP triggers further responsibilities within his/her role and can contribute to strategic alignment in multiple ways through exchanges with their business counterparts: They can capture business requirements, align them with **IT governance, risk & compliance** policies and transfer the requirements into operative **demands / projects** (strategy execution). They can receive **strategic demands**, consolidate them, and derive necessary IT capabilities, IT processes, implications for the **enterprise architecture**, and the **IT portfolio** together with IT (technology transformation). They can communicate technological **innovations** and new IT capabilities and identify possible strategic application and necessary **organizational changes** together with the business (competitive potential). They can identify opportunities for the operational use of new technologies in business within **IT services** and discuss **cost saving** potential (service level). Similar considerations are discussed for the ITBPO and their impact on strategic alignment [18]. This contribution to strategic alignment underscores the critical importance of the ITBP role, as business value can only be realized through the effective alignment of IT with strategic objectives [4].

Overall, a role description of the ITBP must be the baseline to define how they operate within the organization [23]. This description first needs to determine the organizational embedment (division, function, or region) and the capacity (primary or secondary). Also, it needs to include tasks and responsibilities (e.g. following the descriptions in Table IV), give a ration about time effort for each task, and list required hard as well as soft skills. A success factor is a clear distinction to other roles and designing internal alignment mechanisms (according to the adjacent disciplines as shown in Figure 2). This role description can be used to challenge existing ITBPs or during recruitment to find suitable candidates.

VI. Conclusion

The role of the ITBP is increasingly being implemented in organizations to bridge the gap between business and IT, thereby driving strategic alignment. Through the triangulation of a case study and expert interviews, this study provides an in-depth exploration of ITBP's role design and challenges. The findings reveal that the role is highly versatile, with responsibilities that vary depending on the organizational context. We identify eleven key tasks that ITBPs perform to different extents, with relationship management and advice and guidance at the core. These core tasks enable ITBPs to drive strategic business-IT alignment. To be effective in this role, 14 essential skills are identified, including communication, networking, business acumen, and IT, data, and process literacy. However, challenges such as organizational integration, skill gaps, and excessive operational workload often limit the ITBP's ability to drive the business-IT partnership and foster strategic alignment and should therefore be part of a clear role description.

This study contributes to both research and practice. From an academic perspective, the study advances research on business-IT alignment by operationalizing the role of the ITBP within organizations and contributing to the broader discourse on IT professions and their interaction with business functions. For organizations, it provides guidance on structuring the ITBP role and defining key responsibilities that support business-IT alignment. For HR professionals, the identified skill sets provide valuable insights for hiring, career development, and training initiatives for ITBPs. In addition, CIOs gain an understanding of the challenges of the role, particularly the need to allocate sufficient capacity for strategic alignment efforts.

Despite its contributions, the study has limitations. It primarily examines the ITBP role from an IT perspective, whereas its success is also influenced by organizational culture and business counterparts. Future research should include both business and IT perspectives to provide a more holistic understanding of the role. Additionally, the chosen case is a large, international company with a centralized IT department. Future studies could examine other organizational setting such as more decentralized structures or small-and-medium-sized companies for comparison and drawing conclusion based on the organizational set-ups. While this qualitative study provides in-depth insights, a follow-up quantitative study could help validate the identified challenges on a broader scale. Further research could also examine the impact of corporate culture and individual business counterparts on ITBP effectiveness, analyze different job descriptions within IT organizations, and differentiate roles based on IT-relevant tasks.

References

[1] C. Riedinger, M. Netscher, and S. Zimmermann, "Organizational Capabilities for Business-IT Integration in Digital Enterprises," in *Proceedings of the 19th FedCSIS*, 2024, http://dx.doi.org/10.15439/2024F9800.

[2] LeanIX GmbH, "LeanIX IT Cost Optimization Report 2023," 2023. [Online]. Available: https://www.leanix.net/en/download/it-cost-optimization-survey-2023

[3] Gartner Inc., "Leveraging the BRM role is all about business value," 2018.

[4] S. De Haes, W. Van Grembergen, A. Joshi, and T. Huygh, *Enterprise Governance of Information Technology*. in Management for Professionals. Cham: Springer International Publishing, 2020.

[5] C. Riedinger, M. Huber, and N. Prinz, "Factors for Effective Communi cation of IT Costs and IT Business Value," in *Proceedings of the 18th FedCSIS*, Sep. 2023, pp. 677–687, http://dx.doi.org/10.15439/2023F7224.

[6] AXELOS Limited, Ed., *ITIL Foundation, ITIL 4 edition*, 1st edition, TSO, 2019.

[7] T. Tucker, "Technology Business Management: The Four Value Conversations CIOs Must Have With Their Businesses," 2016.

[8] Y. E. Chan and B. H. Reich, "IT alignment: what have we learned?," *J Inf Technol*, vol. 22, no. 4, pp. 297–315, 2007, http://dx.doi.org/10.1057/pal-grave.jit.2000109.

[9] J. C. Henderson and H. Venkatraman, "Strategic alignment: Leveraging IT for transforming organizations," *IBM Systems Journal*, vol. 32, no. 1, pp. 472–484, 1993, http://dx.doi.org/10.1147/sj.382.0472.

[10] J. Cybulski and S. Lukaitis, "The impact of communications and under standing on the success of business/IT alignment," presented at the ACIS 2005 Proceedings, Deakin University, 2005.

[11] J. Luftman, "Assessing Business-IT Alignment Maturity," *CAIS*, vol. 4, 2000, http://dx.doi.org/10.17705/1CAIS.00414.

[12] ISACA, *COBIT 2019*®. Schaumburg: ISACA, 2018.

[13] C. E. Oehlhorn, C. Maier, S. Laumer, and T. Weitzel, "Human resource management and its impact on strategic business-IT alignment: A literature review and avenues for future research," *The Journal of Strategic Information Systems*, vol. 29, no. 4, p. 101641, Dec. 2020, http://dx.doi.org/10.1016/j.jsis.2020.101641.

[14] W. B. Helouani, "Business-it alignment: the emerging role of it business partner," EAESP FGV, 2022. [Online]. Available: https://hdl.handle.net/10438/31794

[15] P. Weill and J. W. Ross, *IT Governance - How Top Performers Manage IT Decision Rights for Superior Results*. Boston: Harvard Business Review Press, 2004. [Online]. Available: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5181980

[16] S. D'Onofrio, "IT-Business-Partner-Management: Begriffsklärung, Abgrenzung und Business-Alignment," *HMD*, vol. 62, no. 2, pp. 259–273, Apr. 2025, http://dx.doi.org/10.1365/s40702-025-01178-4.

[17] A. Eckhardt and C. Rosenkranz, "Lost in translation?!: the need for a boundary spanner between business and it," in *Proceedings of the SIGMIS-CPR '10: 2010 Computer Personnel Research Conference*, Vancouver BC Canada: ACM, May 2010, pp. 75–82, http://dx.doi.org/10.1145/1796900.1796930.

[18] C. Riedinger, C. Rentrop, M. Huber, and S. Zimmermann, "Partnerschaft auf Augenhöhe: Faktoren für eine erfolgreiche Gestaltung der IT-Business-Partner-Organisation," *HMD*, vol. 62, no. 2, pp. 274–293, Apr. 2025, http://dx.doi.org/10.1365/s40702-025-01151-1.

[19] SFIA Foundation, "SFIA – The Skills Framework for the Information Age (Version 9)." 2024. Accessed: Feb. 06, 2025. [Online]. Available: https://sfia-online.org

[20] Scaled Agile, Inc., "Scaled Agile Framework (SAFe)," Scaled Agile Inc., 2025. Accessed: Jan. 28, 2025. [Online]. Available: https://scaledagileframework.com

[21] C. Grawe, *Business/IT-Integration: Neuausrichtung der IT-Funktion in Organisationen im Kontext der Digitalisierung*. Wiesbaden: Springer Fachmedien Wiesbaden, 2023, http://dx.doi.org/10.1007/978-3-658-40132-0.

[22] S. Murer, B. Bonati, and F. J. Furrer, *Managed Evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, http://dx.doi.org/10.1007/978-3-642-01633-2.

[23] C. Rentrop, *IT-Governance - Erfolgsfaktor der digitalen Transformation*. Berlin: Erich Schmidt Verlag GmbH & Co. KG, 2023.

[24] C. Gellweiler, "IT architects and IT-business alignment: a theoretical review," *Procedia Computer Science*, vol. 196, pp. 13–20, 2022, http://dx.doi.org/10.1016/j.procs.2021.11.067.

[25] I. Aguilar Alonso, J. Carrillo Verdún, and E. Tovar Caro, "Description of the structure of the IT demand management process framework," *International Journal of Information Management*, vol. 37, no. 1, pp. 1461–1473, Feb. 2017, http://dx.doi.org/10.1016/j.ijinfomgt.2016.05.004.

[26] R. Foorthuis and S. Brinkkemper, "Best Practices for Business and Sys tems Analysis in Projects Conforming to Enterprise Architecture," *Enterprise Modelling and Information Systems Architectures*, vol. Vol 3, pp. 36-47 Pages, Dec. 2015, http://dx.doi.org/10.18417/EMISA.3.1.4.

[27] U. Flick, *Triangulation*. Wiesbaden: VS Verlag für Sozialwissenschaften, 2011, http://dx.doi.org/10.1007/978-3-531-92864-7.

[28] R. K. Yin, *Case study research: design and methods*, 5. edition. Los Angeles London New Delhi Singapore Washington, DC: SAGE, 2014.

[29] M. Saunders, *Research Methods for Business Students*, 8th ed. Harlow, UK: Pearson, 2019.

[30] P. Baxter and S. Jack, "Qualitative Case Study Methodology: Study De sign and Implementation for Novice Researchers," *TQR*, Jan. 2015, http://dx.doi.org/10.46743/2160-3715/2008.1573.

[31] M. Meuser and U. Nagel, "Das Experteninterview - konzeptionelle Grundlagen und methodische Anlage," in *Methoden der vergleichenden Politik- und Sozialwissenschaft*, S. Pickel, G. Pickel, H.-J. Lauth, and D. Jahn, Eds., Wiesbaden: VS Verlag für Sozialwissenschaften, 2009, pp. 465–479.

[32] J. M. Corbin and A. L. Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 4th ed. Thousand Oaks, CA: Sage Publications, 2015.

[33] C. Riedinger, S. Zimmermann, C. Rentrop, and C. Felden, "The Operationalization of Strategic Alignment: Driving IT Value through Business-IT Dialogues," in *Proceedings of the 58th Hawaii International Conference on System Sciences*, Hawaii, 2025. [Online]. Available: https://hdl.handle.net/10125/109571

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 605–614
DOI: 10.15439/2025F3752
ISSN 2300-5963 ACSIS, Vol. 43

Optimizing Deep Learning for Cotton Leaf Disease Detection Using Meta-Heuristic Feature Selection Algorithms

Naeem Ullah*, Francisco Martínez-Álvarez§, Ivanoe De Falco‡ and Giovanna Sannino‡
*Department of Electrical Engineering and Information Technology - University of Naples "Federico II", Naples, Italy
Email: naeem.ullah@unina.it
§Data Science & Big Data Lab, Pablo de Olavide University, Seville, Spain
Email: fmaralv@upo.es
‡Institute for High-Performance Computing and Networking (ICAR) - National Research Council (CNR), Naples, Italy
Email: {ivanoe.defalco, giovanna.sannino}@icar.cnr.it

***Abstract*—Effective and efficient disease detection is crucial, particularly for economically important crops like cotton. In this paper, we move from the initial development of a deep-learning model for cotton leaf disease detection, called Deep-CCNet, to a more comprehensive comparison of different feature selection algorithms, such as RainWater Algorithm, Particle Swarm Optimization, Bee Evolutionary Algorithm, Genetic Algorithm, and Binary Dragonfly Algorithm. Although Deep-CCNet achieved satisfactory classification performance, the goal of this study is to improve the classification performance and efficiency of deep learning models with meta-heuristic feature selection techniques. This study aims to determine which feature selection method achieves the best balance between performance and computational efficiency. We used the Kaggle "cotton leaf disease dataset", which has 1,711 images from four classes (namely curl virus, bacterial blight, fusarium wilt, and healthy leaf images), to compare these techniques systematically. Our research attempts to find the most effective method that maximizes model performance while minimizing computing resources, in addition to benchmarking the computational and performance parameters of each approach. The results of this study provide a new approach for the choice of feature selection methods in plant pathology, leading to better early disease diagnosis and increased crop resilience via efficient farming practices.**

I. Introduction

COTTON, also referred to as the "silver fiber", is among the most significant world agricultural products, the cornerstone of the textile industry, and a means of livelihood for millions of farmers, exporters, processors, and producers, particularly in developing nations [1], [2]. Cotton production, however, is facing threats from numerous pests and diseases, which considerably reduce the quantity and quality of the cotton crop, resulting in significant economic losses [3], [4]. Some of these issues are cotton leaf diseases, which are directly harmful to the health of a plant and ultimately its fiber production. Some familiar infections include fusarium wilt, bacterial blight, and cotton leaf curl virus. These diseases reduce crop yield and weaken plants, increasing vulnerability to pests and secondary infections. [3].

Studies show that pests and diseases cause a loss of approximately 42% of world agricultural production each year [4]. Traditional disease diagnosis methods, which are mainly based on manual inspection and chemical treatment, are labor-intensive, time-consuming, and often prone to errors. For rural areas where agricultural experts are not available, a precise manual diagnosis is even harder. Therefore, there is a critical need for improved, more efficient, accurate, and timely disease diagnosis techniques [5], [6].

To address the above limitations, the farm sector has witnessed a strong shift towards applying computer vision (CV) and artificial intelligence (AI) technology to automatic, early detection of plant disease [7]–[9]. Leaves have been discovered to be most suitable for such analysis due to their ease of access and visibility compared to other plant components, including stems and roots [10]. Modern CV pipelines typically involve processes such as dataset collection, image pre-processing, feature extraction, feature selection, and classification [11], [12].

Deep convolutional neural networks (DCNNs) have been found effective for automatic feature extraction and classification [13]. In particular, the fully connected layers in DCNNs play an important role by integrating spatial features, such as edges, textures, and color intensities, from the preceding layers to predict the correct class labels [14]. Nevertheless, deep models tend to generate an enormous number of redundant or irrelevant features, thereby leading to overfitting and inefficiencies.

To enhance model resilience and performance, optimization-based feature selection algorithms have been gaining momentum [15]. Techniques such as Particle Swarm Optimization (PSO) [16], Binary Dragonfly Algorithm (BDA) [17], and various meta-heuristic techniques are well suited to select the most salient features with the elimination of noisy or uninformative features. This process of feature selection not only enhances classification accuracy but also reduces the computational cost and results in models that are efficient and lightweight.

In the context of these advancements, this study proposes an integrated approach in which a novel Deep-CCNet deep

learning (DL) model is coupled with meta-heuristic feature selection methods to better improve cotton leaf disease classification performance. Through optimization of feature subsets derived using DeepCCNetFS, we aim to develop a robust and computationally capable model with real-time usage potential in precision farming.

A. Motivation

The key aim of this work is to fulfill the essential requirement of proper early diagnosis of cotton leaf diseases with the least dependency on chemical treatment and facilitate farmers in taking wise decisions using real-time tools. Being highly susceptible to diseases, cotton crops need precise and timely diagnosis for sustainable agriculture, enhanced crop management, and improved yield.

But the use of DL models in such tasks is generally hindered by high-dimensional feature spaces that not only incur computational expense but also introduce redundant information, leading to overfitting and reduced generalization capability. Thus, efficient feature selection is very important in avoiding these constraints.

Lastly, this research seeks to facilitate sustainable agriculture by reducing the application of unnecessary pesticides through efficient early disease detection while making such AI tools accessible and feasible for in-field implementation in cotton farming.

B. Feature Selection Strategies and Proposed Framework

Feature selection is a key component of the modern machine learning pipeline, particularly when dealing with high-dimensional learned feature spaces from deep neural networks. It is crucial to select a subset of the most critical features in building lightweight, efficient, but very accurate models [18]. Without effective feature selection, models can overfit, become computationally more expensive, and be less generalizable, characteristics particularly undesirable in real-world applications such as smart agriculture.

In this context, the deep features extracted from the DeepCCNet model need to be passed from the feature selection step. Since the extracted feature vectors have a high dimensionality, an optimal feature optimization method is essential to enhance the model's performance with computational efficiency.

To address this, we comprehensively evaluate and compare five meta-heuristic feature selection algorithms with different approaches to dynamic exploration and exploitation of the vast and complex feature space: Rainwater Algorithm (RWA) [19], Genetic Algorithm (GA) [20], Particle Swarm Optimization (PSO) [21], Binary Dragonfly Algorithm (BDA) [22], and Bee Evolutionary Algorithm (BEA) [23].

All of these methods are implemented within the DeepCCNet pipeline to optimize feature subsets derived from DeepCCNet for the highest classification performance and the lowest model complexity. Comparative evaluation of these feature selection algorithms is performed based on two key criteria:

1) Classification Performance: Assessment of how well the selected features enable accurate disease diagnosis over the cotton leaf disease dataset.
2) Computational Efficiency: Measuring the overall execution time, feature selection time, and classification time, which is important for effective deployment in resource-constrained agricultural settings.

Furthermore, we used the grid search to carefully optimize the hyperparameters, guaranteeing the best possible model efficiency and performance. For empirical validation, we employ the publicly available Cotton Leaf Disease dataset with 1710 images corresponding to four classes of diseases.

II. Related Works

Recently, the accurate and early detection of cotton leaf diseases using computational methods has been the subject of much research because of its potential effects on agricultural production and sustainability. Conventional ML methods have been the mainstay of previous research, but deep learning (DL) and evolutionary algorithms (nature-inspired feature selection) techniques have lately gained traction and provided notable gains in automation and accuracy.

A. Traditional Methods

Image processing methods and traditional ML algorithms were the main tools used in the early stages of plant disease identification efforts. For example, Sarangdhar et al., [24] proposed a method for the classification and control of diseases on cotton leaf along with soil quality monitoring. The study suggested a regression approach based on support vector machines (SVM) for the diagnosis of five diseases that affect cotton leaves: Alternaria, Bacterial Blight, Cereospra, Fusarium wilt, and Gray Mildew. Farmers were given access to the identification of disease along with its cures through an Android app upon illness detection. Along with the water level in a tank, the Android app was also utilized to display the values of soil factors, including temperature, moisture content, and humidity. Shakel et al., [25] divided the images of Cercospora cotton leaves into clusters using the K-means clustering technique. The hybrid approach for extracting texture and color characteristics was used to extract features. Ultimately, Cercospora cotton leaves were classified using SVM. Similarly, a Multi-SVM-based categorization system for cotton leaf diseases was proposed by Jenifa et al. [26]. The suggested algorithm processed the crops automatically at every stage, identifying defects at an early stage and preventing losses to the farmer in terms of products and money. When provided with the necessary data, the aforementioned Multi-SVM algorithm models, which are either human or robot-supervised, enable the segmentation and classification of various disease kinds. Despite being innovative, these methods have several difficulties. The quality and diversity of the input photos had a significant impact on the accuracy of disease identification, and the manual feature extraction procedure was labor-intensive and prone to human error, which made it challenging to scale and adapt to various plant diseases.

B. Deep Learning Advances

To address the shortcomings of manual feature extraction, the shift towards DL, particularly DCNNs, has significantly

enhanced the robustness and accuracy of plant disease detection systems. These models automate the feature extraction process, capturing complex patterns more effectively than traditional methods. Compared to conventional techniques, these models capture complicated patterns more successfully by automating the feature extraction process [27]. For example, Rai et al. [28] concentrated on developing an enhanced DCNN-based model to address the issue of disease detection and classification in cotton plants. Three distinct experimental setups were examined to examine the effects of varying data split ratios, pooling layer selections (max-pooling vs. average-pooling), and epoch length. A collection including 2293 photos of plants and cotton leaves was used to train the models. Four different leaf classes, plant disease combinations, and the corresponding categories were included in the data. The paper in [29] introduces a dynamic feature selection methodology applied to deep neural networks for univariate and multivariate time series forecasting, based on the Coronavirus Optimization Algorithm [30]. The proposed method, evaluated against random forest models, demonstrated superior performance in terms of mean squared error, highlighting the efficiency and effectiveness of dynamic feature selection. The network that Rajasekar et al. [31] proposed outperformed previous state-of-the-art methods by combining the advantages of the Xception component with the ResNet pre-trained on ImageNet. Kotian et al. [32] used the ResNet50 and KNN ML model for the classification of cotton leaf diseases. Numerous studies have lately used DL methods for the categorization of cotton leaf diseases, as Table 1 illustrates. However, for real-world agricultural applications, their high computing complexity and opaque decision-making procedures often make them more difficult to trust and less accessible. In addition to the studies discussed above, several other DL-based approaches are summarized in Table I for completeness and broader context.

C. Limitations

Our research intends to solve the following important shortcomings of current approaches, despite significant advances in agricultural diagnostics, especially in the diagnosis of cotton leaf disease:

1) Limited Integration of Feature Selection with DL: DL models are commonly used in plant pathology research to detect diseases, although adaptive and dynamic-feature selection methods are not fully integrated. DL models are extremely strong, but their efficacy is sometimes limited by static feature sets that don't change in response to fresh information or intricate disease patterns. Moreover, evolutionary algorithms are rarely used in research, despite their potential to improve model accuracy and computing efficiency by constantly optimizing and refining feature selection.

2) Computational Efficiency and Practical Deployment: The computing needs of training large-scale models are often overlooked by DL techniques, which restricts their application in resource-constrained situations common to many agricultural contexts. This disparity emphasizes the requirement for optimal models that strike a compromise between operational viability and performance.

D. Solutions

Presently, our research methodology rigorously examines and assesses several feature selection algorithms inside the Deep-CCNet approach, including RWA, PSO, BDA, GA, and BEA. The goal of this comparison study is to evaluate how each strategy affects the model's accuracy, computing efficiency, flexibility, and understandability. By doing this, we close important knowledge gaps regarding feature selection efficiencies in agricultural applications and establish new standards for the creation of systems that are scalable, trustworthy, and understandable for the diagnosis of cotton leaf disease. This effort opens the door for more sophisticated AI-driven agriculture systems to be used on a global scale.

III. Methodology

We performed two types of experiments. Firstly, we conducted end-to-end training, validation, and testing of our Deep-CCNet model. Secondly, to improve the performance and reduce the computational cost of our approach, we extracted deep features from the last fully connected layer of the Deep-CCNet model and subjected them to various feature selection algorithms to determine their effectiveness in reducing feature dimensionality and improving classifier performance. We used SVM, KNN, Decision Tree, and Naive Bayes classifiers for the classification of the selection features. The basic workflow of the proposed method is given in Fig. 1.

A. Image Resizing

To maintain uniformity in input size throughout the dataset, all input images are resized to 224 by 224 pixels.

B. Dataset Partitioning

The Deep-CCNet model is evaluated with a twofold approach: 10% of the images are used as a hold-out test set for testing, and the rest, 90% are used in a five-fold cross-validation setup, where in each iteration a different fold (data) is used as a validation set. This approach ensures strong testing against overfitting and encourages generalization.

C. Deep-CCNet Architecture details

The architecture of the Deep-CCNet model is based on the ResNet18's architectural principles; however, it also includes notable modifications specifically designed to maximize the classification performance and reduce computational costs (Table II). The proposed Deep-CCNet architecture is a lightweight convolutional neural network that is optimized for efficiency and fast convergence, ideal for small or medium-sized datasets such as cotton leaf disease classification. Deep-CCNet possesses approximately 5.4 million learnable parameters, significantly lower than ResNet18's 11.18 million parameters. Unlike ResNet18, which uses default ReLU activations, Deep-CCNet uses Leaky ReLU activation units throughout the network to correct the dying neuron problem and facilitate gradient flow.

TABLE I
RECENT WORK ON AUTOMATIC COTTON LEAF DISEASE DETECTION USING DL AND ML TECHNIQUES

Reference	Number of classes	Method	Images in the dataset	Advantages	Limitation	Year
Lakshmi et al. [33]	Seven	Pre-trained ResNet50 and region based segmentation	2,384	Satisfactory results	No Explainability (NE) and computationally expensive approach	2024
Bhujade et al. [34]	Ten	Hybrid approach (CNN with Multi-Resolution Feature Optimization)	990	Good generalization ability and efficient	No Unseen Samples Testing (NUST) and NE	2024
Mohammad et al. [35]	Six	Pre-trained models and custom CNN	1,200	Proposed a new dataset with high-quality images, satisfactory results on unseen samples	NE	2024
Irfan et al. [36]	Eight	Deep CNN with modified dragonfly optimization	4,107	Good generalization ability and satisfactory performance on four datasets	NE and manual hyper-parameter selection	2,024
Manurkar et al. [37]	Six	CNNs	3,600	Balanced dataset used and efficient approach	NUST, low-performance results and NE	2024
Rai et al. [28]	Four	Improved Alexnet	2,293	Satisfactory results, testing on unseen samples, less complex and efficient	Low generalization ability	2023
Arathi et al. [38]	Four	Pre-trained DenseNet-121 model	781	Simple approach	Computationally expensive and NE	2023
Zafar et al. [39]	Four	Feature extraction using EfficientNet-b0 and Inception-v3, feature selection using Emperor Penguin Optimizer, and classification using Quadratic Discriminant Analysis	1,786	Satisfactory performance and good generalization ability	NE	2023
Naeem et al. [40]	Four	VGG16	700	Custom dataset and satisfactory results with a small dataset	NE, low generalization ability, and NUST	2023
Govin-dasamy et al. [41]	Five	Filated convolutional feature pyramid network	26,100	Efficient and reliable approach	Low classification accuracy	2023
Pandey et al. [42]	Four	Hybrid model (CNN plus SVM)	1,710	Accurate and effective approach of classification	NE, computationally expensive, and low generalization ability	2023

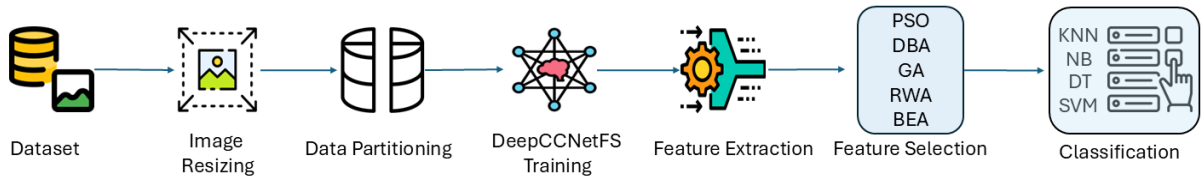


Fig. 1. Overview of the proposed methodology for cotton leaf disease classification.

Furthermore, Deep-CCNet reduces the residual block architecture by eliminating every other convolutional block, creating a shallower but effective model without over-sacrificing feature learning. Deep-CCNet introduces an additional fully connected layer of 1096 neurons preceded by a dropout layer (probability = 0.5) to mitigate overfitting, which is not included in ResNet18.

Moreover, although both networks employ strided convolutions to downsample, Deep-CCNet is specifically architected to predict for 4 classes (i.e., three types of cotton leaf diseases and a healthy class), as opposed to ResNet18's native 1000-class ImageNet configuration. This change in structure yields a computationally more lightweight model that is simpler to train and better suited to domain-specific classification tasks.

D. Feature Extraction

Following preprocessing, images are sent to the DeepCC-NetFS, where feature extraction is handled by many layers. Images are subjected to several filter applications as they go through the convolutional layers, such as conv1 of the first residual unit, which produces feature maps that capture different

TABLE II
KEY DIFFERENCES BETWEEN RESNET18 AND DEEPCCNETFS.

S. No	Property	DeepCCNetFS	ResNet18
1	Number of Learnable Layers	12	18
2	Number of Fully Connected (FC) Layers	2	1
3	Number of Convolutional Layers	12	17
4	Number of Dropout Layers	1	0
5	Activation Function (AF)	Leaky ReLU (LReLU)	ReLU
6	Focus	Optimized for specific patterns in cotton leaf diseases	General-purpose image recognition
7	Computational Efficiency	Enhanced by reducing convolutional blocks and simplifying architecture	Standard efficiency with more layers
8	Layer Configuration	Customized layers and fewer residual blocks for specific disease feature extraction	Standardized residual blocks for diverse feature extraction
9	Total learnable parameters	5.4 Million	11.18 Million
10	Residual Blocks	4	8

aspects of the images, such as edges and textures, and can be represented as:

$$\mathrm{f}_{ij}^{k} = \sum((c^{k} \times x) + b_k) \tag{1}$$

Here, x represents the local region of the image being convolved, k is used to represent the kth layer, f denotes the feature's value, (i, j) represents pixel coordinates, c^k represents the convolution kernel of current layer, and b_k denotes the bias. Batch normalization layers are then used to normalize these feature maps, which reduces internal covariate shifts and allows for faster and more stable training.

Max pooling layers, such as pool1, assist in lowering the computational burden and model complexity by controlling overfitting and summarizing the convolutional layer outputs into more abstract representations. This reduces the spatial dimension. The network can learn identification functions from residual connections, like res2a, to make sure that no data is lost between layers. Because these connections give gradients direct paths during backpropagation, they are essential for enabling the training of very deep networks. The function $f(x)$ yields the ideal feature map.

$$f(x) = max(x1, x2, x3, ...x_n) \tag{2}$$

where $x_1, x_2, x_3, \ldots, x_n$ are the activation values within a local region (e.g., 3×3 window) of the feature map. This operation retains the most prominent features while discarding less significant ones, making the representation more robust and compact. Higher-level reasoning in the network takes place in the FC layers, where fc7 and fc8 are essential building blocks for condensing the distilled information into a format appropriate for categorization. The most important characteristics in Deep-CCNet are effectively highlighted by a global average pooling layer (pool5) that first reduces the feature maps to a single value per map.

The output from the last fully connected layer (fc8) serves as the extracted feature set for downstream tasks. Finally, the softmax layer computes class probabilities, enabling the model to classify cotton leaf images into specific disease categories.

E. Feature Selection Algorithms

Our study's feature selection procedure uses sophisticated algorithms that combine local and global search techniques to extract the most pertinent characteristics from high-dimensional data. By removing redundant and unnecessary features, these techniques decrease the dimensionality of the data and improve the classification models' performance.

1) Binary Dragonfly Algorithm (BDA): The BDA is a nature-inspired optimization method based on the foraging behavior of dragonflies. This algorithm uses a swarm of candidate solutions as binary vectors. It iterates in steps where each solution is updated according to local best (exploitation) and global best (exploration) strategies. The algorithm uses a combination of velocity updates and binary transformations (with sigmoid-like updates) to search the feature space for an optimal feature subset. The fitness function evaluates the feature subsets using a classification model, with the best features selected based on minimum classification error.

2) Bee Evolutionary Algorithm (BEA): The BEA is inspired by the foraging process of bees, i.e., employed, onlooker, and scout stages. The algorithm starts by initializing a population of binary solutions (representations of feature subsets) and evaluating their fitness based on a classification model (KNN). The employed bees attempt to mutate solutions to reach better fitness, and onlooker bees select solutions probabilistically based on fitness. Scout bees search randomly in the solution space when they cannot find better solutions. These steps are repeated by the algorithm iteratively, tracking convergence and improving the population for better feature selection.

3) Genetic Algorithm (GA): The GA is a metaheuristic algorithm inspired by natural selection that is used to solve search and optimization problems. In this algorithm, every member of the population is a binary vector where every bit corresponds to a feature being included or not. The algorithm determines the fitness of every member by training a classifier (for example, KNN) using the selected features. In each generation, the population is refreshed using selection, crossover, and mutation. Selection serves as a promise of

carrying forward superior individuals to the next generation. Crossover exchanges traits among individuals, and mutation introduces randomness using bit flipping. The iterations are iterated over a defined number of generations, and the optimal individual (set of features) is employed in model construction.

4) Particle Swarm Optimization (PSO): The PSO algorithm is an optimization method that is based on the social behavior of bird flocking. The PSO uses a population of particles, and each particle is a possible feature subset. Each particle updates its position according to its own best position and the global best position discovered by the swarm. The algorithm balances exploration (moving towards new areas of the solution space) and exploitation (refining current good solutions) through velocity updates. The particles converge to an optimal subset of features after iterating around the solution space. Fitness is evaluated in terms of classification accuracy on a selected set of features.

5) RainWater Algorithm (RWA): RWA is inspired by raindrop movement, where raindrops "move down" to lower levels, representing a natural minimization search for the location of the minimum in the feature space. It simulates raindrop behavior through iterative solution movement by fitness in trying to achieve the best set of features, reducing classification error. The solutions (sets of features as subsets) are refined through multiple iterations. The algorithm initializes with a random population of feature subsets, computes their fitness from their sum of feature values, and shifts them to a new position by adjusting velocities toward the optimal solution. A dynamic threshold guarantees that the correct features are selected, even when the initial threshold does not yield any selection.

IV. Results and Discussion

A. Dataset

In this work, a publicly available Kaggle data set named the "Cotton leaf disease data set" is used [43]. The dataset consists of 1,711 images representing four classes: Curl Virus (418 images), Bacterial Blight (448 images), Fusarium Wilt (419 images), and Healthy (426 images). The images were captured under real-world conditions in the fields of the Southern Punjab region of Pakistan, a major cotton-producing area. Some of the images were also downloaded from the internet. This dataset aims to assist in the identification and classification of cotton leaf diseases, which are critical for the agriculture industry, particularly for cotton cultivation and export to Europe and the US. The representative samples of the dataset are given in Figure 2.

B. Performance of Deep-CCNet without Feature Selection

The Deep-CCNet model is first evaluated without feature selection as a baseline performance. The five-fold cross-validation results are shown in Table III, and the test results on unseen data are shown in Table IV.

Deep-CCNet achieved a good mean accuracy of 98.06% under cross-validation, with precision, recall, and F1 score all well aligned around 98.0% (Table III). The model is very stable in its performance, indicated by the low standard deviations (below 0.7 for all metrics) and low variance values. The best accuracy achieved is 99.03%, indicating the model generalized well when trained.

On unseen test samples (Table IV), the model performed consistently with a mean accuracy of 96.13%, precision of 96.09%, recall of 96.26%, and F1-score of 96.15%. The results show minimal performance loss between the validation and testing phases, indicating the generalizability of the model. Furthermore, the values of standard deviation on the test set remained low, pointing towards the stability of the model across different data splits.

Overall, the Deep-CCNet model alone without feature selection already exhibits a very good classification capacity for the detection of cotton leaf disease and offers a high baseline for achievable improvement through feature selection techniques.

TABLE III
Five-Fold cross-validation results for the cotton leaf disease dataset (validation).

Metric	Accuracy	Precision	Recall	F1-score
Average	98.06	97.97	98.02	97.99
Median	97.73	97.70	97.80	97.75
Maximum	99.03	98.72	98.72	98.71
Minimum	97.09	97.07	97.08	97.08
Standard deviation	0.63	0.69	0.22	0.62
Variance	0.50	0.40	0.48	0.39

TABLE IV
Testing on unseen samples for the cotton leaf disease dataset.

Metric	Accuracy	Precision	Recall	F1-score
Average	96.13	96.09	96.26	96.15
Median	96.43	96.43	96.43	96.43
Maximum	96.51	96.47	96.59	96.53
Minimum	95.35	95.26	95.55	95.37
Standard deviation	0.48	0.37	0.37	0.44
Variance	0.20	0.23	0.14	0.20

C. Performance Evaluation of Meta-Heuristic Feature Selection Methods

The objective of this experiment is to evaluate systematically the performance of various meta-heuristic feature selection algorithms, including RWA, PSO, DBA, GA, and BEA, in improving the performance of classification models of cotton leaf disease. The classification using four different classifiers—SVM, Naïve Bayes, KNN, and Decision Tree is shown in Table V.

Among the tested methods, the BEA algorithm combined with Naïve Bayes yielded the best overall performance, achieving the highest accuracy (97.66%), precision (97.62%), recall (97.65%), and F1-score (97.63%). The next best results are obtained using BEA with KNN (accuracy: 97.08%, F1-score: 97.07%) and SVM (accuracy: 95.32%, F1-score: 95.30%)

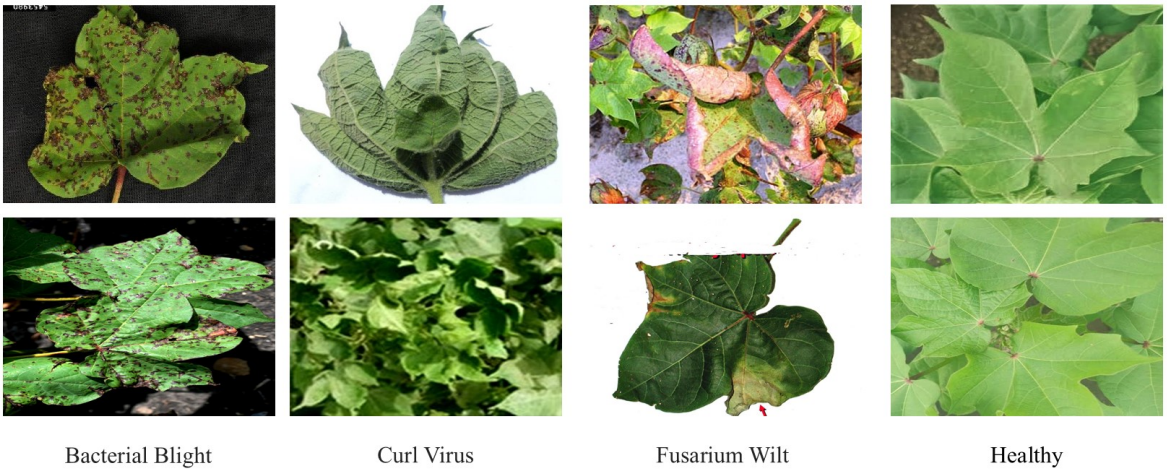


Fig. 2. Sample images from the Cotton Leaf Disease Dataset.

classifiers, respectively. The DBA method also demonstrated strong performance, particularly with KNN, which achieved a precision of 97.22% and an F1-score of 97.56%. Similarly, the GA algorithm combined with Naïve Bayes produced competitive results, with an accuracy of 97.14%, precision of 97.22%, recall of 97.22%, and F1-score of 97.22%. These results suggest that while BEA leads in overall classification performance, both DBA and GA also offer robust alternatives when paired with appropriate classifiers.

On the other hand, RWA generated the poorest classification performances of all the classifiers, with accuracies between 58.00% and 72.00%, which means that its feature selection quality is inferior for this task. PSO and GA showed comparable performance, particularly when applied with Naïve Bayes and Decision Tree classifiers, which means that they offer good trade-offs between feature selection quality and model generalization.

In general, BEA and DBA proved to be the best feature selection methods for improving model performance, with RWA being less ideal in this situation.

Furthermore, as shown in Table V, the number of selected features varies across methods, with RWA selecting only 1 feature, PSO selecting 9, and other algorithms like DBA, GA, and BEA consistently selecting 4 features, demonstrating the impact of feature subset size on classification performance.

D. Computational Cost Analysis of Meta-Heuristic Feature Selection Methods

The objective of this experiment is to compare the computational efficiency of all algorithms used for feature selection in terms of feature selection time, classifier training and testing time, and total running time, as illustrated in Table VI.

RWA recorded the fastest total execution time (7.55 seconds) because of its relatively simple optimization process, but poor classification accuracy limits its practical applicability. PSO and DBA recorded moderate execution times (16.65 and 14.71 seconds, respectively) but excellent classification accuracy, making them good options where both efficiency and accuracy matter.

GA, although ranking well, took significantly more overall execution time (60.15 seconds) than PSO and DBA due mainly to the iterative nature of genetic operations. BEA, although showing the best classification performance, took the highest computational cost (451.39 seconds), which highlights a trade-off between predictive performance and computational resource usage. This substantial computational overhead is largely attributed to the ensemble-based nature and nested search operations within BEA, which, while improving feature selection quality, lead to increased time complexity.

From a practical standpoint, the high execution time of BEA raises concerns regarding scalability and real-time deployment feasibility, especially for edge devices or resource-constrained environments. For instance, in large-scale time series applications or streaming scenarios, such computational demand may hinder timely decision-making or exhaust device resources. In contrast, PSO and DBA strike a more favorable balance by offering competitive accuracy at a fraction of the computational cost, making them more viable for real-world deployment.

Therefore, while BEA may be suitable for offline analysis or batch processing tasks where accuracy is prioritized over speed, PSO and DBA are better suited for operational settings requiring rapid inference and low-latency processing. This analysis underscores the need to carefully match algorithm selection with deployment requirements and hardware constraints.

Briefly, although BEA offers higher classification accuracy, its deployment may be impractical in latency-sensitive environments. PSO and DBA are more balanced choices for scenarios where computational expense is a key consideration.

V. Conclusion

In this work, we have introduced an end-to-end deep learning architecture, Deep-CCNet, for detecting cotton leaf disease and

TABLE V
PERFORMANCE COMPARISON OF FEATURE SELECTION METHODS AND CLASSIFICATION ALGORITHMS

Feature Selection Algorithm	Classifier	# Selected Features	Accuracy	Precision	Recall	F1-Score
RWA	SVM	1	72.55	72.66	73.30	72.57
RWA	Naïve Bayes	1	72.55	72.80		72.58
RWA	KNN	1	58.82	58.99	57.73	58.03
RWA	Decision Tree	1	58.82	59.15	58.33	58.69
PSO	SVM	9	91.43	91.32	91.87	91.63
PSO	Naïve Bayes	9	85.71	86.11	88.19	86.65
PSO	KNN	9	82.35	84.99	85.78	84.47
PSO	Decision Tree	9	88.24	88.19	88.54	88.06
DBA	SVM	4	94.12	94.95	94.95	94.95
DBA	Naïve Bayes	4	91.81	92.67	92.22	92.34
DBA	KNN	4	97.06	97.22	97.91	97.56
DBA	Decision Tree	4	94.12	94.95	94.95	94.95
GA	SVM	4	94.12	93.75	95.00	94.21
GA	Naïve Bayes	4	97.14	97.22	97.22	97.22
GA	KNN	4	91.18	90.50	91.87	91.27
GA	Decision Tree	4	91.18	90.62	93.18	91.00
BEA	SVM	4	95.32	95.28	95.33	95.30
BEA	Naïve Bayes	4	97.66	97.62	97.65	97.63
BEA	KNN	4	97.08	97.06	97.16	97.07
BEA	Decision Tree	4	94.15	94.13	94.21	94.16

TABLE VI
COMPUTATIONAL COST ANALYSIS OF FEATURE SELECTION METHODS.

Feature Selection Algorithm	Best Accuracy	Feature Selection Time (s)	Classifier Training and Testing Time (s)	Total Execution Time (s)
RWA	72.55	0.24	6.24	7.55
PSO	91.43	15.73	0.44	16.65
DBA	97.06	12.99	1.01	14.71
GA	97.14	58.92	1.02	60.15
BEA	97.66	448.65	2.61	451.39

have extensively analyzed the impact of various meta-heuristic feature selection techniques, including RWA, PSO, DBA, GA, and BEA, on classification performance and computational expense. The Deep-CCNet model has achieved high baseline accuracy (96.13% on unseen samples without feature selection), which is proof of its effectiveness for this task.

Through comprehensive experimentation, we found that the BEA achieved the highest classification performance among different classifiers, followed by DBA and PSO, providing a better balance between classification accuracy and computational cost. However, RWA, though having low computational cost, reported significantly lower predictive performance and was therefore less suitable for this application.

Aside from these promising results, there are several limitations to this work. First, the study is conducted on a single cotton leaf disease dataset; hence, the generalizability of the Deep-CCNet model and feature selection techniques to other crops and general plant pathology tasks has not been established. Second, while the study achieved high classification performance, it did not incorporate explainability techniques to interpret model predictions or feature selection outcomes, limiting the transparency and trustworthiness of the system for end users such as farmers and agronomists. Finally, the current approach uses traditional classifiers after feature selection; more advanced classifiers or ensemble techniques could further improve results.

Future work can extend this research in many directions. One direction would be to apply the Deep-CCNet model and feature selection steps to larger, multi-crop datasets for their scalability and robustness. Exploring light-weight models optimized for deployment on mobile and edge devices would also enhance the real-world usability of the system in actual agricultural settings. Last but not least, developing explainable AI (XAI) modules for model prediction and feature importances could provide actionable insights to farmers and agronomists and bring the study closer to the overall goals of interpretable and trustworthy AI in agriculture.

CODE AVAILABILITY

The source code used in this study is publicly available[1]. This facilitates reproducibility, enables validation of the results, and encourages further research and development in the field.

[1]https://github.com/Engr-Naeem-Ullah/Optimizing-Deep-Learning-with-Meta-Heuristic-Feature-Selection-Algorithms

REFERENCES

[1] S. Radhakrishnan, "Sustainable cotton production," in *Sustainable fibres and textiles*, pp. 21–67, Elsevier, 2017.

[2] S. P. Patil and R. S. Zambre, "Classification of cotton leaf spot disease using support vector machine," *International Journal of Engineering Research*, vol. 3, no. 4, pp. 1511–1514, 2014.

[3] A. Farooq, J. Farooq, A. Mahmood, A. Shakeel, K. A. Rehman, A. Batool, M. Riaz, M. T. H. Shahid, and S. Mehboob, "An overview of cotton leaf curl virus disease (clcud) a serious threat to cotton productivity," *Australian Journal of Crop Science*, vol. 5, no. 13, pp. 1823–1831, 2011.

[4] A. Jenifa, R. Ramalakshmi, and V. Ramachandran, "Cotton leaf disease classification using deep convolution neural network for sustainable cotton production," in *2019 IEEE international conference on clean energy and energy efficient electronics circuit for sustainable development (INCCES)*, pp. 1–3, IEEE, 2019.

[5] H. Qayyum, S. T. H. Rizvi, M. Naeem, U. b. Khalid, M. Abbas, and A. Coronato, "Enhancing diagnostic accuracy for skin cancer and covid-19 detection: A comparative study using a stacked ensemble method," *Technologies*, vol. 12, no. 9, p. 142, 2024.

[6] A. Ismail, M. Naeem, U. B. Khalid, and M. Abbas, "Improving adherence to medication in an intelligent environment using reinforcement learning," *Journal of Reliable Intelligent Environments*, vol. 11, no. 1, p. 3, 2025.

[7] N. Ullah, J. A. Khan, S. Almakdi, M. S. Alshehri, M. A. Qathrady, E. A. Aldakheel, and D. S. Khafaga, "A lightweight deep learning-based model for tomato leaf disease classification," *Computers, Materials & Continua*, 2023.

[8] P. Timalsina, S. Bhattarai, S. Paudel, and S. Jha, "Enhancing plant disease detection through image analysis using ssdmobilenetv2 and resnet50," in *Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering* (V. K. Solanki, T. D. Tan, P. Kumar, and M. Cardona, eds.), vol. 42 of *Annals of Computer Science and Information Systems*, p. 79–87, PTI, 2024.

[9] M. H. Calp and V. K. Solanki, "Classification of plant species with iris dataset using ann, knn and k-means algorithms," in *Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering* (P. Kumar, M. Cardona, V. K. Solanki, T. D. Tan, and A. Wahid, eds.), vol. 38 of *Annals of Computer Science and Information Systems*, p. 7–10, PTI, 2023.

[10] N. Ullah, J. A. Khan, S. Almakdi, M. S. Alshehri, M. Al Qathrady, N. El-Rashidy, S. El-Sappagh, and F. Ali, "An effective approach for plant leaf diseases classification based on a novel deepplantnet deep learning model," *Frontiers in Plant Science*, vol. 14, p. 1212747, 2023.

[11] A. Ismail, M. Naeem, M. H. Syed, M. Abbas, and A. Coronato, "Advancing patient care with an intelligent and personalized medication engagement system," *Information*, vol. 15, no. 10, p. 609, 2024.

[12] N. Ullah, F. Guzmán-Aroca, F. Martínez-Álvarez, I. De Falco, and G. Sannino, "A novel explainable ai framework for medical image classification integrating statistical, visual, and rule-based methods," *Medical Image Analysis*, p. 103665, 2025.

[13] N. Ullah, I. De Falco, and G. Sannino, "A novel deep learning approach for colon and lung cancer classification using histopathological images," in *2023 IEEE 19th International Conference on e-Science (e-Science)*, pp. 1–10, IEEE, 2023.

[14] X. Zhang, X. Zhang, and W. Wang, "Convolutional neural network," in *Intelligent Information Processing with Matlab*, pp. 39–71, Springer, 2023.

[15] D. Gutiérrez-Avilés, M. J. Jiménez-Navarro, J. F. Torres, and F. Martínez-Álvarez, "MetaGen: A framework for metaheuristic development and hyperparameter optimization in machine and deep learning," *Neurocomputing*, vol. 637, p. 130046, 2025.

[16] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, ieee, 1995.

[17] M. M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, and S. Mirjalili, "Binary dragonfly algorithm for feature selection," in *2017 International conference on new trends in computing sciences (ICTCS)*, pp. 12–17, IEEE, 2017.

[18] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 1, pp. 56–70, 2020.

[19] T. Biyanto, "Rain water algorithm." https://www.mathworks.com/matlabcentral/fileexchange/65617-rain-water-algorithm, 2025. MATLAB Central File Exchange, retrieved May 14, 2025.

[20] M. Farzaneh, "Feature selection in classification using genetic algorithm." https://www.mathworks.com/matlabcentral/fileexchange/74105-feature-selection-in-classification-using-genetic-algorithm, 2025. MATLAB Central File Exchange, retrieved May 14, 2025.

[21] J. Too, A. R. Abdullah, and N. Mohd Saad, "A new co-evolution binary particle swarm optimization with multiple inertia weight strategy for feature selection," *Informatics*, vol. 6, no. 2, p. 21, 2019.

[22] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural computing and applications*, vol. 27, pp. 1053–1073, 2016.

[23] S. M. H. Mousavi, "Bees cnn evolutionary algorithm." https://github.com/SeyedMuhammadHosseinMousavi/Bees-CNN-Algorithm, 2025. GitHub repository, retrieved May 14, 2025.

[24] A. A. Sarangdhar and V. Pawar, "Machine learning regression technique for cotton leaf disease detection and controlling using iot," in *2017 international conference of electronics, communication and aerospace technology (ICECA)*, vol. 2, pp. 449–454, IEEE, 2017.

[25] W. Shakeel, M. Ahmad, and N. Mahmood, "Early detection of cercospora cotton plant disease by using machine learning technique," in *2020 30th International Conference on Computer Theory and Applications (ICCTA)*, pp. 44–48, IEEE, 2020.

[26] A. Jenifa, R. Ramalakshmi, and V. Ramachandran, "Classification of cotton leaf disease using multi-support vector machine," in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1–4, IEEE, 2019.

[27] N. Ullah, J. A. Khan, L. A. Alharbi, A. Raza, W. Khan, and I. Ahmad, "An efficient approach for crops pests recognition and classification based on novel deeppestnet deep learning model," *IEEE Access*, vol. 10, pp. 73019–73032, 2022.

[28] C. K. Rai and R. Pahuja, "Classification of diseased cotton leaves and plants using improved deep convolutional neural network," *Multimedia Tools and Applications*, vol. 82, no. 16, pp. 25307–25325, 2023.

[29] M. J. Jiménez-Navarro, C. Restrepo-Estrada, L. Melgar-García, and D. Gutiérrez-Avilés, "Feature Selection Guided by CVOA Metaheuristic for Deep Neural Networks: Application to Multivariate Time Series Forecasting," *Lecture Notes in Networks and Systems*, vol. 749, p. 209–218, 2023.

[30] F. Martínez-Álvarez, G. Asencio-Cortés, J. F. Torres, D. Gutiérrez-Avilés, L. Melgar-García, R. Pérez-Chacón, C. Rubio-Escudero, A. Troncoso, and J. C. Riquelme, "Coronavirus Optimization Algorithm: A bioinspired metaheuristic based on the COVID-19 propagation model," *Big Data*, vol. 8, no. 4, pp. 308–322, 2020.

[31] V. Rajasekar, K. Venu, S. R. Jena, R. J. Varthini, and S. Ishwarya, "Detection of cotton plant diseases using deep transfer learning.," *J. Mobile Multimedia*, vol. 18, no. 2, pp. 307–324, 2022.

[32] S. Kotian, P. Ettam, S. Kharche, K. Saravanan, and K. Ashokkumar, "Cotton leaf disease detection using machine learning," *Proceedings of the Advancement in Electronics & Communication Engineering*, 2022.

[33] M. D. Lakshmi, B. K. Sahoo, and R. K. Ganiya, "Analysis of cotton leaf curl diseases using advanced learning model," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 2s, pp. 188–194, 2024.

[34] V. G. Bhujade and V. Sambhe, "Multi-disease classification and severity estimation of cotton and soybean leaves using convolutional neural network," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 11s, pp. 584–594, 2024.

[35] S. Mohmmad, G. Varshitha, E. Yadlaxmi, B. S. Nikitha, D. Anurag, C. Shivakumar, and D. M. S. Isan, "Detection and classification of various diseases in cotton crops using advanced neural network approaches," in *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–6, IEEE, 2024.

[36] I. Haider, M. A. Khan, M. Nazir, A. Hamza, O. Alqahtani, M. T.-H. Alouane, and A. Masood, "Crops leaf disease recognition from digital and rs imaging using fusion of multi self-attention rbnet deep architectures and modified dragonfly optimization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[37] V. Manurkar, S. Kulkarni, S. Rokade, and R. Mirajkar, "Cotton plant disease prediction and remedy recommendation system," in *2024 ASU*

International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), pp. 1616–1620, IEEE, 2024.

[38] B. Arathi and U. N. Dulhare, "Classification of cotton leaf diseases using transfer learning-densenet-121," in *Proceedings of third international conference on advances in computer engineering and communication systems: ICACECS 2022*, pp. 393–405, Springer, 2023.

[39] M. Zafar, J. Amin, M. Sharif, M. A. Anjum, S. Kadry, and J. Kim, "Cnn based features extraction and selection using epo optimizer for cotton leaf diseases classification," *Computers, Materials and Continua*, vol. 76, no. 3, pp. 2779–2793, 2023.

[40] A. B. Naeem, B. Senapati, A. S. Chauhan, S. Kumar, J. C. O. Gavilan, and W. M. Abdel-Rehim, "Deep learning models for cotton leaf disease detection with vgg-16," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 550–556, 2023.

[41] S. Govindasamy and D. Jayaraj, "Collaborative ant colony optimization-assisted support vector machine for accurate cotton leaf disease classification and yield prediction," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 15, 2023.

[42] B. N. Pandey, R. P. Singh, M. S. Pandey, and S. Jain, "Cotton leaf disease classification using deep learning based novel approach," in *2023 International Conference on Disruptive Technologies (ICDT)*, pp. 559–561, IEEE, 2023.

[43] S. K. Noon, M. Amjad, M. Ali Qureshi, and A. Mannan, "Computationally light deep learning framework to recognize cotton leaf diseases," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 6, pp. 12383–12398, 2021.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 615–624
DOI: 10.15439/2025F6466
ISSN 2300-5963 ACSIS, Vol. 43

Implementation of random number generator service with IoT device

Rafał Wojszczyk, Aneta Hapka, Kacper Akdağ-Ochnik
0000-0003-4305-7253, 0000-0003-4613-2570
Koszalin University of Technology
ul. §niadeckich 2, 75-453 Koszalin, Poland
Email: {rafal.wojszczyk, aneta.hapka, u18898}@tu.koszalin.pl

***Abstract*—The paper focuses on the problem of generating random numbers, which on the surface appears to be a very simple problem. However, Most software tools used for this purpose produce pseudorandom numbers. This means that, if certain conditions are met, it is possible to reproduce successive sequences of generated numbers. This paper addresses the problem by developing an IoT device to generate true random numbers based on selected physical properties. The device communicates with a server that provides a web service and an API for generating random values in multiple variations.**

I. Introduction

IN THE DIGITAL AGE, randomness plays a crucial role in a wide array of applications, ranging from cryptography and secure communications [4] to simulations of unexpected events [3], gaming, art and GUID or UUID [5]. The quality of randomness directly impacts the performance and reliability of these applications. True Random Number Generators (TRNGs) are particularly valuable in this context, as it derive their randomness from physical processes, ensuring a higher degree of unpredictability compared to pseudo-random number generators (PRNGs), which rely on deterministic algorithms.

Randomness, as a concept, has intrigued researchers and practitioners due to its wide-ranging utility in fields such as mentioned simulations, gaming, and also statistical sampling, and experimental design. While randomness is often associated with cryptographic applications, this project focuses on non-cryptographic use cases, where high-quality random numbers are equally essential but do not require the stringent security guarantees of cryptographic systems. This distinction allows for the exploration of randomness in a broader context, emphasizing accessibility, scalability, and ease of use.

Existing solutions for generating random numbers often come with limitations, such as high costs, restricted access, or reliance on pseudo-random number generators, which, while sufficient for many applications, lack the inherent unpredictability of true randomness derived from physical processes. This paper seeks to bridge this gap by developing a scalable, cost-effective, and user-friendly web service that leverages true random number generators to deliver high-quality random numbers for non-cryptographic purposes. By focusing on applications such as simulations, Monte Carlo methods, random sampling, and artistic content generation, the service aims to provide a reliable and accessible source of randomness for a global audience, fostering innovation and experimentation across diverse domains.

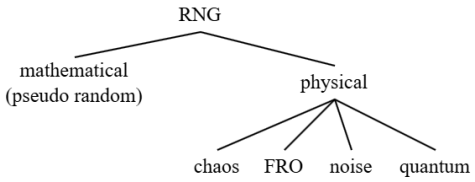


Fig. 1. Classification of random number generators.

Section II discusses how to generate random numbers. Section III describes the most important elements from the construction of the software and the device. Section IV presents and validates the system operation of the whole system and possible applications. The V section provides a summary of the work.

II. Generating random numbers

A. Classification

Random number generators can be divided into two major groups: Mathematical (pseudo random) and Physical. Pseudo random generators can be further divided into several categories based on the type of algorithm used, and physical random number generators are divided into 4 categories: noise based RNGs, free running oscillator (FRO) RNGs, chaos RNGs and quantum RNGs based on the entropy source [7].

B. Pseudo-random numbers

Pseudo-random number generators are algorithms that produce sequences of numbers approximating randomness using deterministic mathematical formulas. They rely on an initial state called a seed, which is processed into a longer sequence of values. PRNG output typically follows a uniform distribution, enabling transformations to other distributions (e.g., Gaussian) [6]. While not truly random, PRNGs are sufficient for applications like simulations, gaming, and randomized algorithms. Cryptographically secure PRNGs (CSPRNGs) enhance security by ensuring outputs are computationally indis-

tinguishable from true randomness, making them essential for cryptographic protocols [7], [8].

PRNGs excel in speed, generating millions of random numbers per second, far surpassing true random number generators limited by physical entropy sources. Their period, the length of the sequence before repetition, varies by algorithm. For instance, the Mersenne Twister has a period of $2^{19937}-1$, making it suitable for large-scale simulations, while simpler PRNGs like linear congruential generators (LCGs) have shorter periods. CSPRNGs feature periods so large that repetition is practically impossible, ensuring long-term reliability [7].

The advantages of PRNGs include low computational cost, ease of implementation, and reproducibility. By using fixed seeds, PRNGs enable consistent results for debugging and verification, making them indispensable in scientific and engineering applications [6], where repeatability is important.

There are a large number of PRNG algorithms implemented. The most popular are

1) **Middle-square method** [9]
   One of the earliest methods, presented by *J. von Neumann* in 1949 at a conference. Highly flawed, with very short period. The numbers are created by squaring the n-digit initial value, adding leading zeroes if the result has fewer than $2n$ digits and extracting middle n digits. The process can be repeated to generate more numbers.
2) **Linear Congruential Generator (LCG)** [10] One of the oldest and most widely used PRNGs, based on the Lehmer generator and published in 1958 by *W. E. Thomson* and *A. Rotenberg*. It is defined by the recurrence relation:

$$X_{n+1} = (aX_n + c) \mod m \tag{1}$$

   where:
   - X is the sequence of random values;
   - $m, 0 < m$ is the modulus;
   - $a, 0 < a < m$ is the multiplier;
   - $c, 0 \leq c < m$ is the increment;
   - $X_0, 0 \leq X_0 < m$ is the seed.

   LCGs are simple and efficient but suffer from lattice structures in higher dimensions, limiting their use in modern applications.
3) **Mersenne Twister** [11] Developed in 1997 by *Makoto Matsumoto* and *Takuji Nishimura*, this PRNG is renowned for its high-quality output and long period. Named after the Mersenne prime $2^{19937} - 1$, which defines its period length, the Mersenne Twister is widely used in simulations and statistical applications. It balances speed and statistical robustness, though its large state size can be a drawback in memory-constrained environments.
4) **Xorshift** [12] Created by *George Marsaglia*, a pioneer in random number generation, Xorshift is an extremely fast PRNG that produces sequences of $2^{32} - 1$ integers. It relies on repeated XOR operations with shifted versions of its state, making it lightweight and efficient. Variants like Xorshift+ and Xorshift* improve statistical properties while maintaining speed.
5) **WELL (Well Equidistributed Long-period Linear)** [13] An improvement over the Mersenne Twister, WELL generators offer better equidistribution and faster recovery from zero states. Developed by *F. Panneton, P. L'Ecuyer, and M. Matsumoto*, WELL variants are suitable for applications requiring high-quality randomness and long periods.
6) **PCG (Permuted Congruential Generator)** [14] A modern family of PRNGs designed by *M. E. O'Neill*, PCG combines linear congruential generators with permutation functions to enhance output quality. PCG variants are compact, fast, and statistically robust, making them popular in gaming and procedural generation.
7) **WELLDOC** [15] Aperiodic PRNGs based on infinite words techniques, developed by *L. Balkova, M. Bucci, A. de Luca, J. Hladky,* and *S. Puzynina* in 2013. These generators are theoretically interesting but less commonly used in practice due to their complexity.
8) **Blum Blum Shub** [16] A cryptographically secure PRNG based on quadratic residues modulo a product of two large primes. While highly secure, it is computationally intensive and rarely used outside cryptographic applications.
9) **Multiply-with-Carry (MWC)** [17] Developed by *George Marsaglia*, MWC generators combine multiplication and carry operations to produce long-period sequences. They are simple to implement and perform well in statistical tests.
10) **Philox and Threefry** [18] Counter-based PRNGs designed for parallel computing environments. They use cryptographic primitives to ensure high-quality randomness and are optimized for GPU and multi-core CPU architectures.

C. Physical random numbers generators

Random Number Generators that rely on physical phenomena generate inherently unpredictable and provably random outcomes. These hardware-based RNGs, often referred to as true random number generators, operate by exploiting naturally occurring random processes, such as electronic noise, quantum effects, or chaotic systems. Unlike pseudo-random number generators, which use deterministic algorithms to produce sequences that only appear random, physical RNGs are typically standalone devices connected to a computer via interfaces like USB or PCI. However, physical RNGs are not without limitations; they often exhibit bias and short-range correlations, meaning the output may not always be uniformly distributed. Bias, defined as the deviation in the probabilities of generating ones and zeros, is quantified by the equation:

$$b = \frac{p(1) - p(0)}{2} \tag{2}$$

where $p(1)$ and $p(0)$ represent the probabilities of generating a one and a zero, respectively. To mitigate these issues, post-processing techniques, such as von Neumann debiasing or

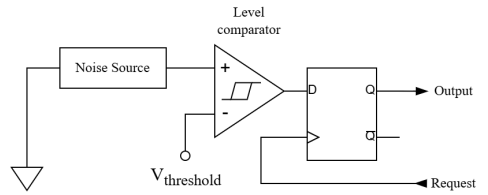


Fig. 2. The diagram shows a basic noise-based RNG comprising a noise source (e.g., thermal noise), a level comparator (with threshold $V_{threshold}$), and a D flip-flop (DFF) to sample and digitize the noise signal into random bits. This design leverages stochastic physical phenomena to generate true randomness.

cryptographic hash functions, are often applied to the raw output of physical RNGs. These methods help ensure the final output is both unbiased and suitable for applications requiring high-quality randomness, such as cryptography or secure communications.

Noise-based random number generators exploit stochastic physical phenomena to extract entropy for true randomness. These systems sample analog noise signals — unpredictable fluctuations inherent to electronic, optical, or quantum processes — and convert them into statistically independent bit sequences. Common noise sources include thermal agitation, quantum tunneling effects, and chaotic dynamics, each offering distinct advantages in entropy density and sampling rates [19], [7].

Thermal noise (Johnson-Nyquist Noise) arises from the random motion of electrons in conductive materials at equilibrium, as described by the fluctuation-dissipation theorem. Its power spectral density is given by:

$$S_V(f) = 4k_BTR \quad (\text{V}^2/\text{Hz}), \tag{3}$$

where k_B is Boltzmann's constant, T is the absolute temperature (K), and R is the resistance (Ω). This white noise exhibits a Gaussian amplitude distribution, making it a robust entropy source. Practical implementations sample thermal noise via high-gain amplifiers and analog-to-digital converters (ADCs), though challenges include mitigating $1/f$ noise at low frequencies and temperature-induced drift [7].

Zener noise (Avalanche Breakdown Noise) originates from quantum tunneling and impact ionization in reverse-biased Zener diodes operating near the avalanche breakdown regime. Individual electron-hole pair generation events produce discrete current spikes, yielding a shot noise process with Poissonian statistics. The noise current I_n follows:

$$\langle I_n^2 \rangle = 2qI_{\text{DC}}\Delta f, \tag{4}$$

where q is the electron charge, I_{DC} is the bias current, and Δf is the bandwidth. Zener-based RNGs achieve high bitrates (>100 Mbps) but require precise voltage regulation to stabilize the breakdown region and avoid deterministic oscillations [7].

Chaos-based random number generators leverage nonlinear dynamical systems governed by chaos theory to harvest entropy. These systems exhibit deterministic yet unpredictable behavior due to their sensitivity to initial conditions (the "butterfly effect") and topological mixing. Common implementations exploit chaotic phenomena in optical, electronic, optoelectronic, or mechanical domains, where small perturbations amplify exponentially over time, yielding outputs that are statistically indistinguishable from true randomness [21]. Other example of optical chaos-based RNGs employs distributed feedback (DFB) lasers with self-feedback mechanisms.

Free-Running Oscillators [7] (FROs) are a class of hardware-based random number generators that exploit the inherent jitter and metastability in electronic circuits to produce randomness. An FRO is a simple yet effective circuit that consists of a logical inverter whose output is fed back into its input, creating an unstable oscillatory state. The oscillation frequency is determined by the internal propagation delays of the inverter and the stray capacitances in the circuit. This unpredictability makes FROs a popular choice for generating random numbers in embedded systems and hardware security applications. The randomness in an FRO-based RNG is extracted by sampling the oscillating signal at irregular intervals or by comparing the phases of multiple FROs. One common approach is to use a pair of FROs with slightly different frequencies. The phase difference between the two oscillators drifts over time due to jitter, and this drift is sampled to produce random bits.

Quantum Random Number Generators (QRNGs) [8], [20] leverage the inherent unpredictability of quantum mechanics to generate true randomness. Unlike classical random number generators, which rely on deterministic algorithms or physical processes that may be predictable in principle, QRNGs exploit the fundamental indeterminism of quantum systems. This makes them ideal for applications requiring high levels of security, such as cryptography and secure communications. QRNGs operate by measuring quantum phenomena that are intrinsically random. One of the most common implementations uses photons, the elementary particles of light, and their quantum properties. For example, a photon in a superposition of states will collapse randomly into one of two possible states when measured. This randomness is used to generate binary bits (0s and 1s) [7], [8].

Spatial QRNGs [8], [7] rely on the physical path a photon takes after passing through a beam splitter. These systems use two detectors, one for each output port of the beam splitter. While conceptually simple, spatial QRNGs require careful calibration to ensure that the detectors are equally sensitive and that the beam splitter is perfectly aligned. Any imbalance in the system can introduce bias into the output, requiring post-processing to correct.

Temporal QRNGs [8], [7], on the other hand, exploit the randomness in the timing of photon arrivals. Instead of measuring the path a photon takes, these systems measure the time intervals between successive photon detections. For example, the system might compare two consecutive time intervals. This approach has the advantage of requiring only a single detector, simplifying the hardware design and reducing calibration requirements.

Another class of QRNGs exploits the randomness of radioactive decay [8]. In these systems, a radioactive source emits particles (e.g., alpha or beta particles) at random intervals. A detector registers the arrival of these particles, and the timing of the detections is used to generate random bits.

D. TRNG Type Comparison

The development and evaluation of TRNGs have been extensively studied in the literature. Quantum-based TRNGs, in particular, have gained significant attention due to their provable security based on the laws of quantum mechanics. Herrero-Collantes and Garcia-Escartin [8] provide a comprehensive review of quantum random number generators (QRNGs), highlighting their theoretical foundations and practical implementations. They emphasize the role of quantum phenomena, such as photon detection and vacuum fluctuations, in generating true randomness.

Cirauqui et al. [20] further explore the challenges and benchmarking of QRNGs, discussing the trade-offs between speed, security, and resource requirements. Their work underscores the importance of device-independent QRNGs, which offer enhanced security by minimizing assumptions about the underlying hardware.

Stipčević and Koç [7] provide a broader perspective on TRNGs, covering chaos-based, noise-based, and FRO-based designs. They discuss the sources of entropy, post-processing techniques, and security considerations for each type. Their work highlights the importance of environmental robustness and cost-effectiveness in practical TRNG implementations.

True Random Number Generators vary in design and application. Chaos-based TRNGs use nonlinear systems for high entropy but require complex calibration, making them suitable for cryptography. FRO-based TRNGs, leveraging electronic jitter, are simple and fast but require post-processing, which is ideal for IoT and FPGAs. Noise-based TRNGs rely on thermal or shot noise, offering reliability but requiring analog components, often used in hardware security modules (HMS) and military systems. Quantum-based TRNGs, exploiting quantum phenomena, provide ultimate security but are costly and sensitive, limiting their use to high-security applications. Each TRNG type balances performance, cost, and complexity for specific needs.

The choice of TRNG type depends on the specific application requirements:

- Chaos-based TRNGs are well-suited for cryptographic applications due to their high entropy and nonlinear dynamics. However, they are sensitive to environmental conditions and require careful calibration.
- FRO-based TRNGs are ideal for resource-constrained environments, such as IoT devices and FPGAs, due to their simplicity and low power consumption. However, they often require post-processing to improve randomness quality.
- Noise-based TRNGs offer a balance between reliability and security, making them suitable for hardware security modules and military systems. Their reliance on analog components increases the complexity of the design.
- Quantum-based TRNGs provide unparalleled security, making them ideal for high-stakes applications such as quantum communication and cryptography. However, their high cost and sensitivity to environmental factors limit their widespread adoption.

III. Implementation

A. Motivation

Randomness is seemingly underestimated, but it plays a key role in many areas: from simulations and Monte Carlo methods, through computer games, to statistical sampling and the generation of unique identifiers. The quality of random numbers directly affects the reliability of calculation results and system security, which is why generators based on real physical processes, rather than solely on deterministic algorithms, are increasingly being used. There is no common belief among programmers that it is necessary to use truly random generators. Popular libraries offer good-quality pseudo-random generators, but their deterministic nature makes it possible to reproduce the sequence if the initial state is known. True random generators, using the previously described thermal noise, quantum phenomena, or oscillator jitter, provide a higher degree of unpredictability. Building your own TRNG allows you to eliminate dependence on external services and gives you full control over the quality and availability of the entropy source.

The implementation of the true random number generator tool described below is a response to the needs of programmers. The basic access layer for this type of user is WebAPI, hence a lot of attention was paid to the preparation of the API and documentation when developing the solution. On the other hand, it is valuable to be able to test the solution in advance without having to use typical developer tools. In order to meet this requirement, a web interface was developed.

B. System architecture

The system must accommodate a moderate volume of simultaneous requests while maintaining low latency, high reliability, and scalability, ensuring robust performance under variable operational conditions. Additionally, the implementation will incorporate cryptographic protocols to secure data transmission and prevent adversarial manipulation during distribution.

Components presented in Fig. 3 work together in order to generate random numbers and serve them on the internet. The web app acts as a graphical interface for users. It displays the services available on the website as interactive forms that users use to send requests to the API and presents the results - the generated random numbers. The Hardware Random Number Generator is the source of random numbers. It generates random 32 bit words as quickly as possible and sends them over a WebSocket connection to the web API. The web API in pair with the database performs the back-end duties: converting random bits into numbers, processing user

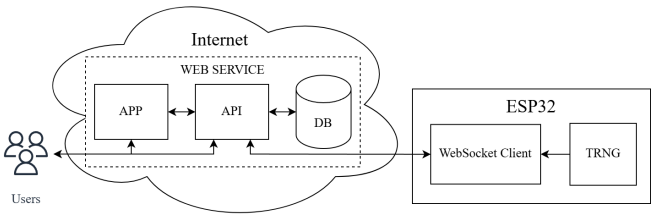


Fig. 3. The web service is a containerized application deployed in the cloud. The TRNG is a physical module connected to the internet. Users access the app through the internet.

requests, handling authorization, authentication and sending the results to users. The app, API and database components are deployed in a containerized environment in the cloud.

The development of the system utilizes a diverse set of tools and technologies tailored to each component. For the Web App, HTML structures web content, CSS handles styling, and JavaScript adds interactivity, while JQuery simplifies DOM manipulation and Bootstrap ensures responsive design. The Web API relies on Python and Django, with Django Rest Framework enabling RESTful API development, Open API providing standardization, and ASGI with Daphne supporting asynchronous communication. HTTP and WebSocket protocols facilitate seamless data exchange between the server and clients. The Database uses SQLite for its lightweight, high-reliability SQL engine, with SQL managing data manipulation and retrieval. For the TRNG, it use an IoT device [2], it means the ESP32 microcontroller with an embedded TRNG is programmed using the Arduino IDE and C++, ensuring efficient and secure random number generation. Together, these tools and technologies provide a robust foundation for the system's development. The ESP32 microcontroller was chosen because of its built-in Wi-Fi module (which simplified communication implementation), low purchase cost, and ease of programming in the Arduino environment. The IoT device can also be built on other microcontrollers that support Arduino programming, such as those from the AVR and NRF families.

C. Software

The web application's front end, which serves as the graphical interface accessible via a web browser, is integrated into the Python Django application using the Django templating engine. While the front end is primarily written using HTML, CSS, JavaScript and the Django Template Language (DTL), the Django templating engine dynamically renders the pages based on the context provided by the backend. This integration allows the front end to seamlessly interact with the Django framework.

The Web API is a distributed system designed to provide secure, scalable, and efficient access to random number generation services. It is composed of three interconnected applications, each responsible for a distinct functional domain. These applications work in concert to handle user authentication, random number generation and data delivery, ensuring a robust, modular architecture.

The Web API is organized into the following components:

1) **API (Application Programming Interface):** The core application responsible for handling user requests and delivering random number generation services. It implements the following modules:
   - **Views and Models:** Defines the endpoints and data structures for random tools, such as integer sequences, floating-point numbers, and custom distributions.
   - **Random Tool Functions:** Implements algorithms for transforming raw random bits into user-requested formats.
   - **OpenAPI Documentation:** Provides a standardized, machine readable specification of the API endpoints, request/response formats, and authentication mechanisms. This documentation is automatically generated using Redoc, ensuring clarity and consistency for developers integrating with the API. Available at https://trng.kacperochnik.eu/api/
2) **Auth (Authorization and Authentication Module):** A dedicated service for managing user identities, access control, and resource quotas. Key features include:
   - **User Authentication:** Implements Session and JWT-based authentication to verify user credentials and issue access tokens.
   - **Authorization:** Enforces role-based access control (RBAC) to restrict access to sensitive endpoints (e.g., administrative tools).
   - **Points System:** Tracks and deducts user points for each random number generation request, ensuring fair usage and preventing abuse.
3) **Trng (True Random Number Generator Interface):** A WebSocket-based service that interfaces with physical TRNG hardware to harvest entropy and deliver it to the API. Its components include:
   - **WebSocket Consumer:** Manages real-time connections with TRNG devices, enabling low-latency data transfer.
   - **Circular Buffer:** Stores incoming random bits in a fixed-size, ring-shaped buffer to ensure continuous availability and prevent data loss during high-throughput operations.
   - **Bit Disposal Interface:** Provides a standardized interface for delivering random bits to the API's random tool functions, ensuring compatibility with diverse TRNG hardware.

The system employs SQLite as its relational database management system (RDBMS), leveraging Django's Object-Relational Mapping (ORM) layer for data persistence operations. SQLite was selected for its minimal configuration requirements, serverless architecture, and atomic transaction support - characteristics that align with the application's need for simplicity and rapid development iteration.

D. Hardware

The Hardware Random Number Generator (HRNG) subsystem comprises an ESP32-WROOM-32D microcontroller configured as a dedicated entropy source. This IoT-enabled device leverages the chip's native True Random Number Generator peripheral to harvest physical noise, transmitting raw random data to the web API via secure WebSocket protocol. The complete design files and firmware are available in the project repository.[1]

The ESP32 incorporates a True Random Number Generator that generates 32-bit random numbers suitable for cryptographic operations [23]. The TRNG relies on physical entropy sources so that no number within the specified range is more or less likely to appear than any other. In terms of key characteristics, the generator obtains its entropy from thermal noise produced by the high-speed ADC and SAR ADC, as well as from an asynchronous clock mismatch provided by the RC_FAST_CLK (8 MHz internal RC oscillator). Its output rate can theoretically reach up to 5.2 Mbps, however, to ensure optimal performance, the recommended read rate is 500 kHz when using SAR ADC noise and 5 MHz when using high-speed ADC noise. The TRNG is accessed via the 'RNG_DATA_REG' memory-mapped I/O (MMIO) register at the address '0x3FF75144', and it incorporates built-in Von Neumann debiasing with a 2:1 bit compression ratio to remove bias from the raw entropy stream.

Functionally, the TRNG generates true random numbers based on two primary sources of entropy. The first source, Thermal Noise, is generated by both the high-speed ADC and the SAR ADC, when enabled, these ADCs produce bit streams that serve as random seeds for the TRNG. The second source, Asynchronous Clock Mismatch, is derived from the RC_FAST_CLK (8 MHz internal RC oscillator), which introduces timing jitter and clock drift. These sources are combined using XOR logic, as expressed in the following equation:

$$b_{out} = \bigoplus_{i=0}^{n} (s_i \oplus t_i) \tag{5}$$

where:

- s_i represents sampled oscillator states (e.g., from the SAR ADC or high-speed ADC).
- t_i represents timing jitter measurements (e.g., from the RC_FAST_CLK).

Regarding the entropy feed mechanism, the TRNG uses two distinct noise sources. SAR ADC Noise provides 2 bits of entropy per clock cycle of the RC_FAST_CLK (8 MHz), with a maximum recommended read rate of 500 kHz to ensure maximum entropy. In contrast, High-Speed ADC Noise offers 2 bits of entropy per APB clock cycle (typically 80 MHz) and has a maximum recommended read rate of 5 MHz to similarly ensure optimal entropy extraction.

The Arduino program is fairly simple, it sends generated 32 bit random words, over WebSocket, as fast as possible, in a loop. It uses esp32 random API function - *esp_random()* to retrieve random bits from the register. Source code are presented in example 3.1.

[1]Repository: https://github.com/TeriyakiGod/esp-arduino-trng

Example 3.1 (C++):

```
void loop() {
  webSocket.loop();// Maintain WebSocket conn.
  // Check if WiFi is still connected
  if (WiFi.status() != WL_CONNECTED) {
    Serial.println("WiFi lost connection");
    connectToWiFi();
  }
  // Gen. rnd. number and send over websocket
  uint32_t randomValue = esp_random();
  webSocket.sendBIN((uint8_t*)&randomValue(),
  sizeof(randomValue));
}
```

E. Security

The developed solution belongs to the class of IoT devices and is constantly connected to the Internet, so adequate security is required [2]. The Web API incorporates several integrated mechanisms to ensure both security and scalability. It encrypts all communications between clients and the API using TLS 1.3, which protects against eavesdropping and man-in-the-middle attacks. In addition, the API employs rate limiting with token bucket algorithms to throttle excessive requests, thereby preventing denial-of-service (DoS) [4] attacks and ensuring fair distribution of resources. Load balancing is also applied, as incoming requests are distributed across multiple API instances via round-robin or least-connections strategies to maintain high availability and fault tolerance. Complementing these measures, robust authentication is achieved through the use of session cookies (maintained via a 'sessionid') and API tokens supplied in the 'Authorization' header, while standard error responses—such as '403 Forbidden' for insufficient points or privileges, '503 Service Unavailable' when the TRNG is offline, and '429 Too Many Requests' when the rate limit is exceeded—provide clear feedback to users.

The device further implements a custom binary protocol over WebSocket Secure (WSS) that is optimized for both efficiency and security. This protocol defines its frame format using 32-bit little-endian unsigned integers and supports a transmission rate of around 20 KB/s. Error handling is addressed by incorporating an exponential backoff strategy for connection failures, and security is reinforced by using TLS 1.3 in conjunction with the ESP32's root certificate bundle. Moreover, additional security considerations include physical isolation through a Faraday-shielded enclosure, runtime protection ensured by secure boot with flash encryption, and enhanced transport security via Perfect Forward Secrecy (PFS) using ECDHE-RSA. Together, these strategies result in an implementation that achieves 128-bit security under the Dolev-Yao threat model, provided that there is no physical access to the TRNG unit.

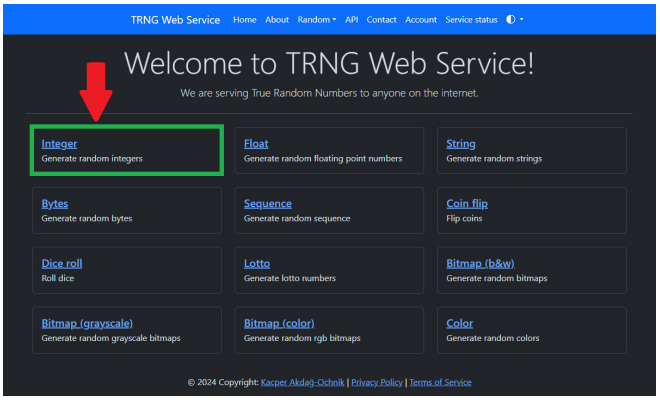


Fig. 4. Homepage with the Integer Generator button highlighted

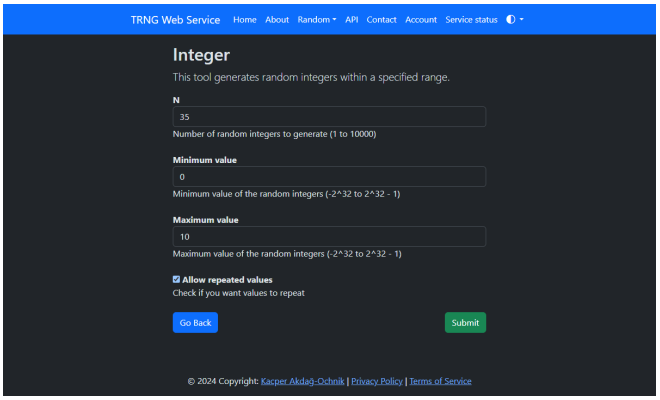


Fig. 5. Integer generator form filled with values.

IV. Verification

A. Access via GUI and API

Users begin by visiting the web service through a browser, where the homepage presents a selection of random number generators. To generate random integers, the user navigates to the integer generator using a button on the homepage, as shown in Fig. 4. Once on the integer generator page, the user can configure the tool to their preferences. For example, to generate 35 integers in the range from 0 to 10, the user sets N (the number of random integers) to 35, the minimum value to 0, and the maximum value to 10, as illustrated in Fig. 5. After submitting the form, the application returns the generated random integers. The distribution of these integers is visualized in a histogram in Fig. 6, showing a roughly uniform spread of random values across the specified range.

For programmatic access, users can send a GET request to the API endpoint[2]. The API responds with a JSON object containing the generated random integers, a timestamp, and the number of bits used, as shown in the code listing 4.1. The histogram in Fig. 7 visualizes the distribution of the integers generated via the API.

[2]https://serveraddess/api/rand/int?n=35&min=0&max=10&repeat=True

Example 4.1 (JSON):


```
"values":
    [4, 8, 2, 9, 8, 9, 10, 10, 5, 3, 1, 0,
    10, 6, 10, 3, 6, 10, 7, 1, 10, 3, 4,
    2, 10, 6, 8, 5, 3, 0, 9, 0, 0, 6, 7],
"timestamp": "2025-03-20T17:14:12.492409",
"bits": 1120
```

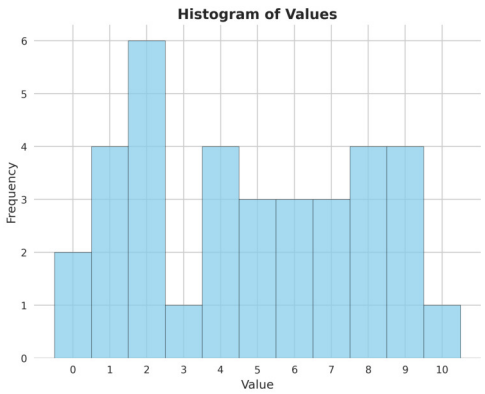



Fig. 6. Histogram of the first result

The application provides a seamless experience for generating random numbers, whether through the intuitive UI or the programmatic API. The histograms in Figs. 6 and 7 demonstrate that the generated integers are roughly uniformly distributed and random each time, validating the effectiveness and reliability of the random number generator. This ensures that users can trust the application to produce unbiased and balanced results for their needs. All functions are listed in the V appendix section.

B. Statistical Testing with Dieharder

The randomness of the TRNG output was evaluated using the Dieharder test suite, a comprehensive collection of statistical tests designed to assess the quality of random number generators. Dieharder analyzes the output for patterns,

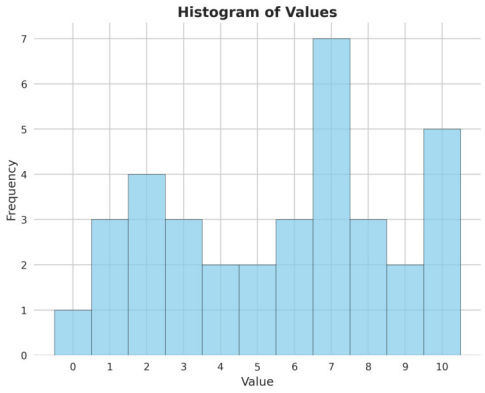


Fig. 7. Histogram of the second result

TABLE I
DIEHARDER TEST RESULTS. ALL OF THE TESTS PASSED, ONLY 2 WEAK RESULTS: RGB_BITDIST AND RGB_LAGGED_SUM.

Test Name	ntup	tsamples	psample	p-value
diehard_birthdays	0	100	100	0.01192945
diehard_operm5	0	1000000	100	0.67160533
diehard_rank_32x32	0	40000	100	0.88503575
diehard_rank_6x8	0	100000	100	0.44944912
diehard_bitstream	0	2097152	100	0.07796698
diehard_opso	0	2097152	100	0.75986760
diehard_oqso	0	2097152	100	0.52389989
diehard_dna	0	2097152	100	0.63689771
diehard_count_1s_str	0	256000	100	0.64745948
diehard_count_1s_byt	0	256000	100	0.48743869
diehard_parking_lot	0	12000	100	0.06943192
diehard_2dsphere	2	8000	100	0.33036560
diehard_3dsphere	3	4000	100	0.69729265
diehard_squeeze	0	100000	100	0.41006273
diehard_sums	0	100	100	0.22268815
diehard_runs	0	100000	100	0.87899505
diehard_craps	0	200000	100	0.01235318
marsaglia_tsang_gcd	0	10000000	100	0.47170575
sts_monobit	1	100000	100	0.72827629
sts_runs	2	100000	100	0.52336996
sts_serial	1	100000	100	0.03008230
rgb_bitdist	6	100000	100	0.00048266
rgb_min_distance	2	10000	1000	0.16471595
rgb_permutations	2	100000	100	0.26856586
rgb_lagged_sum	24	1000000	100	0.99977196
rgb_kstest_test	0	10000	1000	0.27646420
dab_bytedistrib	0	51200000	1	0.52014772
dab_dct	256	50000	1	0.83006246
dab_filltree	32	15000000	1	0.79259371
dab_monobit2	12	65000000	1	0.12926929

biases, and correlations that could indicate deviations from true randomness. The TRNG output was subjected to a variety of tests, and the results are presented in table I.

The Dieharder test results confirm that the TRNG output behaves as expected for a high-quality random sequence. Most tests passed with p-values within acceptable confidence intervals, indicating no significant biases or correlations. However, a few tests, such as rgb_bitdist (for bit distances 4 and 6) and rgb_lagged_sum (for lag 24), returned weak results. The weak result of the two tests is caused by insufficient variability in the sequences of consecutively generated zeros or ones. This may be due to suboptimal threshold alignment in the thermal noise signal digitization circuit, leading to correlation between consecutive samples.

C. Possible applications

The proposed self-hosted true random number generation system addresses diverse applications requiring high-quality, statistically robust, and verifiably unbiased randomness. The scientific field can make use of the web service in many ways, eg. Monte Carlo Simulations, Diffusion Models and Statistical Research.

Large-scale computational models in fields such as quantum chemistry, particle physics, and financial mathematics rely on entropy-rich seeds to minimize correlation artifacts. For instance, lattice Boltzmann simulations of fluid dynamics and option pricing models in quantitative finance demand randomness with certified uniformity to ensure convergence accuracy. The system's post-processing algorithms, which suppress residual biases in raw entropy sources, enhance the reproducibility of results across distributed computational clusters.

Modern generative artificial intelligence frameworks, including stochastic differential equation (SDE)-based diffusion models, utilize random noise sequences to synthesize high-fidelity images or molecular structures. A TRNG-backed service guarantees that the latent space sampling process remains free from deterministic patterns, which could otherwise introduce perceptual biases or reduce generative diversity in outputs.

Randomized controlled trials (RCTs), bootstrap sampling, and permutation tests require auditable randomness to preserve methodological integrity. By providing cryptographically signed random streams, the system enables researchers to verify the provenance of experimental data, mitigating concerns about inadvertent selection biases or adversarial tampering in peer-reviewed studies.

There are multiple applications in the field of gaming.

1) **Multiplayer Games:** Massively multiplayer online (MMO) platforms and blockchain-based play-to-earn ecosystems necessitate transparent randomness for procedural content generation, loot box mechanics, and non-fungible token (NFT) attribute assignment. Centralized pseudorandom number generators often face skepticism from users; outsourcing to an independently audited TRNG service enhances player trust and regulatory compliance.
2) **Lotteries and Raffles:** Public or decentralized systems conducting randomized prize allocations require algorithmic transparency to demonstrate impartiality. Integrating the TRNG system with blockchain-based smart contracts or verifiable computation frameworks ensures that draw outcomes derive from non-deterministic physical entropy, enabling participants to independently audit the randomness generation process. This eliminates skepticism toward algorithmic fairness, particularly in decentralized or distributed systems where trust in centralized authorities is limited.
3) **Casino Gaming:** Digital gaming platforms depend on unpredictability for mechanics such as shuffling virtual decks, generating roulette ball trajectories, or determining slot machine payouts. Hardware-backed TRNGs mitigate vulnerabilities inherent to algorithmic pseudorandomness—such as state reconstruction or seed manipulation—by anchoring outcomes to physically irreproducible entropy sources. This ensures that game mechanics cannot be reverse-engineered or biased, even by privileged insiders with access to the software stack.

The web service can be applied to create art.

1) **Generative Art:** Algorithmic art platforms (e.g., Art Blocks) and fractal geometry engines depend on entropy to drive procedural variations. High-dimensional randomness ensures that procedurally generated artworks—such as those minted as NFTs—exhibit unique traits without deterministic repetition, thereby preserving artistic value and collector interest.
2) **Algorithmic Music:** Stochastic composition tools, such as Markov chain-based melody generators or granular synthesis systems, leverage random parameters to explore unconventional harmonic structures. TRNG-derived sequences facilitate organic, non-repetitive soundscapes, distinguishing outputs from those produced by conventional PRNGs with limited cycle lengths.
3) **Interactive Installations:** Immersive environments and augmented reality (AR) experiences use real-time randomness to adapt narrative pathways or visual elements based on user interactions. A low-latency TRNG service enables dynamic, audience-responsive installations while maintaining statistical unpredictability across prolonged exhibitions.

While the system is unsuitable for high-throughput cryptographic operations like TLS handshakes (handling a single request takes almost 380 ms, for example from section IVa), it may supplement key derivation for non-real-time applications, such as creating master keys for offline data vaults and random vectors [1]. However, dedicated hardware security modules remain preferable for latency-sensitive tasks.

However, in production use, the scalability issue can be solved by duplicating instances through containerization and implementing a load balancer. Compared to competing solutions, it is worth highlighting the open-source approach and documentation of all endpoints. In addition, the proposed solution meets functional security requirements, including TLS and JWC.

V. Summary

Generating true random numbers is often underestimated by programmers. However, it turns out that a great deal depends on it, and it should be as random as possible. The paper demonstrates a software and hardware solution for generating such numbers. The IoT device developed uses noise, which serves as the basis for the generation. The generator's functions are then accessed via an API and GUI on a website.

The randomness of the generated numbers was verified using automatic tests, manual tests (histograms presented earlier) and a dedicated randomness verification solution, Dieharder. The results obtained show that the whole in the sense as a system generates data with a high level of randomness. Functional tests confirm that the system as a whole generates and provides numbers in accordance with user requirements and handles errors correctly. Performance tests show that the selected architecture (cloud containers, WebSocket, rate limiting) ensures low latency and high scalability. Statistical tests show that bit strings from TRNG do not exhibit significant deviations from randomness: no bias, appropriate distribution, no autocorrelation. As a result, they can be considered "truly random."

Obtaining a high level of randomness makes it possible to excavate the solution in many areas of IT. However, it should be pointed that the speed of data generation is average (less than 3 requests per second), so it cannot be used in real time. The solution to this problem is to multiply the IoT device, which, thanks to the use of ESP32, will not be expensive. Other risks arise from the analog circuit and the analog-to-digital converter. There is a risk that after generating a strong and steady signal, the system will receive the same grain all the time, hence work on improving the analog part is being considered. Other developments are related to expanding the software layer to generate random data for sharpening in artificial intelligence tools. Future work also could involve testing larger datasets to ensure the TRNG's performance scales appropriately. Additionally, other statistical test suites, such as the NIST Statistical Test Suite [22] or TestU01, could be used to further validate the TRNG's randomness.

Appendix

The system performs following functions:

1) Generating random numbers:
   a) Generating n random integers in range $[a, b]$.
   b) Generating n random floating point numbers in range $[0, 1]$ with p precision.
   c) Generating n random bytes represented in selected positional numeral system:
      i) Binary
      ii) Octal
      iii) Hexadecimal
      iv) Decimal
   d) Generating n random strings m long, with options:
      i) With repeating characters.
      ii) Without repeating characters.
      iii) With character set selection:
         A) Letters
         B) Digits
         C) Special characters
   e) Generating a random permutation of integers in range $[a, b]$
   f) Generating n coin tosses.
   g) Generating n, m sided dice throws.
   h) Generating n lottery tickets.
   i) Generating a random bitmap with dimensions x by y in PNG format, with the following types:
      i) Monochrome (1 bit per pixel)
      ii) Grayscale (8 bits per pixel)
      iii) Color (24 bits per pixel)
   j) Generating n random HEX color codes.
2) Ability to communicate with True Random Number Generator Modules using WebSocket protocol.

3) Ability to exchange data with users using HTTP protocol.
4) Encrypted connections using SSL/TLS.
5) Access to the service through the Web App or Web API.
6) Access control.
7) Anonymous user access with IP registry.
8) Web API keys for registered users.
9) Credit system allowing users to exchange points for random numbers.

REFERENCES

[1] Rak, Tomasz, and Dariusz Rzońca. 2024. "Security and Privacy in Networks and Multimedia" Electronics 13, no. 15: 2887. doi: https://doi.org/10.3390/electronics13152887

[2] Felkner, Anna, Jan Adamski, Jakub Koman, Marcin Rytel, Marek Janiszewski, Piotr Lewandowski, Rafał Pachnia, and Wojciech Nowakowski. 2024. "Vulnerability and Attack Repository for IoT: Addressing Challenges and Opportunities in Internet of Things Vulnerability Databases" Applied Sciences 14, no. 22: 10513. doi: https://doi.org/10.3390/app142210513

[3] Łukasz Sobaszek and Arkadiusz Gola and Edward Kozłowski, Application of survival function in robust scheduling of production jobs, editors M. Ganzha and L. Maciaszek and M. Paprzycki, Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, vol. 11, 2017, doi: 10.15439/2017F276

[4] Daniel Czyczyn-Egird, Rafał Wojszczyk, "The effectiveness of data mining techniques in the detection of DDoS attacks". In: Omatu, S., Rodríguez, S., Villarrubia, G., Faria, P., Sitek, P., Prieto, J. (eds) Distributed Computing and Artificial Intelligence, 14th International Conference. DCAI 2017. Advances in Intelligent Systems and Computing, vol 620. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-62410-5_7

[5] T. Nowicki, K. Chlebicki, D. Pierzchała, R. Waszkowski and K. Worwa, "Simulation method of reliability evaluation for RFID based restricted access administrative office," 2017 IEEE International Conference on RFID Technology & Application (RFID-TA), Warsaw, Poland, 2017, pp. 89-94, doi: 10.1109/RFID-TA.2017.8098877.

[6] Wolfgang Hörmann, Josef Leydold, and Gerhard Derflinger. Automatic nonuniform random variate generation. Springer Science & Business Media, 2013.

[7] Mario Stipčević and Çetin Koç. "True Random Number Generators". In: Nov. 2014, pp. 275–315. isbn: 978-3-319-10682-3. doi: 10.1007/978-3-31910683-0_12.

[8] Miguel Herrero-Collantes and Juan Carlos Garcia-Escartin. "Quantum random number generators". In: Reviews of Modern Physics 89.1 (Feb. 2017). issn: 1539-0756. doi: 10.1103/revmodphys.89.015004. url: http://dx.doi.org/10.1103/RevModPhys.89.015004

[9] John von Neumann. "Various techniques used in connection with random digits". In: Monte Carlo Method. Ed. by A. S. Householder, G. E. Forsythe, and H. H. Germond. Vol. 12. National Bureau of Standards Applied Mathematics Series. Washington, D.C.: U.S. Government Printing Office, 1951, pp. 36–38

[10] Derrick H. Lehmer. "Mathematical methods in large-scale computing units". In: Proceedings of 2nd Symposium on Large-Scale Digital Calculating Machinery. 1951, pp. 141–146.

[11] Makoto Matsumoto and Takuji Nishimura. "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator". In: ACM Trans. Model. Comput. Simul. 8.1 (Jan. 1998), pp. 3–30. issn: 1049-3301. doi: 10.1145/272991.272995..

[12] George Marsaglia. "Xorshift RNGs". In: Journal of Statistical Software 8.14 (2003), pp. 1–6. doi: 10.18637/jss.v008.i14.

[13] François O Panneton, Pierre l'Ecuyer, and Makoto Matsumoto. "Improved long-period generators based on linear recurrences modulo 2". In: ACM Transactions on Mathematical Software 32.1 (2006), pp. 1–16. doi: 10.1145/1132973.1132974.

[14] Melissa E. O'Neill. PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation. Technical Report HMCCS-2014-0905. Harvey Mudd College, Sept. 2014.

[15] L'ubomíra Balková et al. "Aperiodic pseudorandom number generators based on infinite words". In: Theoretical Computer Science 647 (Sept. 2016), pp. 85 - 100. issn: 0304-3975. doi: 10.1016/j.tcs.2016.07.042.

[16] Lenore Blum, Manuel Blum, and Michael Shub. "A Simple Unpredictable Pseudo-Random Number Generator". In: SIAM Journal on Computing 15.2 (1986), pp. 364–383. doi: 10.1137/0215025.

[17] George Marsaglia. "Random number generators". In: Journal of Modern Applied Statistical Methods 2.1 (May 2003), pp. 2–13. doi: 10.22237/jmasm/1051747320.

[18] John Salmon et al. "Parallel random numbers: as easy as 1, 2, 3". In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. 2011, Article No. 16. doi: 10.1145/2063384.2063405.

[19] C. D. Motchenbacher and J. A. Connelly. Low-noise electronic system design. Wiley Interscience, 1993. isbn: 0-471-57742-1.

[20] David Cirauqui and Miguel Ángel García-March and Guillem Guigó Corominas and Tobias Graß and Przemysław R. Grzybowski and Gorka Muñoz-Gil and J. R. M. Saavedra and Maciej Lewenstein. Quantum Random Number Generators : Benchmarking and Challenges. 2022.

[21] X. Li et al. "Scalable parallel physical random number generator based on a superluminescent LED". In: Opt. Lett. 36 (2011), pp. 1020–1022.

[22] Darren Hurley-Smith, Julio Hernandez-CastroJ, ulio Hernandez-Castro, Bias in the TRNG of the Mifare DESFire EV1, RFIDSec 2016

[23] Espressif Systems. ESP32 Technical Reference Manual. Accessed: 2023-10-01. 2023. url: https://www.espressif.com/sites/default/files/documentation/esp32_technical_reference_manual_en.pdf\#rng.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 625–634
DOI: 10.15439/2025F1656
ISSN 2300-5963 ACSIS, Vol. 43

The price of customer presence in Attended Home Delivery with Customer Availability Profiles

Roberto Zanotti
0000-0002-3073-4895
Dept. of Clinical and Experimental Sciences,
Università degli Studi di Brescia,
viale Europa 11, 25123 Brescia, Italy
E-mail: roberto.zanotti@unibs.it

Daniele Manerba, Renata Mansini
0000-0002-3502-5289, 0000-0002-2194-0339
Dept. of Information Engineering,
Università degli Studi di Brescia,
via Branze 38, 25123 Brescia, Italy
E-mails: {daniele.manerba, renata.mansini}@unibs.it

***Abstract*—Attended Home Delivery is a last-mile distribution paradigm in which the customer must be present at home to receive the goods in person. In this context, we study a vehicle routing and scheduling problem in which the customers' availability is given as a time-dependent probability profile, and the company incurs a penalty cost proportional to the probability of not finding the customer at home during the selected timeslot for delivery. Using an efficient Mixed-Integer Linear Programming formulation for the problem as a black-box tool and lexicographic optimization procedures, we develop an economic analysis to support the company in exploiting the trade-off between basic optimization and the possibility of increasing the customer presence probabilities by paying additional costs. Managerial insights are derived from the change in the value and structure of optimal solutions under different budget levels allocated to improving customer availability profiles.**

I. Introduction and literature review

SINCE the advent of e-commerce, the delivery of goods purchased online directly to customers' homes has rapidly become the leading business model for retail companies and, in turn, one of the most important logistics operations within urban areas, and beyond. The largest share of deliveries performed in e-commerce consists of parcels, small packages, food, and groceries. The volume of this business is impressive and continues to grow rapidly. Over the past 10-12 years, the global e-commerce market has grown tenfold, reaching a value of 3.3 trillion dollars in 2022, approximately 22% of total retail sales. The COVID-19 pandemic significantly boosted this trend: the share of online retail rose from 15% in 2019 to 21% in 2021, exceeding even the most optimistic forecasts (see, e.g., [11]). Experts now agree that the sector still holds considerable long-term growth potential, with projections estimating it could reach 5.4 trillion dollars within the next 3–4 years [21].

Despite being very convenient (and sometimes necessary) for customers, home delivery services pose significant logistical challenges for companies and municipalities, in particular concerning last-mile deliveries. Last-mile delivery refers to the final stage of the distribution process, in which goods are transported from a local hub or depot to the end customer's address. From an economic point of view, last-mile operations represent by far the most expensive part of the fulfillment chain, accounting for more than 50% of the total shipping costs. On average, companies charge the consumers up to 80% of the shipping cost per package to cover them, taking the rest from the profit margins of sold products. Given the costs and the high impact in terms of emissions [8][15] and social implications [19], there is a considerable amount of literature devoted to the evaluation of alternative ways of handling the last mile, from parcel lockers [5] to drones [13], from autonomous robots [7] to crowd-shipping [2][3][12][22]. However, home delivery is one of the predominant delivery paradigms in the industry, whether because it is necessary given the types of goods being purchased (grocery, high-value items) or because customers perceive value when goods are delivered directly to their home. In this paper, we focus on a specific implementation of home delivery operations known as Attended Home Delivery (AHD). In the AHD setting, unlike standard home deliveries, the presence of the customer at home is required when the courier arrives. This requirement prominently brings the time dimension into the delivery operations, since couriers and customers must be synchronized with each other.

From a practical perspective, delivery synchronization is typically managed at the time of ordering by collecting not only the customer's location, but also a *time window* (*timeslot*) on a specific day for the expected delivery. While, in principle, the width of a customer's service time window can be arbitrarily chosen, it is intuitive that narrower time windows result in higher delivery costs for companies [18], which are often passed on to customers through additional delivery fees. Moreover, the delivery company needs to maintain a certain level of control over the time windows schedule available for its services, i.e., it is not realistic to leave an arbitrary choice to the online customer. For this reason, AHD services are often based on a discrete set of available time windows from which customers can select their favorite one. This way, the company still provides customers with a certain degree of flexibility while allowing itself the possibility of finding, at a tactical level, the best schedule to propose. An active stream

This work has been supported by "ULTRAOPTYMAL - Urban Logistics and sustainable TRAnsportation: OPtimization under uncertainTY and MAchine Learning", a PRIN2020 project funded by the Italian University and Research Ministry (grant number 20207C8T9M).

in the literature has addressed the optimization of this AHD tactical aspect, using approximated or explicit evaluation of the corresponding routing costs (see, e.g., [1][9][23]).

A recent alternative to use multiple time windows (thus avoiding the need to solve tactical timeslot assignment and pricing problems) is the introduction of Customer Availability Profiles (CAPs), presented in [10]. After partitioning a working day into timeslots, a CAP assigns to each timeslot the probability that the customer will be at home. This framework grants the company full flexibility in planning delivery routes and schedules, at the cost of potentially incurring a penalty when a delivery is attempted in a timeslot during which the customer is actually not at home. For the company, such a penalty may represent an estimation of the costs to sustain for the recovery of the service missed (e.g., re-delivery at a different time or day) and/or a compensation to the customer for having lowered the quality of their service. Since the customer presence in a timeslot is not deterministically known a-priori but just estimated through a probability, then it makes sense to calculate the penalty as proportional to the probability of not finding the customer at home in the timeslot selected by the company.

AHD optimization problems that explicitly include CAPs are more tailored to support a company in the decision-making process at a day-ahead tactical level (see, e.g., the multi-attribute approach proposed in [4]). They can better assess the trade-off that derives from simultaneously minimizing the traveling cost and maximizing the so-called service *hit rate* (equivalent to minimizing the corresponding penalty cost). The total hit-rate cost reflects the quality level of the delivery service and should be considered by the company together with the operating costs. However, in practice, the company may adopt additional strategies, such as offering discounted fees or other forms of compensation to encourage customers to be available during timeslots that are operationally desirable but associated with low presence probabilities. To the best of our knowledge, this integrated approach, where CAPs are treated with a certain degree of flexibility, has not yet been explored in the context of AHD problems. Hence, in this work, by developing mathematical tools including Mixed-Integer Linear Programming (MILP) models and tailored lexicographic solution methodologies, we provide a detailed economic analysis on the trade-offs related to investing in customers' behavior modification along with classical routing and scheduling decisions.

A. Contribution

Building on the current state of the art, this paper provides three main contributions:

- We propose a novel integrated problem in the AHD context, along with its mathematical formulation, that extends existing settings by allowing the company to actively invest a budget to improve customer presence probabilities. In this framework, availability profiles are no longer treated as fixed inputs but as strategic decision variables within the optimization process.
- We develop a multi-step analytical methodology based on lexicographic optimization that enables a structured economic analysis of the trade-offs involved, first by determining a maximum budget for influencing customer availability and then by systematically evaluating the impact of investing partial amounts of that budget.
- We conduct an extensive computational analysis on a large set of instances to derive actionable managerial insights. In particular, we quantify the strategic trade-off in which the company accepts higher penalties on some deliveries to obtain significant savings in travel costs, making the *price of customer presence* a clearly identifiable and decision-relevant concept.

B. Structure of the paper

The remainder of the paper is organized as follows. Section II introduces the optimization problem and provides its mathematical formulation. Section III describes the improved decisional setting in which the company has additional flexibility regarding the CAPs. Section IV details our experimental campaign, describing both the adopted methodology and the instances used, while Section V shows the results obtained and comments on them, deriving managerial insights on the process. Conclusions are drawn in Section VI.

II. Problem statement and mathematical formulation

In this section, we state the optimization problem addressed in this work by building upon an analogous framework presented in the literature [6]. We then formulate the problem as a Mixed-Integer Linear Programming (MILP) model.

A. Problem definition

We consider the problem of addressing daily delivery operations for a set P of geographically dispersed customers. The deliveries are performed by a uniform fleet of vehicles, $K = \{1, \dots, \lambda_K\}$, which originates from and returns to a central depot $\mathcal{D}$. Each vehicle has a uniform capacity, Q, and all routes must be completed within a maximum operational period, t_{max}. For each customer $p \in P$, there is a specific demand d_p, and a constant service time s is required for every stop. The road network defines the feasible routes along which vehicles can travel, both from the depot to each customer and between any two customer locations. Let t_{ij} denote the travel time between any two locations i and j, where each location may be either the depot or a customer. To incorporate customer availability profiles, the planning horizon is discretized into a set of sequential, non-overlapping timeslots, $T = \{1, \dots, \lambda_T\}$. Each timeslot $t \in T$ is defined by a start time a_t and an end time b_t. Let ρ_{pt} represent the likelihood that customer p is available at home during timeslot t. A key assumption is that $0 < \rho_{pt} < 1$ for all customers and timeslots, ruling out complete certainty of presence or absence. If a delivery attempt fails, a penalty is incurred. This penalty is calculated as the product of a failure cost parameter α_p, potentially different for each customer $p \in P$, multiplied by the probability $(1 - \rho_{pt})$

of the customer p not being home during the delivery during timeslot $t \in T$. The failure costs are assumed to be calculated in time units, thus representing the additional time spent by the courier to recover the service (e.g., by scheduling a second visit or by moving the goods to a collecting point) and/or the additional time spent by the customer to retrieve their parcel (e.g., roundtrip from the house to the collecting points where the goods have been left), which must be compensated by the company as well. Finally, while in many operational settings the failure costs can be seen as constant for all the customers, in this work we prefer to consider the general case in which they depend on the customer (e.g., on their distance to the depot, fidelity program, or delivery priority). The optimization objective is to determine a set of vehicle routes, along with their corresponding delivery schedules, that minimizes a composite cost function. This function aggregates the total vehicle travel costs with the overall penalty cost arising from potential synchronization failures (i.e., missed deliveries).

Unlike some prior studies that incorporate numerous operational details (see [6][10][24]), our model focuses solely on the core features of the optimization problem to keep the analysis clear and concise. For this reason, some simplifying assumptions are made concerning, e.g., the homogeneity of the vehicle fleet, the constant service time for all the customers, and the presence of a unique depot where all the vehicles depart and come back. Note, however, that these assumptions are realistic in many AHD services, as the e-grocery ones.

B. Problem formulation

The problem is modeled on an *extended* directed graph $G = (V, A)$. The ordered node set V is constructed from three distinct subsets:

- the ordered set $N_P = \{1, \ldots, \lambda_N\}$ of λ_N nodes, each representing a customer.
- the ordered set $N_S = \{\lambda_N + 1, \ldots, \lambda_N + \lambda_K\}$ of λ_K nodes, each being a replica of the unique depot $\mathcal{D}$. This duplication allows us to have one starting node per vehicle, thereby distinguishing each vehicle's route without the need to introduce additional vehicle-indexed arc variables.
- a single node $e = \lambda_N + \lambda_K + 1$, representing the common terminal node for all the routes of the vehicles. This node is a further replica of the unique depot $\mathcal{D}$, and its presence allows us to model vehicle routes as elementary paths.

The arc set A is defined as the set of all feasible connections, i.e., $A = \{(i,j) : i \in N_P, j \in N_P\} \cup \{(i,j) : i \in N_S, j \in N_P\} \cup \{(i,e) : i \in N_P\}$. A weight corresponding to the travel time t_{ij} is assigned to each arc $(i,j) \in A$. For those arcs starting or arriving in a replica of the depot, the corresponding travel time is replicated as well. Given the above graph, the problem asks for a set of λ_K directed elementary paths, one for each vehicle $k \in K$, which originate at the corresponding starting depot node $\lambda_N + k$ and terminate at e. Figure 1 shows an example of the creation of our extended graph starting from a standard network. The total travel time of a vehicle is the sum of the weights of the arcs along its path, and the total travel cost is the aggregate travel time of all vehicles.

The following decision variables are defined to model the problem:

- x_{ij}: a binary variable that is equal to 1 if the direct arc from node i to node j is part of a vehicle's route, and 0 otherwise;
- z_{ij}: a continuous variable denoting the time of arrival at node j, when it is reached directly from node i via arc (i,j);
- y_{it}^k: a binary variable that is equal to 1 if vehicle k services node i during timeslot t, and 0 otherwise.

Then, the problem under study can be efficiently formulated through the following MILP model, in which, given a subset of nodes $\bar{V} \subset V$, $\delta^-(\bar{V}) = \{(i,j) \in A : i \notin \bar{V}, j \in \bar{V}\}$ and $\delta^+(\bar{V}) = \{(i,j) \in A : i \in \bar{V}, j \notin \bar{V}\}$ represent the set of arcs entering and leaving the set of nodes $\bar{V}$, respectively:

$$\min \sum_{(i,j)\in\delta^-(\{e\})} z_{ij} + \sum_{p\in P}\sum_{t\in T}\sum_{k\in K} \alpha_p\,(1-\rho_{pt})\,y_{pt}^k \quad (1)$$

subject to

$$\sum_{(i,j)\in\delta^+(\{h\})} x_{ij} \le 1, \quad h \in N_S \quad (2)$$

$$\sum_{(i,j)\in\delta^-(\{e\})} x_{ij} = \sum_{k\in K}\sum_{(i,j)\in\delta^+(\{\lambda_N+k\})} x_{ij} \quad (3)$$

$$\sum_{(i,j)\in\delta^+(\{h\})} x_{ij} = \sum_{t\in T}\sum_{k\in K} y_{ht}^k, \quad h \in N_P \quad (4)$$

$$\sum_{(i,j)\in\delta^-(\{h\})} x_{ij} = \sum_{t\in T}\sum_{k\in K} y_{ht}^k, \quad h \in N_P \quad (5)$$

$$\sum_{k\in K}\sum_{t\in T} y_{ht}^k = 1, \quad h \in N_P \quad (6)$$

$$y_{\lambda_N+k,1}^k = \sum_{(i,j)\in\delta^+(\{\lambda_N+k\})} x_{ij}, \quad k \in K \quad (7)$$

$$y_{e,\lambda_T}^k = \sum_{(i,j)\in\delta^+(\{\lambda_N+k\})} x_{ij}, \quad k \in K \quad (8)$$

$$\sum_{t\in T}\sum_{p\in P} d_p y_{pt}^k \le Q, \quad k \in K \quad (9)$$

$$\sum_{t\in T}\sum_{k\in K} a_t y_{ht}^k \le \sum_{(i,j)\in\delta^-(\{h\})} z_{ij} \le \sum_{t\in T}\sum_{k\in K} b_t y_{ht}^k, \quad h \in N_P \quad (10)$$

$$\sum_{(i,j)\in\delta^+(\{h\})} z_{ij} - \sum_{(i,j)\in\delta^-(\{h\})} z_{ij} = \sum_{(i,j)\in\delta^+(\{h\})} (s + t_{ij}) x_{ij}, \quad h \in N_P \quad (11)$$

$$z_{ij} = t_{ij} x_{ij}, \quad h \in N_S, (i,j) \in \delta^+(\{h\}) \quad (12)$$

$$(t_{\lambda_N+1,i} + t_{ij} + s)x_{ij} \le z_{ij} \le (t_{max} - t_{je} - s)x_{ij}, \quad (i,j) \in A \mid i,j \notin N_S \cup \{e\} \quad (13)$$

$$\sum_{t\in T} y_{jt}^k \ge \sum_{t\in T} y_{it}^k + x_{ij} - 1, \quad (i,j) \in A, k \in K \quad (14)$$

$$x_{ij} \in \{0,1\}, \quad (i,j) \in A \quad (15)$$

$$y_{it}^k \in \{0,1\}, \quad i \in V, t \in T, k \in K \quad (16)$$

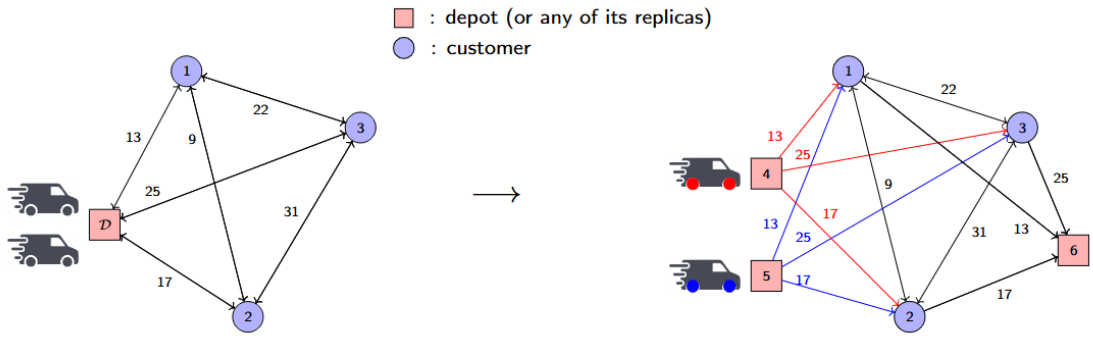


Fig. 1: Example of creation of our extended graph (on the right) starting from a standard network (on the left) involving a single depot, three customers, and two vehicles. In the extended graph, some arcs can be traversed only by a certain vehicle (red and blue arcs), while some others are in common (black arcs). Each vehicle starts from its own depot node, but they both arrive at the same final depot node e.

$$z_{ij} \geq 0, \quad (i,j) \in A. \tag{17}$$

The objective function (1) minimizes the total cost for the company, which is composed of the overall traveling time of the vehicles (first term) and the overall penalty cost paid for visiting customers in timeslots where there is a non-null probability of missing them (second term). Note that, for the sake of simplicity, this cost function is expressed in terms of time units as the scheduling variables and, therefore, also the penalties are measured in time.

Constraints (2) and (3) impose that all vehicle tours must originate from their specified start nodes and end at the common depot, e. The requirement for each customer in the set N_P to be visited exactly once is established through the flow conservation constraints (4) and (5). Constraints (6) guarantee that each customer visit is assigned to a single vehicle $k \in K$ and occurs within a single timeslot $t \in T$. If a vehicle $k \in K$ is used, constraints (7) and (8) ensure its corresponding start (i.e., $y^k_{\lambda_N+k,1} = 1$) and end (i.e., $y^k_{e,\lambda_T} = 1$) depot timeslots are activated. Furthermore, vehicle capacity limits are handled by constraints (9), which prevent the total demand of customers on any given route from exceeding the vehicle capacity, Q. The temporal feasibility of the routes is governed by the set of constraints from (10) to (13). Specifically, (10) enforces the arrival time within the time window $[a_t, b_t]$ for each service performed within timeslot $t \in T$, while (11) defines the progression of time between consecutive visits to customers i and j, accounting for service and travel times. The initial arrival time for each route's first customer is set by constraints (12). Constraints (13) introduce instead lower and upper bounds on the start of the service in a specific node j. This time has to be greater than the time required to reach the previous node from the depot, plus the time needed to reach j from the previous node, plus the service time in the previous node. The start of the service in node j also needs to make a return to the depot feasible, i.e., it needs to be lower than the maximum working time minus the time to serve the node minus the time to return to the depot. Together, these temporal constraints also serve the crucial function of eliminating subtours. Finally, the link between the routing variables (x) and the vehicle assignment variables (y) is established by constraints (14), to ensure that any two consecutively visited nodes are served by the same vehicle. These last constraints are needed since our formulation determines vehicle assignments without needing an explicit vehicle index on the x variables, a feature enabled by the unique starting depot for each vehicle. The domains of decision variables are defined in constraints (15) through (17).

The above formulation is similar to that presented in [4] and adopts a perspective that differs from that previously available in the literature [10][14][20][24]. In particular, the presented extended graph allowed us to obtain a more compact and more efficient MILP model through the introduction of:

- two-indexed binary variables for arc selection, which embed vehicle-specific route information without requiring a third index, similar to the method in [16];
- continuous scheduling variables to enforce time-based subtour elimination, drawing inspiration from modeling techniques for ordering problems [17].

The model efficiency obtained allows us to adopt state-of-the-art MILP solvers to find optimal solutions for real-world instance sizes.

III. An Improved Decisional Setting Including CAPs Flexibility

As stated in the introduction, the above formulation assumes that customer availability profiles are fixed input data and cannot be modified. However, in practice, companies often have the opportunity to foster some specific customer behavior of interest by offering some form of incentives (e.g., a discount on delivery fees). Assuming that the company has a budget B to support such modifications and a communication channel to reach the customers (e.g., email), it becomes possible to evaluate when and to what extent it is convenient for the company to implement such an influence on the customers' behavior.

To this aim, let us introduce an additional continuous decision variable γ_{pt} representing the increment (induced by the company) with respect to ρ_{pt} of the presence probability of customer $p \in P$ during timeslot $t \in T$[1]. Then, it is possible to create a new optimization model by modifying the one already presented in Section II-B, namely model (1)–(17). More precisely, the objective function (1) is substituted with the following one

$$\min \; f_{TC} + f_{HRC} - f_{HRI} \tag{18}$$

where

$$f_{TC} = \sum_{(i,j)\in\delta^-(\{e\})} z_{ij} \tag{19}$$

$$f_{HRC} = \sum_{p\in P}\sum_{t\in T}\sum_{k\in K} \alpha_p \; (1-\rho_{pt}) \; y^k_{p,t} \tag{20}$$

$$f_{HRI} = \sum_{p\in P}\sum_{t\in T} \alpha_p \gamma_{pt} \tag{21}$$

and the following inequalities

$$0 \le \gamma_{pt} \le (1-\rho_{pt}) \sum_{k\in K} y^k_{pt}, \quad p \in P, t \in T \tag{22}$$

$$\sum_{p\in P}\sum_{t\in T} \alpha_p \gamma_{pt} \le B \tag{23}$$

are added to the system of constraints (2)–(17).

The new objective function (18), in which the overall traveling cost is named f_{TC} and the overall hit-rate cost based on the pure CAPs initially given by the customers is named f_{HRC}, now incorporates a third term representing the saving on the hit-rate cost when considering the CAPs modified by the company investment, which is named f_{HRI}. In constraints (22), the new decision variables are allowed to take non-negative values up to the residual presence probability of a customer in a certain timeslot. However, since it makes sense to pay for improving presence probabilities only for timeslots included in the visiting schedule, constraints (22) set to 0 the corresponding variables γ when it is not the case. Finally, constraint (23) limits the company's investment in the modification of the CAPs to a certain budget B.

With respect to the model just stated, it is important to clarify two points:

- despite the definition of the decision variables γ, in reality, the company cannot directly adjust the customers' probability profiles. The model instead quantifies the amount of additional presence probability the company is aiming for and returns the proportional compensations to be offered to the customers for being at home during the selected timeslot. While the customers still have the freedom to choose what to do, the offered compensation is meant to convince them to stay at home during the desired timeslot.
- the third term of objective function (18) represents a saving only for the problem as it is modeled, not on the actual overall cost for the company, which is of course made up by the total objective function value plus the amount invested to influence the customers, i.e., the same amount of the saving. Hence, the company wants to evaluate the possibility of investing more than the traveling and expected hit-rate costs to better balance them.

From now on, we will refer to the MILP model (18) subject to constraints (2)–(17) and (22)–(23) as the *Attended Home Delivery Problem with CAPs Flexibility* (AHDP-CF).

IV. Experimental setting

This section describes the experimental setup adopted to evaluate how the budget constraint affects the optimization problem considered. Section IV-A outlines the generation of benchmark instances, while Section IV-B and Section IV-C describe the analytical methodology developed and the corresponding implementation details, respectively.

A. Instances generation

To obtain statistically meaningful results, our experiments consider a very large set of benchmark instances, in which many of their characteristics are randomly varied. More precisely, we generated 500 instances involving 10 customers (i.e., $|P| = 10$) and 3 vehicles (i.e., $\lambda_K = 3$). Customers are randomly distributed within a 20000×20000 m^2 area. Euclidean distances are computed between each node[2], assuming a constant speed of 11.11 m/s (i.e., 40 km/h). They are then truncated to integer values. Customer demands d_p range from 5 to 30 units, with a fixed service time of 7 minutes. Vehicle capacity is set to $Q = 3 \cdot \frac{\sum_{p\in N_P} d_p}{\lambda_K}$ to align total fleet capacity with aggregate demand. Each vehicle's work shift is limited to $t_{max} = 14400$ seconds (i.e., 4 hours), while the time horizon is divided into 8 equal timeslots (i.e., $\lambda_T = 8$). This time horizon has been chosen to mimic the typical half-day shift of many couriers; however, the model can tackle wider time horizons and denser timeslot division of the shift. Each customer is assigned to one of seven CAPs, each representing different customer presence behaviors during the day. Six profiles are adapted from [10] and eventually discretized, namely, *V-shape*, *A-shape*, *M-shape*, *W-shape*, *Linear-Increase*, and *Linear-Decrease*, whereas the last profile is randomly generated. A detailed visualization of the six non-random profiles is reported in Figure 2. Clearly, each profile is intended to mimic a peculiar behavior of a customer during the time horizon. Failure costs α_p are drawn from a Normal distribution with mean equal to the service time s and standard deviation equal to $s/2$, truncated in $[2, 15]$ minutes.

The generated instances are publicly available at https://or-dii.unibs.it/instances/AHDPCF.zip.

[1]Note that this variable represents an increment and not a free variation with respect to ρ_{pt} since it never makes sense for the company to pay for diminishing the presence probability of a customer during a timeslot.

[2]While real distances between customers seldom correspond to Euclidean ones, the absence of a real street map and the randomness used to generate the customer locations justify the adoption of this simple metric.

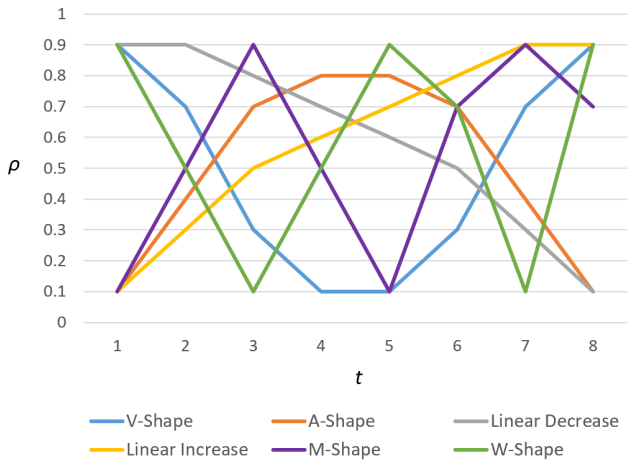


Fig. 2: A plot of the six non-random discrete Customer Availability Profiles adopted, across 8 timeslots.

B. Analytical methodology

The methodology adopted for our quantitative analysis is mainly based on the possibility to solve the MILP model AHDP-CF proposed in Section III and a two-step lexicographic approach that interprets the proposed multi-attribute optimization problem as an actual multi-objective one. The analysis includes a systematic change of the available budget B. To this aim, let B_{max} represent the maximum sensible budget that the company can consider and let β be a real value in $[0, 1]$. Then, let us define $B = \beta B_{max}$ as the available budget. In our analysis, β is used as a sensitivity parameter to evaluate the behavior of the optimal solution for the company against different budgets. The value of B_{max}, instead, which is instance-dependent, is calculated during the optimization process described below.

More precisely, for each of the generated instances, the following optimization steps are implemented:

- **Step 1:** A lexicographic optimization is performed on AHDP-CF where, first, the priority is given to the minimization of the overall objective function (18) and, then, the sole minimization of the hit-rate investment is pursued. This second optimization step is needed since there could be multiple optimal solutions, all with the same objective function value but different proportions of the three objective function terms f_{TC}, f_{HRC}, and f_{HRI}. Among these solutions, the one with the lowest possible value of f_{HRI} is of our interest. More in detail:
  - **Step 1.a**: The MILP model AHDP-CF is solved to optimality without considering the budget constraint (23). This way, the model returns a solution in which the presence probabilities of interest (i.e., those corresponding to the timeslots visited) are increased at least up to the most convenient value with respect to the traveling time minimization. Let us call $f^{Step1.a}$ the value of the optimal solution just obtained.
  - **Step 1.b**: The MILP model AHDP-CF is solved to optimality again without considering the budget constraint (23) but with the new objective function

$$\min \ f_{HRI} \tag{24}$$

and with the following additional constraint

$$f_{TC} + f_{HRC} - f_{HRI} \leq f^{Step1.a}. \tag{25}$$

Constraint (25) ensures that the value of the primary objective function obtained in the previous step is not deteriorated. Let us call $f_{HRI}^{Step1.b}$ the value of f_{HRI} calculated in the optimal solution.

- **Step 2**: The value of B_{max} is set to $f_{HRI}^{Step1.b}$, since it does not make sense to consider larger budgets for the current instance.
- **Step 3**: The parameter β is set to a value in $[0, 1]$ and the actual budget B is calculated.
- **Step 4**: The MILP model AHDP-CF is solved to optimality, now including also the budget constraint (23). Let us call sol_β the optimal solution obtained. This solution is the one analyzed in the later discussion of the results.

C. Implementation and resolution details

Experiments were performed on a machine equipped with an *Intel Xeon Gold* 6140x (18 physical cores, 36 logical threads) CPU, 64GB RAM, and running a *Windows 11* 64-bit operating system. The solution framework, including the lexicographic procedure and MILP models, was implemented in Java 16. Gurobi v12.0.0 was adopted for solving the MILPs; the default parameter setting is used, and no time limit is imposed on the resolution, thus ensuring termination only upon optimality.

V. Results and Discussion

This section presents and discusses the results of the computational experiments conducted to evaluate the impact of investing a certain budget for improving the expected hit-rate (representing the quality of the delivery service) on our AHDP-CF. Our analysis focuses on the trade-offs between the investment in higher customer presence probabilities and the resulting operational costs and solution structure. To provide clear managerial insights, we always compare the solutions sol_β obtained by using an increasing budget (from $\beta = 0.05$ up to $\beta = 1$, with a 0.05 increment) against solution sol_0, which represents the baseline scenario where no budget is available (i.e., when $\beta = 0$).

Our analysis is designed to provide quantitative insights into how different budget levels β reshape the optimal solution. In each part of the discussion, results are reported as sequences of boxplots, one for each β value that has been tested, thus aggregating the results obtained over all the 500 instances generated. In a boxplot, the cross denotes the average value, while the thick line indicates the median. The upper part of the colored box corresponds to the 75th percentile, while the lower one is associated with the 25th percentile. Moreover, the whiskers span from the minimum to the maximum value,

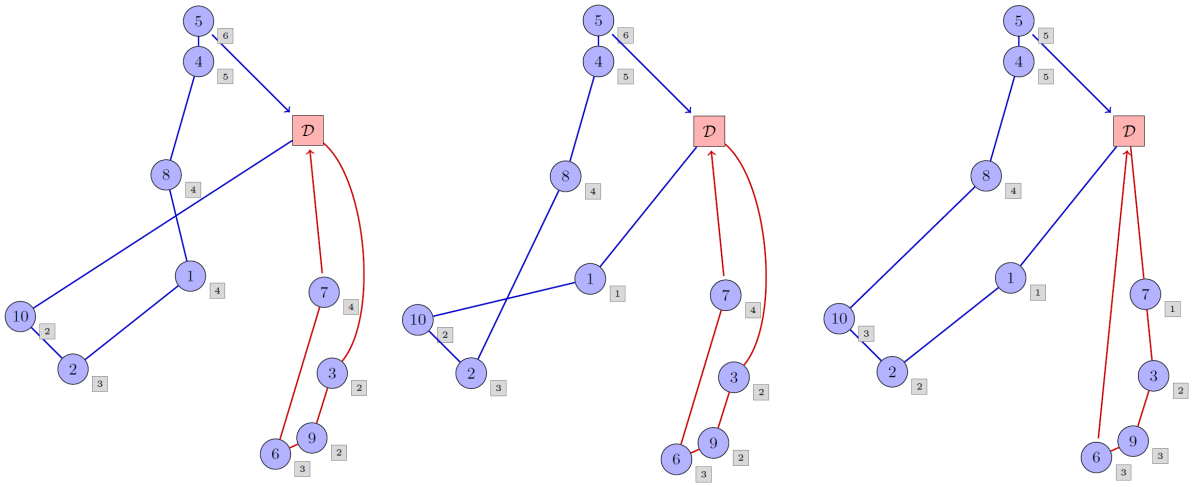


Fig. 3: Solution examples for different β values: $\beta = 0$ (left), $\beta = 0.50$ (center), $\beta = 1.00$, (right). The number in a gray square near each node indicates the timeslot selected for the corresponding visit.

excluding the outliers (which are reported as simple dots). Finally, we want to assess the impact of the available budget from several perspectives, each of which is related to a different Key Performance Indicator (KPI) that can be calculated on the solutions obtained by implementing the analytical method proposed in Section IV-B. In this discussion, we decided to focus on the following KPIs of interest:

- CSS (change in start times): represents, averaged over all the customers, the absolute value of the difference in seconds between the service start times scheduled in sol_β and those scheduled in sol_0.
- PCT (percentage of changed timeslots): represents the percentage of customers for which the timeslot scheduled for the delivery in sol_β is different from that scheduled in sol_0.
- PAC (percentage of affected customers): represents the percentage of customers in sol_β for which the company has invested at least something greater than 0 (clearly, in sol_0 where the budget is null, the baseline comparison value is always 0).
- PCTC (percentage change in traveling cost): represents the difference between the traveling cost of sol_β and that of sol_0, as a percentage of the traveling cost of sol_0.
- PCHC (percentage change in hit-rate cost): represents the difference between the hit-rate cost f_{HRC} of sol_β and that of sol_0, as a percentage of the available budget B.

To illustrate how investment drives structural changes in a single instance, Figure 3 compares the routes obtained with three different β values, namely, 0.0, 0.50, and 1.00. Although node-to-vehicle assignments remain fixed, increasing β reorders visits. The solution with $\beta = 1.00$ corresponds to the solution that could be obtained by solving a standard VRP with travel time minimization. The other solutions show more complicated routes, since the cost of visiting customers in less favorable timeslots cannot be fully compensated by the budget.

Figure 4 illustrates the distribution of the change in the service start times (CSS) for customers across various budget levels. A noticeable trend emerges, characterized by three

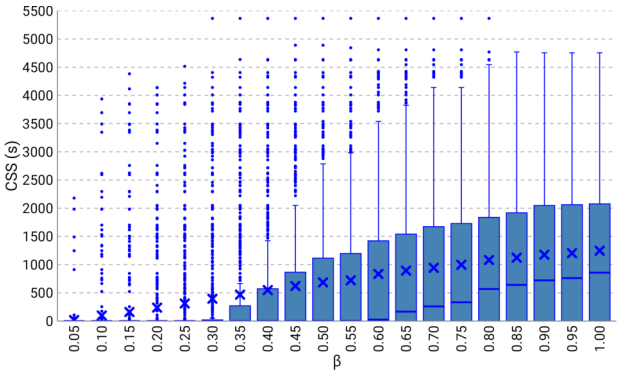


Fig. 4: Change of the start of the service (CSS), averaged over all customers.

distinct phases. Initially, for low budget levels ($\beta \leq 0.30$), CSSs are negligible, as the median, 25th, and 75th percentiles are all concentrated close to zero. As the budget increases into the intermediate range of $0.35 \leq \beta \leq 0.60$, the median CSS remains at zero, but the distribution skews significantly upward. This indicates that, while the start time for more than half of the customers is unaffected, a growing portion of them experiences a substantial shift, as seen in the 75th percentile progressively increasing to a value of nearly 1500 seconds. For the highest budget levels ($\beta \geq 0.65$), a different operational behavior emerges where the entire distribution shifts upwards. In particular, the median CSS itself starts to increase, reaching nearly 900 seconds for $\beta = 1$, signifying that the investment now alters the start times for the majority of the customers. This tripartite behavior is a pattern that can be noticed in most analyzed KPIs, indicating a fundamental shift in the solution's

characteristics based on the amount of budget invested. In contrast to such a behavior of the median, the average CSS has an almost linear increase from 0 to around 1250 seconds.

Figure 5 presents the distribution for the percentage of customers (PCT) whose selected timeslot changes with respect to the zero-budget baseline. The analysis reveals the same clear

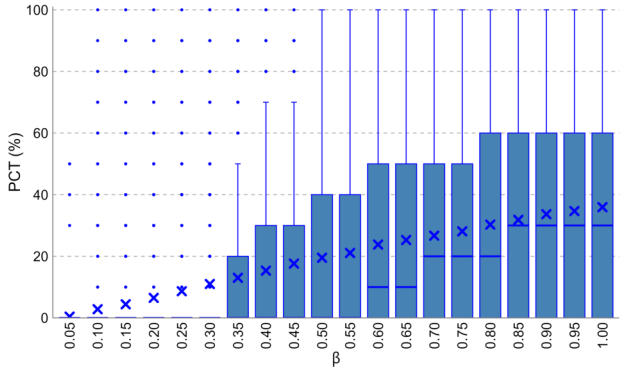


Fig. 5: Percentage of customers for which the timeslot selected by the company changes (PCT).

threshold-based behavior as in the previous KPI. The average PCT still increases linearly across β values, going from 0 to about 35%. The median's behavior, however, is more distinctly segmented. In the first segment ($\beta \leq 0.30$), the median PCT is null, thus indicating that most customers are unaffected. A sharp transition then occurs at $\beta = 0.35$ where the 75th percentile rises while the median stays at 0. For $\beta \geq 0.60$, the median increases in three steps and stabilizes at 30% for $\beta \geq 0.85$.

To understand the mechanism driving these outcomes, Figure 6 illustrates the investment strategy identified in the optimal solutions, by showing the percentage of customers (PAC) for which the company invests in changing their behavior. In

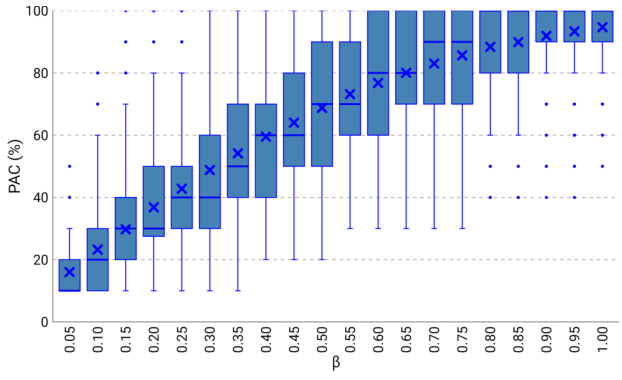


Fig. 6: Percentage of customers on which the company invests its budget (PAC).

contrast to the complex, threshold-based behavior seen in the previous KPIs, the application of the budget is remarkably direct and nearly linear. As β increases, the percentage of affected customers also rises monotonically and with relatively low variability. The median and average PAC values remain closely aligned throughout the entire range, starting near 15% and climbing steadily until they approach 100% for higher investment levels. A notable insight emerges when comparing this strategy to its impact on rescheduling (Figure 5), i.e., the percentage of customers receiving investment is considerably higher than the percentage whose timeslot is changed. For instance, if we consider median values at $\beta = 0.70$, the budget is invested in a 90% of the customers, while the percentage of customers with an altered timeslot is just 20%. This implies that, for a significant portion of customers (around 70%), the budget is invested not to change their assigned timeslot, but to increase the likelihood of delivery success for the scheduled timeslot in sol_0.

Figure 7 evaluates the impact of the budget on the routing decisions by showing the percentage change in traveling costs (PCTC). Note that, differently from the previous ones, this KPI

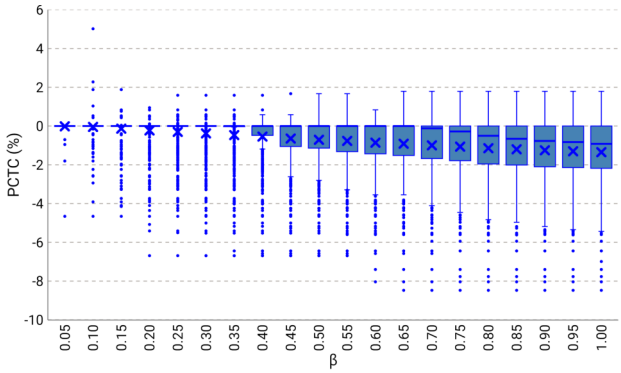


Fig. 7: Percentage change in traveling costs (PCTC).

may result in both negative and positive values. When they are negative, the f_{TC} derived from the solution including some investment is lower than the no-budget solution. This is the more frequent behavior. The resulting trend reveals a complex behavior that strongly mirrors (even if horizontally specular) the tripartite behavior of the first two KPIs analyzed rather than the linear investment strategy itself. Such a behavior appears, in this case, with thresholds at $\beta = 0.40$ and $\beta = 0.70$. The average PCTC decreases linearly from 0% to -1.4%, while the median PCTC value reaches a minimum of about -1%. Interestingly enough, there are some cases in which the budget investment increases the traveling cost, meaning that the model favors some decrease in hit-rate cost at the expense of a smaller increase in traveling cost. These situations correspond to noisy results, due to the possible existence of multiple optimal solutions for the problem. In fact, when a certain non-null budget is available, it is never optimal to invest in changing the solution structure without improving the traveling cost, since the hit-rate saving must always be compensated by the budget investment. This means, for every solution with an increase in traveling costs, there must be an equally optimal solution in which the traveling cost remains the same or improves compared to sol_0. Anyway, across all

the β values, these cases fall beyond the 75th percentile, and, therefore, they do not disrupt the validity of our analysis.

The final analysis, presented in Figure 8, examines the portion of the budget dedicated to compensating for changes in the objective function component f_{HRC}, as a percentage of the available budget to invest (PCHC). As before, this KPI may

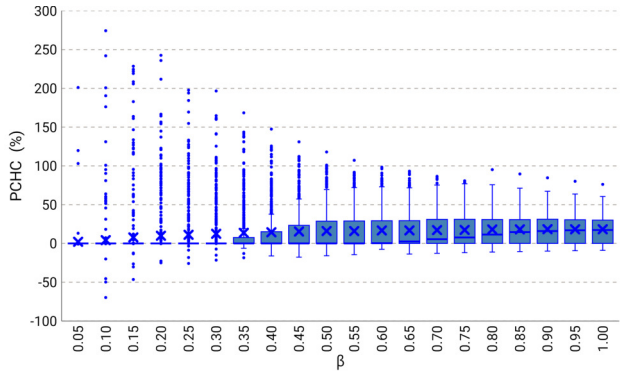


Fig. 8: Percentage change in hit-rate cost (PCHC).

result in both negative and positive values. Here, a positive value means that the value of f_{HRC} when a budget is available is greater than that obtained with zero budget. This is the more common situation. The trend follows a clear phased pattern, like in most previous KPIs. The median PCHC is initially zero, becomes positive and rises through the mid-range of β values, and finally stabilizes just under 20%. Notably, the high variability observed at low-to-moderate investment levels decreases significantly as the solution stabilizes. Interpreting these results provides a deeper insight into the solution's trade-offs. When the median PCHC value is close to zero, it suggests that the budget is primarily used to increase success probability within the timeslots scheduled for visits in sol_0. Instead, when the median PCHC stabilizes just under 20% for high β values, it indicates that roughly a fifth of the budget is now used to offset the corresponding increase in the hit-rate cost. This increase reveals a strategic selection of different timeslots for the visits that allows the traveling cost reduction seen previously for the same β values. The PCHC distribution also reveals a wide range of outcomes. Extreme cases show that the PCHC can reach almost 275%, where the hit-rate cost increase is 2.75 times the available budget. Conversely, very few times, negative values may also occur. These situations reveal the presence of multiple optimal solutions, as seen for the previous KPI. Again, the frequency of these cases does not affect the insights derived.

VI. Conclusions

In this paper, we studied the Attended Home Delivery problem with Customer Availability Profiles and introduced a novel and managerially relevant dimension, namely, the ability for a company to invest a budget to actively influence customer presence probabilities during certain timeslots of interest. We proposed the corresponding new optimization setting together with its MILP formulation, called the Attended Home Delivery problem with CAPs flexibility (AHDP-CF). Through a structured analytical methodology based on lexicographic optimization, we conducted an extensive computational study to quantify the impact of the possible investment. Our analysis revealed that the relationship between the investment in service quality and the resulting operational outcomes is characterized by two distinct thresholds that define three different ranges in which the invested budget has different impacts. We showed that small, incremental investments (within the first range) yield negligible changes to the solution, whereas substantial budgets (within the third range) enable a fundamental restructuring of the delivery routes and schedules. The primary managerial insight derived is the possible quantification of a strategic trade-off. Many times, it is optimal for the company to leverage the budget to accept higher penalties on some customer visits in exchange for significant savings in overall traveling costs. This work, therefore, provides a clear and quantitative understanding of the *price of customer presence* in AHD logistics. The framework and insights presented can serve as a valuable decision-support tool for logistics companies, allowing them to move beyond treating customer availabilities as fixed data and instead see them as a strategic lever that can be priced and optimized.

Future research could extend this work in several directions. Developing scalable heuristic or metaheuristic algorithms would allow the AHDP-CF to be applied to larger, real-world instances. Moreover, exploring dynamic settings where investment decisions can be adjusted during the operational day presents another promising avenue for future investigation. Finally, one may also want to incorporate into the problem the *willingness to accept* of a customer against the company's offers by adopting more sophisticated models specifically tailored for individual choices.

Acknowledgment

The authors gratefully acknowledge the financial support of "ULTRAOPTYMAL - Urban Logistics and sustainable TRAnsportation: OPtimization under uncertainTY and MAchine Learning", a PRIN2020 project funded by the Italian University and Research Ministry (grant number 20207C8T9M, website: https://ultraoptymal.unibg.it).

References

[1] N. Agatz, A. Campbell, M. Fleischmann, and M. Savelsbergh. Time slot management in attended home delivery. *Transportation Science*, 45(3):435–449, aug 2011. doi: 10.2307/23018537.

[2] M. Bierlaire, E. Fadda, L. F. Tiotsop, and D. Manerba. A chance-constraint approach for optimizing social engagement-based services. page 301–304, 2022. doi: 10.15439/2022F235.

[3] M. Bierlaire, E. Fadda, L. F. Tiotsop, and D. Manerba. Workforce allocation for social engagement services via stochastic optimization. *Studies in Computational Intelligence*, 1158 SCI:71–87, 2024. doi: 10.1007/978-3-031-57320-0_5.

[4] V. Bonomi, D. Manerba, R. Mansini, and R. Zanotti. Optimizing attended home delivery: Multiple recovery options and customer availability profiles to face synchronization failures. *International Journal of Production Economics*, 279, 2025. doi: 10.1016/j.ijpe.2024.109463.

[5] V. Bonomi, R. Mansini, and R. Zanotti. Last mile delivery with parcel lockers: evaluating the environmental impact of eco-conscious consumer behavior. *IFAC-PapersOnLine*, 55(5):72–77, 2022. doi: 10.1016/j.ifacol.2022.07.642.

[6] V. Bonomi, R. Mansini, and R. Zanotti. Mediating governance goals with patients and nurses satisfaction: a multi-actor multi-objective problem including fairness. *International Journal of Production Research*, 62(10):3753–3780, 2023. doi: 10.1080/00207543.2023.2248522.

[7] N. Boysen, S. Schwerdfeger, and F. Weidinger. Scheduling last-mile deliveries with truck-based autonomous robots. *European Journal of Operational Research*, 271(3):1085–1099, 2018. doi: 10.1016/j.ejor.2018.05.058.

[8] A. Cassia, O. Jabali, F. Malucelli, and M. Pascoal. The electric vehicle shortest path problem with time windows and prize collection. pages 313–322, 2022. doi: 10.15439/2022F186.

[9] J.-F. Côté, R. Mansini, and A. Raffaele. Multi-period time window assignment for attended home delivery. *European Journal of Operational Research*, 316(1):295–309, 2024. doi: 10.1016/j.ejor.2024.01.021.

[10] A. M. Florio, D. Feillet, and R. F. Hartl. The delivery problem: Optimizing hit rates in e-commerce deliveries. *Transportation Research Part B: Methodological*, 117:455–472, nov 2018. doi: 10.1016/j.trb.2018.09.011.

[11] Forrester Research, Inc. Forrester Data: Online Retail Forecast, 2017 to 2022 (Western Europe), 2016.

[12] M. Gajda, O. Gallay, R. Mansini, and F. Ranza. Optimizing last-mile delivery through crowdshipping on public transportation networks. *Transportation Research Part C: Emerging Technologies*, 179:105250, 2025. doi: 10.1016/j.trc.2025.105250.

[13] V. Garg, S. Niranjan, V. Prybutok, T. Pohlen, and D. Gligor. Drones in last-mile delivery: A systematic review on efficiency, accessibility, and sustainability. *Transportation Research Part D: Transport and Environment*, 123:103831, 2023. doi: 10.1016/j.trd.2023.103831.

[14] A. Gobbi, D. Manerba, R. Mansini, and R. Zanotti. A kernel search for a patient satisfaction-oriented nurse routing problem with time-windows. volume 52, page 1669 – 1674, 2019. doi: 10.1016/j.ifacol.2019.11.440.

[15] A. Lorenz, N. Madeja, and C. Leyh. A framework for assessing the sustainability impact of intelligent transport systems in the smart city context. pages 161–169, 2023. doi: 10.15439/2023F6002.

[16] Z. Luo, H. Qin, C. Che, and A. Lim. On service consistency in multi-period vehicle routing. *European Journal of Operational Research*, 243(3):731–744, 2015. doi: 10.1016/j.ejor.2014.12.019.

[17] F. Maffioli and A. Schiomachen. A mixed-integer model for solving ordering problems with side constraints. *Annals of Operations Research*, 69:277–297, 1997. doi: 10.1023/A:1018989130169.

[18] D. Manerba, R. Mansini, and R. Zanotti. Attended home delivery: reducing last-mile environmental impact by changing customer habits. *IFAC-PapersOnLine*, 51(5):55–60, 2018. doi: 10.1016/j.ifacol.2018.06.199.

[19] R. Mansini, F. Ranza, and R. Zanotti. Towards sustainable last-mile delivery: Introducing an off-peak urban policy to mitigate environmental and social impacts. *IFAC-PapersOnLine*, 58(2):186–191, 2024. doi: 10.1016/j.ifacol.2024.07.112.

[20] E. Marzal and L. Sebastia. Team orienteering problem with time windows and variable profit. page 347–350, 2022. doi: 10.15439/2022F158.

[21] Morgan Stanley. Here's why e-commerce growth can stay stronger for longer. https://www.morganstanley.com/ideas/global-ecommerce-growth-forecast-2022, 2022. Accessed: 2025-07-18.

[22] K. Mousavi, M. Bodur, and M. J. Roorda. Stochastic last-mile delivery with crowd-shipping and mobile depots. *Transportation Science*, 56(3):612–630, 2022. doi: 10.1287/trsc.2021.1088.

[23] R. Spliet and G. Desaulniers. The discrete time window assignment vehicle routing problem. *European Journal of Operational Research*, 244(2):379–391, jul 2015. doi: 10.1016/j.ejor.2015.01.020.

[24] S. Voigt, M. Frank, P. Fontaine, and H. Kuhn. The vehicle routing problem with availability profiles. *Transportation Science*, 57(2):531–551, mar 2023. doi: 10.1287/trsc.2022.1182.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 635–639
DOI: 10.15439/2025F5496
ISSN 2300-5963 ACSIS, Vol. 43

Adiabatic Quantum Computing for the Subset Sum Problem: Preliminary Studies

C.A.F. Bernardes,* P. Castellucci,* D.S. Gonçalves,* E.I. Duzzioni* A. Mucherino,†

†Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

ces.aug.fre@gmail.com, pedro.castellucci@ufsc.br,

douglas.goncalves@ufsc.br, eduardo.duzzioni@ufsc.br

†IRISA, Université de Rennes, Rennes, France

antonio.mucherino@irisa.fr

***Abstract*—The Subset Sum Problem (SSP) is one of those combinatorial problems that are very easy to understand (take a bunch of integer numbers and verify whether there exists a subset of these numbers which sums up to a given target integer), but it can be very difficult to solve. The SSP is actually an NP-complete problem, but it is "weakly" NP-hard, implying that there are instances of SSP that can be solved in polynomial time. For this particular problem, the instance hardness can be measured by evaluating the so-called "density" index, which basically compares the number of involved integer numbers to the number of bits we need for their binary representation. In our preliminary study on the use of adiabatic quantum computing for the SSP, we investigate the actual feasibility in solving hard instances of the problem. In fact, hard SSP instances are those requiring a large number of bits for the representation of the integers, while the analog nature of the quantum computer does not allow us to ensure highly accurate integer representations. Some preliminary computational experiments performed on D-Wave quantum annealer are presented and compared to standard solvers for classical computers.**

I. Introduction

GIVEN a set $\mathcal{A}$ of n positive integer numbers and a positive target t, the Subset Sum Problem (SSP) asks whether there is a subset $\mathcal{S}$ of $\mathcal{A}$ such that the elements of $\mathcal{S}$ sum up exactly to t. The original set $\mathcal{A}$ may be a *multiset*, but in this work we assume, for simplicity, that it contains no repeated elements. In a typical problem variant, the SSP asks to enumerate *all* possible solutions (i.e. all subsets) of a given SSP instance. But again, for simplicity, we make the assumption in this work that only one solution is required.

The SSP is one of the famous Karp's NP-complete problems [12]. It is considered to be *weakly* NP-hard, because there exist large classes of SSP instances for which algorithms admitting a pseudo-polynomial complexity have been proposed [6]. In order to find out the actual hardness of a given SSP instance, the so-called *density* index can be employed:

$$\rho := \frac{n}{\ell(n)},$$

where

$$\ell(n) = \log_2 \max_{a \in \mathcal{A}} a.$$

SSP instances having a density index ρ equal to 1 are the hardest to solve, and indeed, to date, no pseudo-polynomial algorithm for these instances have been found. In these instances, the number of bits necessary to represent the integers in $\mathcal{A}$ equals the number of integers in $\mathcal{A}$. In this work, we will focus our attention on these density-1 instances.

There is a large scientific literature on the SSP. A large review of the existing literature is out of the scope of this short paper, so that we will limit ourselves to cite only a few recent works aiming at improving the performance of SSP algorithms for hard "density-1" instances. The algorithm that has been considered to be the *fastest* for these instances for a very long time is one proposed by Horowitz and Sahni in 1974 [11], which has, however, an exponential theoretical complexity. In more recent times, the works presented in [3] and [19], for example, showed that there exist algorithmic solutions which can exhibit better performance. It is important to remark that the research is progressing in parallel in both areas of deterministic algorithms (such as in [3]) and heuristics (such as in [19]). The most recent effort in the former category of algorithms was presented in the pre-print [18]. To date, however, algorithms having polynomial complexity have not been found yet for solving density-1 instances (a discovery that could imply that **P**=**NP**, which is unlikely).

Many researchers are exploring nowadays alternative approaches for tackling the SSP. Some of us are for example currently working on a novel approach in optical computing [10], where the use of the spatial properties of light allow us to speedup the solution of hard SSP instances (very recently, moreover, the same optical circuit has been adapted to perform single-key cryptography in [7]). Even if these optical circuits are not "quantum devices", they exploit physical phenomena normally used in quantum computing, such as interference.

Some recent theoretical studies aiming at exploiting the potential "quantum advantage" of modern quantum technologies for solving the SSP can be found for example in [2], [5], [9].

In this work, we study the feasibility of solving hard density-1 instances of the SSP by adiabatic quantum computing. To the best of our knowledge, this alternative approach to the SSP was only a little studied in recent times. One example of such studies can be found in [20], where a mathematical model similar to ours is introduced, whose solution is however attempted in a different manner. Moreover, in our study, we will focus our attention on the analog nature of the considered quantum computer, and on the consequent representation issues for the elements of $\mathcal{A}$ (as well as for the target t). We also point out that several of the publications cited above are mostly theoretical and provide no computational experiments.

The rest of the paper is organized as follows. Two general methods for the SSP are briefly presented in Section II, because they will be involved in the performance comparisons presented with our computational experiments. Section III will briefly describe the physics behind the quantum annealer implemented by the D-Wave quantum processing units, and propose a reformulation of the SSP as a Quadratic Unconstrained Binary Optimization (QUBO), a problem format that can handled by D-Wave quantum computers. Finally, our experiments will be reported in Section IV, and Section V will conclude the paper.

II. Methods for the SSP

A. The BP algorithm

We begin this section by presenting a generic algorithm whose worst-case complexity is exponential, but that is suitable for all kinds of SSP instances, including the hard ones. This algorithm is presented in different guises in textbooks about NP-complete problems, and under different names. We make the choice to refer to it as the Branch-and-Prune (BP) algorithm [15]: this is the name of an algorithm for a more general problem, the Discretizable Distance Geometry Problem (DDGP) [17], which can be reduced to the SSP under special conditions. The DDGP has several important applications in various research fields [16].

The BP algorithm is based on the idea to construct a binary tree, where every path from the root to one of the leaf nodes represents a possible subset $\mathcal{S}$ of the original set $\mathcal{A}$. Every layer of the tree is related to one of the integers in $\mathcal{A}$: if the subsets "opting" for the left-handed branch at a given level do not contain the corresponding integer, then this same integer would instead be contained in all other subsets where the right-handed branch is selected. The tree grows regularly by doubling its nodes as we step from one layer to the next. The full tree contains therefore 2^n nodes, but they do not need to be all explored. Some tree branches can in fact be *pruned* by verifying some feasibility tests. If the integers that are selected in the current branch sum up to a value that is larger than the target t, for example, then we can prune the current branch. Also, if the sum of all remaining (i.e. not yet considered) integers at the current tree layer does not allow the current partial sum to reach the desired target t, then the branch can also be pruned.

An implementation in C programming language of the BP algorithm for the SSP is publicly available on the GitHub[1].

B. MILP formulation

Given the set $\mathcal{A} = \{a_1, \ldots, a_n\}$ and the target t, the SSP can be cast as the following boolean linear equation:

$$\sum_{i=1}^{n} a_i x_i = t, \tag{1}$$

where $x \in \{0,1\}^n$. Since, in principle, the SSP instance at hand may be infeasible, we need to take into consideration the case where the equation above may admit no solutions. For this reason, we introduce a slack variable $z \in \mathbb{R}$ and formulate the following optimization problem:

$$\begin{aligned} \min_{x,z} \quad & |z| \\ \text{s.t} \quad & \sum_{i=1}^{n} a_i x_i + z = t, \\ & x \in \{0,1\}^n, z \in \mathbb{R}. \end{aligned}$$

If the optimal value for the problem above is zero, then the SSP has a solution given by the binary decision variables x_i, where $x_i = 1$ indicates that a_i is a member of the solution subset $\mathcal{S}$.

In order to make the objective function linear, we can represent $z = \Delta^+ - \Delta^-$, where Δ^+, Δ^- are non-negative real variables. Then, we have the following Mixed Integer Linear Program (MILP):

$$\begin{aligned} \min_{x,\Delta^+,\Delta^-} \quad & \Delta^+ + \Delta^- \\ \text{s.t} \quad & a^T x + \Delta^+ - \Delta^- = t, \\ & \Delta^+ \geq 0, \quad \Delta^- \geq 0, \\ & x_i \in \{0,1\}, \quad i = 1, \ldots, n, \end{aligned} \tag{2}$$

where

$$a^T x = \sum_{i=1}^{n} a_i x_i.$$

Such a formulation is ideal for MILP solvers (as for example Gurobi [22], which we will use in our computational experiments) that implement branch-and-bound techniques for MILPs.

III. Quantum Annealing

Quantum annealing is an optimization technique used to find the ground state of a given Hamiltonian through the adiabatic evolution of the quantum system [1]. The problem to be solved can be encoded in the time-dependent Hamiltonian:

$$H(\tau) = A(\tau)H_I + B(\tau)H_F,$$

where the time-dependent coefficients $A(\tau)$ and $B(\tau)$ satisfy the condition $A(0) \neq 0$, $B(0) = 0$ and $B(T) \neq 0$, $A(T) =$

[1] https://github.com/mucherino/DistanceGeometry

0, with T being the total evolution time, which corresponds to the computation time. Following the proposal of quantum annealing in the transverse field Ising model [13], the initial Hamiltonian for n qubits is described by

$$H_I = -\sum_i^n \sigma_x^i$$

and the final one by

$$H_F = \sum_i^n h_i \sigma_z^i + \sum_{i,j}^n J_{i,j} \sigma_z^i \sigma_z^j.$$

In these equations, h_i denotes the bias field acting on the i-th qubit and $J_{i,j}$ is the coupling constant between the qubits i and j. Besides of being an easy-to-prepare Hamiltonian, whose ground state is

$$|\psi(0)\rangle = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes n},$$

H_I acts as a driver Hamiltonian, which is responsible for mixing the states in the computational basis. Therefore, if initially the physical system starts in the ground state of the Hamiltonian H_I and is adiabatically driven until the final configuration determined by H_F, then, by performing measurements in the computational basis, we can find the solution to our problem as a string of bits.

The strategy described above is the one implemented by the quantum computers produced by the D-Wave Systems Inc. company. In practice, the quantum system is working on 20 mK [21], so there is a no null probability of finding the system outside the ground state, indicating the probabilistic character of the solution of the computation. This implies that, in order to solve a given problem using the D-Wave's quantum processing units, it is necessary to run the calculations several times.

Instead of formulating the problem in terms of the Ising Hamiltonian, it is possible to write it in the form of a Quadratic Unconstrained Binary Optimization (QUBO) problem [8]. The mapping of Pauli operators (σ^i) to QUBO operators (q^i) is obtained by $\sigma^i = 2q^i - I$, where I is the identity matrix. Therefore, the QUBO Hamiltonian becomes:

$$H_F = \sum_i^n Q_i q^i + \sum_{i,j}^n Q_{i,j} q^i q^j,$$

where the operators q^i have eigenvalues 0 or 1, provided that the eigenvalues of the Pauli matrices are -1 and 1. From a practical point of view, the problem is in general written in the QUBO formulation and then translated to the Ising model, which is the natural physical formulation.

In order to obtain a QUBO formulation for the SSP, we remark that $x \in \{0,1\}^n$ solves equ. (1) if it is a zero for the quadratic function:

$$q(x) = (a^T x - t)^2, \tag{3}$$

where a is a column vector with elements a_i, and the same goes for $x \in \{0,1\}^n$. We can therefore re-write:

$$q(x) = x^T a a^T x - (2ta)^T x + t^2 = x^T Q x + b^T x + c,$$

with $Q = aa^T$, $b = -2ta$ and $c = t^2$. By ignoring the constant term c, we have a correspondence with the QUBO Hamiltonian where

$$Q_i = -2ta_i, \quad Q_{i,j} = a_i a_j,$$

are its coefficients.

IV. Computational experiments

A. Generating hard SSP instances

In order to generate hard SSP instances, we make sure that its density index is as close as possible to 1. To this end, we define the very first integer to be included in the original set $\mathcal{A}$ as 2^n, where n is the desired size of the instance. All other integers, used to fill the set $\mathcal{A}$ until it reaches the desired size n, are generated by randomly choosing integers with a uniform distribution in the (open) interval $(1, 2^n)$.

In order to control the hardness of our instances, we also follow some additional recommendations given in [4] when generating our instances. The range of the target t is constrained and taken between

$$(n/2 - \sqrt{n}) \cdot 2^{n-1} \quad \text{and} \quad (n/2 + \sqrt{n}) \cdot 2^{n-1},$$

while approximately half of the original integers in $\mathcal{A}$ are chosen to take part to one pre-constructed SSP solution, which needs to satisfy the selected target t.

Our method for a random generation of hard instances begins by selecting the maximum integer in the set, as described above, and then selects the target t in the bounds given above. Only at this point does the random generation of the other integers begin, and the decision on whether or not to include each generated integer (with chance 1/2) in one associated solution is randomly made. However, the probability of defining a valid solution in this way is extremely unlikely (the randomly chosen integers should sum up exactly at t), and therefore a *correction* step is implemented in our generation algorithm, where the values of the integers selected to form an SSP solution are modified in order to have their sum correspond to the pre-defined target t.

Before delivering the generated instance, our method double-checks that the number of integers forming the constructed SSP solution actually concerns exactly the half of the integers in $\mathcal{A}$ (the only tolerated approximation is the one implied by an odd n). When this is not the case, integers belonging to the SSP solution are either randomly split or fusioned, while other integers (not belonging to the solution) may be erased or added in order to keep the cardinality of $\mathcal{A}$ equal to n.

TABLE I
SOME COMPUTATIONAL EXPERIMENTS WHERE THE GENERATED SSP INSTANCES ARE SOLVED BY A C IMPLEMENTATION OF THE BP ALGORITHM, BY GUROBI (MILP MODEL), BY THE SA SIMULATION PROVIDED BY D-WAVE, AND FINALLY BY D-WAVE ITSELF. THE VALUES OF THE TARGETS, FOR THE 4 INSTANCES, ARE $t_1 = 1024$, $t_2 = 3145728$, $t_3 = 5368709120$, AND $t_4 = 7696581394432$. THE VALUE OF ϵ INDICATES THE OBSERVED ERROR WITH THE RESPECT TO THE KNOWN SOLUTIONS. THE COMPUTATIONAL TIME IS GIVEN IN SECONDS.

Size/target	10	t_1	20	t_2	30	t_3	40	t_4
	ϵ	time	ϵ	time	ϵ	time	ϵ	time
BP	0	0.00	0	0.00	0	0.07	0	9.36
Gurobi	0	0.01	0	0.31	0	1.53	0	1138.00
SA	0	2.00	-8	4.00	-249	8.00	21959	15.00
D-Wave	0	0.02	11	0.25	20926	1.1	173294	2.80

B. Preliminary experiments

We provide some preliminary computational experiments where we solve 4 hard instances of the SSP, having 4 different sizes: 10, 20, 30 and 40. We limit ourselves to instances having size 40 because we have soon noticed that we had already hit the limits of the quantum computer when using instances having this relatively small size. Since our instances all have density equal to 1, the number of bits necessary to represent the elements of $\mathcal{A}$ in the 4 instances is 10, 20, 30 and 40, respectively. On standard 64-bit machines, therefore, standard primitive types, in the majority of programming languages, can hold values that are sufficiently large for our purposes. This is, however, not the case in the analog environment of D-Wave QPUs (Quantum Processing Units).

Our experiments are reported in Table I. The value of ϵ indicates the absolute error with respect to the known solutions, while the computational time is given in seconds. All experiments have been performed on a laptop computer equipped with an 11^{th} Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz, running Linux. We used Gurobi version 11, with default parameters except for `IntFeasTol` $= 10^{-9}$ and `NumericFocus` $= 1$. Concerning D-Wave QPUs, we considered two topologies, ZEPHYR and CHIMERA, and varied `num_reads` in $[2000, 4000]$ and `annealing_time` in $[40, 80]$. The results reported in Table I are the best over all the variations considered and over 10 runs/jobs.

We can remark that the BP algorithm, in spite of its worst-case exponential complexity, performs quite efficiently on a standard computer: it is able to solve *exactly* all SSP instances in a reasonable time. Gurobi applied to (2) also solved all instances but its drawback is the computation time. The Simulated Annealing (SA) meta-heuristics [14], implemented in the D-Wave standard Python package, appears to be slower than BP and faster than Gurobi, and provides solutions affected by large errors. This tendency is then amplified when running the experiments on D-Wave. The analog quantum computer is able to output its solution quickly (it is even faster than BP on the largest instances), but the given solutions are affected by much larger errors.

One interesting remark is that most of the solutions found by D-Wave are "local" minimizers, in the following sense. The concept of neighbouring, and hence the one of *locality*, makes sense in the context of the SSP only if we take into account the discrete nature of the search space. When we talk about local solutions, we make reference here to solutions that can be obtained from a provided solution x by performing only one bit flip: in other words, the solutions that are considered to be in the neighbourhood of x are those for which only one integer is added or removed from the current subset. We consider all D-Wave solutions to be local minimizers because they cannot be improved (with respect to the loss function (3)) by one single bit flip.

On D-Wave QPUs, the biases h_i and coupling parameters J_{ij} have specific physical ranges that the hardware can implement. These are typically floating-point numbers rather than integers, and they are restricted to a relatively small range: typically $h_i \in [-2.0, 2.0]$ and $J_{ij} \in [-1.0, 1.0]$. Besides, the D-Wave software interface has an automatic scaling feature which scales the problem data in order to fit in these intervals. Thus, if the problem data has a very large range, scaling them all down can make the smaller differences less distinguishable by the noisy analog hardware, potentially affecting solution quality.

We conjecture that, when D-Wave is actually not able to find the correct solutions, it is because of the representation errors implied by its analog nature. Part of our future works will be devoted to further studying this conjecture.

V. CONCLUSIONS

This paper focuses on the SSP, one of the most important, well-known and largely studied NP-complete problems. We have briefly surveyed two basic methods for the solution of SSP instances, provided a QUBO formulation of the problem, which we have then used for performing some preliminary computational experiments on D-Wave.

The main point in this work is the following: can we actually represent hard SSP instances on the quantum computer? Or is this rather possible only for some of the "easy" SSP instances? Our experiments show that an important difficulty is related to the accuracy in the representation of the integers in the set $\mathcal{A}$. As the number of bits necessary for representing the integers increases, it becomes harder and harder to have an error-free representation in the analog quantum computer. Our preliminary experiments seem to indicate that the primary impediment to quantum advantage in this domain is not algorithmic design but rather the limitations in representation in the analog computer.

Work is currently in progress to design methods for overcoming this representation issue. We will also study the possibility to represent and solve more general problems on D-Wave, such as for example the DDGP mentioned in the Introduction.

ACKNOWLEDGMENTS

This work was partially supported by the ANR project EVARISTE, which financed in May 2025 a visit of AM to the other co-authors at the Federal University of Santa Catarina. This paper is the result of the work performed during the visit. The authors are also grateful to the 5 reviewers for their fruitful comments.

REFERENCES

[1] T. Albash, D.A. Lidar, *Adiabatic Quantum Computation*, Reviews of Modern Physics **90**(1), 015002, 2018.

[2] J. Allcock, Y. Hamoudi, A. Joux, F. Klingelhöfer, M. Santha, *Classical and Quantum Algorithms for Variants of Subset-Sum via Dynamic Programming*, 30^{th} Annual European Symposium on Algorithms (ESA22), Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 28 pages, 2022.

[3] P. Austrin, P. Kaski, M. Koivisto, J. Nederlof, *Dense Subset Sum May Be the Hardest*, Leibniz International Proceedings in Informatics, 33^{rd} Symposium on Theoretical Aspects of Computer Science, Article No. 13, 13:1–13:14, 2016.

[4] D.J. Bernstein, S. Jeffery, T. Lange, A. Meurer, *Quantum Algorithms for the Subset Sum Problem*. In: Post-Quantum Cryptography, Proceedings 5, 16–33, 5^{th} International Workshop PQCrypto 2013.

[5] X. Bonnetain, R. Bricout, A. Schrottenloher, Y. Shen, *Improved Classical and Quantum Algorithms for Subset-Sum*, International Conference on the Theory and Application of Cryptology and Information Security, Springer International Publishing, 633–666, 2020.

[6] M.J. Coster, A. Joux, B.A. LaMacchia, A.M. Odlyzko, C.P. Schnorr, J. Stern, *Improved Low-density Subset Sum Algorithms*, Computational Complexity **2**, 111–128, 1992.

[7] M.G. Damaceno, A. Mucherino, R. Medeiros de Araújo, P.H. Souto Ribeiro, N. Rubiano da Silva, *Experimental Investigation of Optical Processing With Spatial Light Modulation*, arXiv preprint arXiv:2507.03821, 2025.

[8] F. Glover, G. Kochenberger, Y. Du, *A Tutorial on Formulating and Using QUBO Models*, arXiv preprint arXiv:1811.11538, 2018.

[9] A. Helm, A. May, *The Power of few Qubits and Collisions–Subset Sum below Grover's Bound*, Springer International Publishing, International Conference on Post-Quantum Cryptography, 445–460. 2020.

[10] S.B. Hengeveld, N. Rubiano da Silva, D.S. Gonçalves, P.H. Souto Ribeiro, A. Mucherino, *An Optical Processor for Matrix-by-Vector Multiplication: an Application to the Distance Geometry Problem in 1D*, Journal of Optics **24**(1), 015701, 2022.

[11] E. Horowitz, S. Sahni, *Computing Partitions with Applications to the Knapsack Problem*, Journal of the ACM (JACM) **21**(2), 277–292, 1974.

[12] R.M. Karp, *Reducibility Among Combinatorial Problems*. In: R.E. Miller, J.W. Thatcher, J.D. Bohlinger (Eds.), "Complexity of Computer Computations". New York: Plenum. 85–103, 1972.

[13] T. Kadowaki, H. Nishimori, *Quantum Annealing in the Transverse Ising Model*, Physical Review E, **58**(5), 5355, 1998.

[14] S. Kirkpatrick, C.D. Gelatt, M.P Vecchi, *Optimization by Simulated Annealing*, Science **220**, 671–680, 1983.

[15] L. Liberti, C. Lavor, N. Maculan, *A Branch-and-Prune Algorithm for the Molecular Distance Geometry Problem*, International Transactions in Operational Research **15**, 1–17, 2008.

[16] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[17] A. Mucherino, C. Lavor, L. Liberti, *The Discretizable Distance Geometry Problem*, Optimization Letters **6**(8), 1671–1686, 2012.

[18] J. Salas, *Beyond Worst-Case Subset Sum: An Adaptive, Structure-Aware Solver with Sub-$2^{n/2}$ Enumeration*, arXiv preprint arXiv:2503.20162, 2025.

[19] C.P. Schnorr, T. Shevchenko, *Solving Subset Sum Problems of Density close to 1 by Randomized BKZ-Reduction*, Cryptology ePrint Archive, 5 pages, 2012.

[20] Q. Zheng, Y. Miaomiao, Z. Pingyu, W. Yan, L. Weihong, X. Ping, *Solving the Subset Sum Problem by the Quantum Ising Model with Variational Quantum Optimization based on Conditional Values at Risk*, Science China Physics, Mechanics & Astronomy **67**(8), 280311, 2024.

[21] D-Wave Systems Inc., *Technical Description of the D-Wave Quantum Processing Unit*, `https://docs.dwavesys.com`, 2021.

[22] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, `https://www.gurobi.com`, 2024.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 641–646
DOI: 10.15439/2025F8584
ISSN 2300-5963 ACSIS, Vol. 43

Cognitive-Aware Peer Assessment: Design Implications from a Classroom Deployment

Naama Bouskila, Lihi Dery
0009-0001-1996-3451
0000-0002-8710-3349
Ariel University
Ariel, Israel
Email: {naama.bouskila, lihid}@ariel.ac.il

***Abstract*—Peer assessment is widely used in higher education, yet the cognitive demands placed on student assessors, particularly under conditions of overload and repetition, remain poorly understood. We examine how two cognitive factors, information overload and what we term assessment fatigue, influence evaluation behavior and user experience. Assessment fatigue is defined as cognitive strain resulting from repeated evaluative tasks. The study draws on data from a university-level deployment of a structured peer evaluation system. We applied Structural Equation Modeling (SEM) to analyze how behavioral data and self-reported perceptions of overload and fatigue relate to overall system satisfaction. Results reveal a significant indirect pathway from information overload to system satisfaction, mediated by fatigue. Based on these findings, we propose design recommendations for cognitively-aware assessment systems that adapt to students' cognitive constraints, contributing to the development of AI-supported educational tools that are more robust and human-centered.**

I. Introduction

PEER ASSESSMENT is widely used in higher education to enhance learning and promote critical thinking [14, 19]. As educational platforms increasingly integrate automated or semi-automated assessment tools, peer evaluation systems are becoming candidates for AI-supported enhancement [7, 8]. While the design of aggregation methods and elicitation strategies has recently received some attention [18, 21], the *cognitive experience* of the assessors—specifically the effects of overload and fatigue—has been largely overlooked. The cognitive demands placed on student assessors when reviewing multiple peer projects—remain poorly understood. As students progress through a sequence of evaluations, they may experience information overload and fatigue, which can affect the consistency of their ratings and their overall satisfaction with the process. This paper addresses that gap by focusing on two cognitive factors that may degrade the quality of peer evaluations over time: information overload (IOL) and what we term assessment fatigue. There is no single definition of information overload (see e.g., the survey paper by Bawden and Robinson [2]). A commonly accepted definition is that information overload occurs when a user has insufficient cognitive capacity to handle the presented information. IOL has been linked to decreased decision quality and increased inconsistency [9, 11]. These cognitive factors do not only impact the user experience—they may also compromise the accuracy of peer evaluations. Under high cognitive load, students may fail to distinguish between projects, default to repeated ratings, or reduce the depth of evaluative effort, potentially affecting the fairness and validity of assessment outcomes.

We introduce the term assessment fatigue to describe cognitive depletion stemming from repeated evaluative tasks within peer review sessions. While information overload is well established in cognitive psychology and organizational behavior, its role in peer assessment systems remains underexplored. The concept of fatigue in this context has received even less scholarly attention.

We study these phenomena using data collected from the R2R peer assessment platform [8], deployed in a university course on data analysis. In this setting, students assessed peer projects using a structured interface that combined Likert-scale ratings with pairwise comparison queries. Figure 1 illustrates the R2R system flow. Specifically, we ask: (1) What is the impact of information overload on peer assessment? (2) What is the impact of assessment fatigue on peer assessment? (3) How satisfied are users with the R2R peer assessment system? Our goal is to model the relationships between these cognitive factors and evaluation patterns, and to consider how future systems might be designed to account for such effects.

In addition to collecting behavioral data such as score patterns and tie frequency, we administered a post-assessment questionnaire that measured perceived information overload (IOL), assessment fatigue, and satisfaction with the system. To model the relationships among these subjective constructs and evaluation behavior, we applied Structural Equation Modeling (SEM), treating each construct as a latent variable.

We use the results of our behavioral analysis to motivate several design recommendations for improving peer assessment systems. These include incorporating fatigue-sensitive prompts and weighting inputs based on behavioral reliability. The goal is to help future systems better support assessors under cognitive strain and improve the overall consistency and fairness of evaluations. By combining behavioral modeling with system-level implications, this work contributes to the development of human-centered, AI-supported educational

This work was supported by the Ministry of Innovation, Science & Technology, Israel

technologies that account for the cognitive limitations of student evaluators. Specifically, our contributions are threefold:

- We define and estimate assessment fatigue as a latent variable using behavioral and self-report data.
- We apply Structural Equation Modeling (SEM) to analyze the relationships between information overload, assessment fatigue, and peer evaluation behavior.
- We propose design recommendations for peer assessment systems that address cognitive strain during evaluation.

II. Related Work

Peer assessment has been widely used in educational settings to support evaluation and feedback among students. Many systems have been developed to structure this process, aiming to improve reliability, fairness, and ease of use [19, 20, 14]. However, persistent challenges include strategic grading [18], inconsistency among assessors [21], and score inflation [15]. Rating-based systems are vulnerable to calibration issues and leniency bias [8], while systems that rely only on rankings can impose high cognitive demands on students [16].

The R2R platform addresses some of these limitations by combining Likert-scale ratings with pairwise comparisons, triggered only when students assign the same score to multiple projects [8]. This approach collects informative input while keeping the evaluation task manageable. The present study builds on this framework to examine how assessors' cognitive states—specifically overload and fatigue—affect their rating behavior and system satisfaction.

Information overload has been studied in decision-making, education, and organizational behavior [1, 6, 9], but its relevance to peer assessment remains underexplored. We also introduce the concept of assessment fatigue, defined as cognitive strain from repeated evaluation tasks. Although well-studied in cognitive psychology, fatigue has not been explicitly modeled in the peer assessment context. This study examines both constructs and their relationship to behavioral patterns and subjective user experience.

Previous studies in large-scale peer review, such as in MOOCs, have documented variability in rater performance over time [13, 17]. However, such work typically treats inconsistency as noise, rather than exploring its cognitive origins. In contrast, our study models cognitive factors as latent constructs and uses Structural Equation Modeling (SEM) to analyze their connection to evaluation behavior.

Recent work in educational AI explores how systems can adapt to learners' cognitive and emotional states [7], including adaptive pacing [22], feedback support [5], and scaffolding [10]. However, most peer assessment platforms still treat all students as equally capable reviewers. We argue that behavioral indicators of cognitive strain—such as reduced variability, scoring decay, or rushed evaluations—can inform the design of more responsive peer assessment systems.

III. Methodology

The study was conducted in an undergraduate data analysis course at Ariel University. Students participated in a structured peer assessment exercise using the R2R platform [8]. Each student viewed and evaluated 9–12 peer presentations. Evaluations began with assigning 1–5 Likert-scale scores to each project based on predefined rubrics. When multiple projects received the same score, the system triggered pairwise comparison queries to elicit ordinal preferences among them. This hybrid approach—combining cardinal ratings with selective ordinal input—yields a partially ordered set of preferences per student.

Once all data were collected, the system aggregated the individual inputs into a full ranking over all projects. For each median score level, it computed an internal ranking among the tied projects using either Borda or Copeland voting rules. The final output was a global project ranking that preserved both score-based judgments and fine-grained pairwise distinctions [8].

To capture students' evaluation behavior over time, we computed a set of behavioral indicators derived from their interactions with the R2R system. These included the mean score each student assigned across all evaluated projects, as well as a decay coefficient—defined as the difference between the mean of the first third and the last third of scores in the session—to estimate whether students' ratings declined over time. We also recorded the proportion of ties (i.e., repeated scores across projects). Another variable tested is response latency—measured as the average time between rating actions—was also extracted as a behavioral indicator. However, it did not show a significant association with any of the cognitive variables (overload, engagement, or satisfaction), and was therefore excluded from the final model.

Upon completing their peer assessments, students filled out a questionnaire including validated items on information overload, perceived fatigue, and system usability. Responses were measured using a five-point Likert scale and aggregated into three latent constructs: Information Overload, Assessment Fatigue, and Satisfaction. The full questionnaire can be found in the Appendix VI.

A total of 231 students participated in the study. Some did not rate all projects in their session, or failed to complete the questionnaire. We were thus left with a total of 194 students.

We employed Structural Equation Modeling (SEM). The model was pre-specified based on the theoretical assumptions on the relationships between cognitive factors (fatigue and overload), behavioral patterns, and subjective experience (satisfaction with the system). The SEM was implemented in R using the `lavaan` package, with maximum likelihood estimation. Model fit was assessed using standard indices: CFI, RMSEA, and SRMR. The latent variables were modeled as predictors of observed behavioral features to understand how cognitive strain manifests in evaluation patterns.

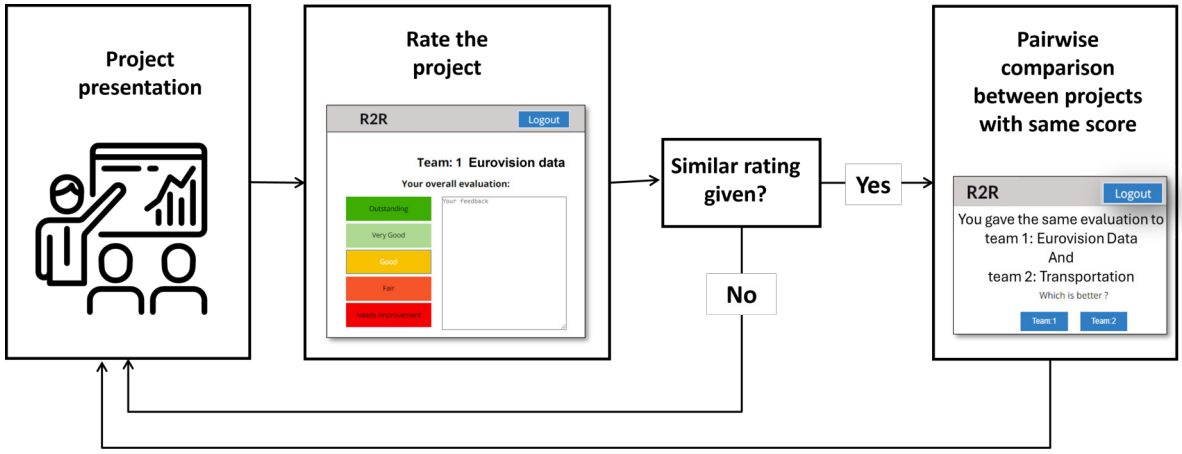


Fig. 1. R2R system flow. Students view project presentations and then rate the projects. If a student assigns the same rating to two or more projects, a series of pairwise comparison queries is executed.

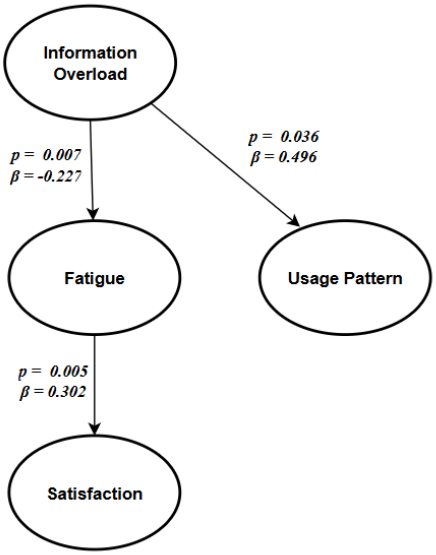


Fig. 2. Structural model results. Information overload predicts both assessment fatigue and usage patterns. Fatigue in turn predicts system satisfaction. A significant indirect effect from overload to satisfaction is mediated through fatigue.

IV. Results

The SEM demonstrated an acceptable to good fit with the data. The scaled chi-square test was marginally significant, $\chi^2(112) = 139.24$, $p = .041$. Additional fit indices supported the model: CFI = 0.934, TLI = 0.921, RMSEA = 0.035, and SRMR = 0.063.

Figure 2 presents a simplified diagram of the structural paths estimated in the model. The most robust finding was a positive and significant path from the latent fatigue construct to system satisfaction ($\beta = 0.302$, $p = .005$). Although labeled as fatigue, the items comprising this factor reflect cognitive engagement (e.g., active reflection and effort). Thus, higher scores indicate lower fatigue or greater evaluative involvement. Interpreted accordingly, the result suggests that students who were more cognitively engaged during the peer assessment process reported higher satisfaction with the system—a pattern consistent with prior findings on self-regulated learning and intrinsic motivation.

Information overload significantly predicted assessment fatigue ($\beta = -0.227$, $p = .007$), indicating that students who reported higher overload tended to also report greater fatigue. However, the direct path from information overload to system satisfaction was not statistically significant ($\beta = 0.012$, $p = .912$), suggesting that the relationship between overload and satisfaction operates indirectly, through its effect on fatigue.

The significance of the indirect path from information overload to satisfaction through fatigue ($ab_3 = -0.048$, $p = .019$) supports a partial mediation model. Fatigue acts as a conduit through which perceived overload affects downstream user experience. This finding aligns with cognitive theories that emphasize the cumulative toll of high information load and the importance of affective and physical strain in shaping task satisfaction [9, 1].

In the context of peer assessment, these mechanisms may help explain how students' evaluative consistency and system satisfaction decline as cognitive demands accumulate over the course of repeated assessment tasks.

In addition to the latent variables, the model incorporated behavioral measures. Information overload significantly predicted the mean rubric usage pattern ($ab_3 = 0.496$, $p = .036$), indicating that cognitive strain manifests not only in subjective reports but also in measurable changes in evaluation behavior. Students experiencing higher overload engaged differently with the interface, suggesting that behavior traces may offer real-time or asynchronous indicators of cognitive state.

Several additional paths were included in the model for completeness but were not statistically significant. These included the direct path from information overload to satisfaction

($ab_3 = 0.012$, $p = .912$), as well as paths from behavioral pattern indicators to satisfaction ($ab_3 = -0.053$, $p = .548$) and fatigue ($ab_3 = -0.070$, $p = .315$).

In summary, the model confirms that perceived information overload predicts fatigue, and that fatigue, in turn, positively predicts system satisfaction. The direct path from overload to satisfaction was not significant, indicating that fatigue mediates this relationship. These findings answer our core research questions and highlight the value of modeling cognitive constructs such as overload and fatigue as latent variables. The use of SEM allowed us to capture latent constructs that are not directly observable, such as perceived fatigue and overload, and to examine their relationships with both self-reported satisfaction and system-logged behavioral patterns. This modeling approach made it possible to detect indirect effects and account for measurement error, offering a more accurate view of how cognitive states shape peer assessment experience.

V. Design Recommendations

Assessment fatigue, i.e., cognitive fatigue during peer assessment, can be addressed through real-time detection mechanisms that rely on unobtrusive behavioral signals. Because the assessment environment already logs granular student actions—such as response times, ratings, and tie-breaking decisions—it is feasible to infer signs of strain or disengagement as students work through their evaluation tasks. Specifically, indicators such as repeated use of a single score (e.g., consistently assigning the highest rating), low variance across rubric dimensions, skipping optional comment fields, or evidence of scoring decay over time may suggest cognitive overload or reduced attentional engagement. The latter can be quantified using a decay coefficient, defined as the difference in mean scores between the first and last segments of a student's evaluation sequence. A consistent downward trend in scores, relative to how other assessors rated the same projects, may reflect increasing fatigue or diminished evaluative attention as the session progresses.

When such patterns are detected in real time, the system can scaffold the rating process to reduce cognitive load. Instead of allowing students to proceed with rapid, unreflective scoring, the interface can present brief prompts encouraging deeper engagement. For example, when repeated use of the same score is detected, the system might ask: "Consider pacing your next few ratings more slowly to ensure accuracy" or "Before submitting, consider whether this score reflects the project's clarity, originality, and completeness." Other prompts might invite rubric-based reflection, such as: "Did this project meet expectations for structure, argumentation, or use of data?". These light-weight nudges are designed to trigger evaluative reappraisal without interrupting workflow. Drawing on principles of adaptive scaffolding and self-regulated learning, such prompts can help students maintain attentional focus and calibrate their ratings more accurately, even under conditions of assessment fatigue.

In addition to in-the-moment interventions, cognitive strain can be addressed retrospectively through post-hoc adjustments to aggregation. After the evaluation session concludes, the system can compute a reliability score for each assessor based on fatigue-related behavioral signals. This score, scaled between 0.1 and 1.0, reflects the inferred trustworthiness of the student's inputs. These weights can then be used to adjust both rating-based and ranking-based aggregation methods. In the case of ratings, a student who assigns the same score to all projects or shows minimal rubric variance may have their evaluations down-weighted in the final average. For ranking-based aggregation, methods such as Borda or Copeland can be adapted to incorporate these weights directly. The adjusted Borda score for a project c_j would be computed as:

$$\text{AdjustedBorda}(c_j) = \sum_{i=1}^{n} w_i \cdot \text{BordaRank}_i(c_j)$$

where w_i is the reliability weight for student i, and $\text{BordaRank}_i(c_j)$ is the rank assigned by that student to project c_j. This post-hoc adjustment strategy requires no change to the frontend interface or assessment process.

VI. Discussion

Real-time scaffolding and reliability-based weighting support student wellbeing by recognizing and mitigating the cognitive toll of extended evaluation tasks. These adaptive mechanisms may also enhance fairness, by reducing the influence of assessors operating under strain. Thus, integrating fatigue-sensitive design into peer assessment systems carries not only cognitive and pedagogical value, but also reflects human-centered AI principles—namely, respect for cognitive limits, equitable treatment, and minimal intrusion.

Several directions remain open for extending this work. One promising avenue is the integration of large language models (LLMs) to enhance feedback analysis and detect signs of fatigue or disengagement through natural language patterns. While the present study focuses on rating behavior and assessment structure, written feedback can offer a complementary signal of cognitive strain. For example, LLMs can be used to analyze comment length, sentence count, lexical richness, or syntactic complexity. Qualitative shifts such as sentiment flattening ("nice job") or use of generic, non-specific phrases ("good project," "well done") may indicate reduced engagement or cognitive effort. Detecting such patterns at scale could enable a multi-modal fatigue detection pipeline that combines behavioral and textual indicators.

A second line of future work involves implementing the design recommendations proposed in this paper within a live peer assessment platform. This includes integrating fatigue-sensitive prompts, real-time scaffolding for ratings, and post-hoc reliability-weighted aggregation. Controlled experiments can then evaluate the effectiveness of these interventions in improving evaluation quality and fairness of outcomes.

ACKNOWLEDGMENT

This research was supported by the Ministry of Innovation, Science & Technology, Israel.

REFERENCES

[1] Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74(5):1252–1265, 1998. https://doi.org/10.1037/0022-3514.74.5.1252.

[2] David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science*, 35(2):180–191, 2009. https://doi.org/10.1177/0165551508095781.

[3] Gee-Woo Bock, Mimrah Mahmood, Sanjeev Sharma, and Youn Jung Kang. The impact of information overload and contribution overload on continued usage of electronic knowledge repositories. *Journal of Organizational Computing and Electronic Commerce*, 20(3):257–278, 2010. https://doi.org/10.1080/10919392.2010.494530.

[4] Chao-hsiu Chen. The implementation and evaluation of a mobile self-and peer-assessment system. *Computers & Education*, 55(1):229–236, 2010. https://doi.org/10.1016/j.compedu.2010.01.008.

[5] Olga Chernikova, Daniel Sommerhoff, Matthias Stadler, Doris Holzberger, Michael Nickl, Tina Seidel, Enkelejda Kasneci, Stefan Küchemann, Jochen Kuhn, Frank Fischer, et al. Personalization through adaptivity or adaptability? a meta-analysis on simulation-based learning in higher education. *Educational Research Review*, page 100662, 2024. https://doi.org/10.1016/j.edurev.2024.100662.

[6] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011. https://doi.org/10.1073/pnas.1018033108.

[7] Amir Darvishi, Hadi Khosravi, Shazia Sadiq, and Dragan Gašević. Incorporating ai and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, 2022. https://doi.org/10.1111/bjet.13233.

[8] Lihi Dery. Interactive and iterative peer assessment. In *Proceedings of the European Conference on Artificial Intelligence (ECAI), https://doi.org/10.3233/FAIA240656*, 2024.

[9] Martin J. Eppler and Jeanne Mengis. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society*, 20(5):325–344, 2004. https://doi.org/10.1080/01972240490507974.

[10] Frank Fischer, Elisabeth Bauer, Tina Seidel, Ralf Schmidmaier, Anika Radkowitsch, Birgit J Neuhaus, Sarah I Hofer, Daniel Sommerhoff, Stefan Ufer, Jochen Kuhn, et al. Representational scaffolding in digital simulations–learning professional practices in higher education. *Information and Learning Sciences*, 123(11/12):645–665, 2022. https://doi.org/10.1108/ILS-06-2022-0076.

[11] Janusz A Hołyst, Philipp Mayr, Michael Thelwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N Kenett, Denis Helic, Aljoša Rehar, Sebastijan R Maček, et al. Protect our environment from information overload. *Nature Human Behaviour*, 8(3):402–403, 2024. https://doi.org/10.1038/s41562-024-01833-8.

[12] Felix Krieglstein, Maik Beege, Günter Daniel Rey, Christina Sanchez-Stockhammer, and Sascha Schneider. Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, 35(1):9, 2023. https://doi.org/10.1007/s10648-023-09738-0.

[13] Chinmay Kulkarni, Kristine P. Wei, Harini Le, Daniel Chia, Katharine Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6):1–31, 2013. https://doi.org/10.1145/2505057.

[14] H. Li, Y. Xiong, C. V. Hunter, X. Guo, and R. Tywoniw. Does peer assessment promote student learning? a meta-analysis. *Assessment & Evaluation in Higher Education*, 45(2):193–211, 2020. https://doi.org/10.1080/02602938.2019.1620679.

[15] Ernesto Panadero, M. Romero, and J.-W. Strijbos. The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, 39(4):195–203, 2013. https://doi.org/10.1016/j.stueduc.2013.10.005.

[16] Kumar Raman and Thorsten Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1046, 2014. https://doi.org/10.1145/2623330.2623654.

[17] Nihar B. Shah, James K. Bradley, Ameet Parekh, Martin J. Wainwright, and Kannan Ramchandran. A case for ordinal peer evaluation in moocs. In *NIPS Workshop on Data Driven Education*, 2013.

[18] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. Catch me if i can: Detecting strategic behaviour in peer assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4794–4802, 2021. https://doi.org/10.1609/aaai.v35i6.16611.

[19] Keith Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.

[20] Keith J. Topping. Peer assessment. *Theory Into Practice*, 48(1):20–27, 2009.

[21] Toby Walsh. The peerrank method for peer assessment. In *ECAI*, pages 909–914, 2014. https://doi.org/10.3233/978-1-61499-419-0-909.

[22] Bahman Zohuri and Farhang Mossavar-Rahmani. Revolutionizing education: The dynamic synergy of personalized learning and artificial intelligence.

International Journal of Advanced Engineering and Management Research, 9(1):143–153, 2024. https://doi.org/10.51505/ijaemr.2024.9111.

APPENDIX: QUESTIONNAIRE

Students completed a questionnaire immediately after finishing the peer assessment tasks. It included two demographic questions, six items from a validated satisfaction scale [4], four items from an information overload scale [3], and five items on fatigue adapted from Krieglstein et al. [12].

Response Scales:

- **5-point scale:** 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree
- **7-point scale:** 1 = Strongly Disagree, 2 = Disagree, 3 = Somewhat Disagree, 4 = Neither Agree nor Disagree, 5 = Somewhat Agree, 6 = Agree, 7 = Strongly Agree

General Information

- Age
- Gender

Satisfaction [4] *(5-point scale)*

- R2R was more efficient for peer and self-evaluation than paper-and-pencil methods would be.
- The operation of R2R evaluations was easy and convenient.
- I tried to grade my peers appropriately and not to be picky or hypocritical.
- I worried that my classmates would give me low grades intentionally.
- I think the instructor's comments and grading would be more professional than my classmates'.
- I felt that I received biased low scores from peers.

Information Overload [3] *(7-point scale)*

- I cannot effectively process all the information I read.
- I am overwhelmed with the information I read.
- I feel anxious that I might have missed an important piece of information I read.
- I cannot assimilate all the information I read.

Fatigue [12] *(7-point scale)*

- The learning content was difficult to understand.
- The learning content included much complex information.
- I actively reflected upon the learning content.
- I made an effort to understand the learning content.
- I was able to expand my prior knowledge with the learning content.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 647–652
DOI: 10.15439/2025F6352
ISSN 2300-5963 ACSIS, Vol. 43

Machine learning for survival analysis: a comparative study on intensive care unit (ICU) patient data and simulations

Lukáš Boček[1, 2]
ORCiD: 0009-0006-9576-715X
[1]Department of Statistics and Probability
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
Email: l.bocek9825@gmail.com
&
[2]DataSentics, Inc.
Washingtonova 1599/17, 11000 Prague, Czech Republic
Email: lukas.bocek@datasentics.com

Lubomír Štěpánek[1, 2, 3]
ORCiD: 0000-0002-8308-4304
[1]Department of Statistics and Probability
[2]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
Email: lubomir.stepanek@vse.cz
&
[3]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, 12000 Prague, Czech Republic
Email: lubomir.stepanek@lf1.cuni.cz

***Abstract*—Survival analysis focuses on modeling the time until a specific event occurs, often in the presence of censored observations. While classical methods like the Cox model are widely used, modern machine learning (ML) approaches offer greater flexibility and predictive power. This paper compares classical and ML-based survival models on both real-world and simulated datasets. We demonstrate that techniques like CoxBoost and penalized Cox regression outperform tree-based models like Random Survival Forests in most settings. Explainable Artificial Intelligence (AI) tools are applied to improve the transparency and interpretability of model predictions.**

***Index Terms*—Survival analysis, machine learning, Cox proportional hazards model, random survival forests, CoxBoost, penalized Cox model, ICU patients, explainable artificial intelligence**

I. Introduction

SURVIVAL analysis is a branch of statistics concerned with the time until an event of interest occurs. The event of interest can be, for example, the death of a patient or the time until the failure of a machine. In this situation, a phenomenon called data censoring is often present, as typically, not all units experienced an event of interest during the monitoring period.

The most widely used model in survival analysis is the Cox model, introduced in 1972 by [1]. The Cox model does not make assumptions regarding the hazard function, which provides necessary flexibility, but makes parametric assumptions about the effect of the covariates on the hazard function. The assumptions of the model are relatively strict and may lead to biased results if violated. This is often pointed out as a potential drawback.

Machine learning is a subfield of artificial intelligence that leverages knowledge from statistics and computer science. In the last decades, it has established itself as a highly flexible and powerful tool capable of solving a wide variety of problems. In general, machine learning requires little to no assumption to be made and is able to discover complex relations in the data. This raises an important question: How can survival analysis benefit from incorporating machine learning methods. Addressing this question is the primary focus of this work.

This paper presents a comprehensive evaluation of six models: Cox proportional hazard model, Cox model with elastic net regularization, CoxBoost, Random Survival Forests (RSF), Conditional Inference Forests (CIF), and Oblique RSF. We evaluated their performance on a real-world Intensive Care Unit (ICU) dataset as well as on simulated datasets with known ground truth.

II. Related work

The use of machine learning and statistical methods for survival analysis has been attracting a lot of attention in recent years. Survival clustering analysis for time-to-event data was proposed in [2], robust nonparametric survival curves comparison in [3], [4], while support vector machines in [5], [6], together with implementation in R. All the above mentioned approaches offer rather a methodological framework on usage the ML in survival analysis.

Another approach how to utilize machine learning in survival analysis appeared in a publication [7], where the authors decomposed the dependent two-dimensional time-event variable into two components. Thereafter, the authors separated

the analysis into two parts, the occurrence of the event being a classification task and time-to-event estimation being a regression task. This approach was later improved and used for the prediction of a COVID-19 blood antibody decrease, which could help to identify individuals who should receive boosting vaccination when a new variant of COVID-19 would appear [8].

Furthermore, the use of deep learning techniques in the field of survival analysis started to seem promising and become one of the main topics for research. In an article [9], the authors utilized neural networks in order to create a personalized treatment recommender system. Authors in an article [10] employed deep learning techniques for competing risk analysis. A comprehensive review about deep learning algorithms used in survival analysis was lately given by [11], which also discusses deep learning algorithms capable of handling time-dependent covariates. Additionally, the review outlines how deep learning can be used to integrate multimodal data – including image, text, and tabular inputs – into survival modeling.

III. Methodology

This section outlines the analytical framework used in our study. We detail the specific traditional and machine learning survival models evaluated, the metrics employed for assessing their predictive performance, and the explainable AI techniques used to interpret their behaviour.

A. Survival and ML Models

We evaluated six survival analysis models from both traditional and machine learning families:

- **Cox Proportional Hazards (Cox PH)**: The most popular model for the analysis of survival data is the Cox proportional hazards model (in literature, it often appears as the Cox PH model or simply the Cox model). The Cox model is a semiparametric model, it makes no assumptions about the hazard function but makes a parametric assumption regarding the effect of the covariates on the hazard function. The fact that the hazard function is not needed provides a significant advantage, as in many situations the true form of the hazard function might be unknown or too complex [12].
  The Cox model is most often stated in the following form,
  $$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}\boldsymbol{\beta}), \tag{1}$$
  where
  - $h(t|\mathbf{x})$ is the hazard function at time t given covariates $\mathbf{x}$,
  - $h_0(t)$ is the baseline hazard function at time t,
  - $\mathbf{x}$ is the vector of covariate (i.e., it is eventually a matrix),
  - $\boldsymbol{\beta}$ is the vector of regression coefficients.
- **Random Survival Forests (RSF)**: Random survival forests (RSF) is a method introduced in article [13], and it presents modification for analysis of right-censored survival data of the popular random forest algorithm. The authors argue that extending the random forest algorithm for the purpose of analyzing survival data is of great value as with traditional methods, nonlinear effects of variables must be modeled by transformations, and ad hoc approaches are often needed. On the other hand, random forests handle these difficulties automatically.
  The RSF algorithm, as described in [13] is as follows:
  (i) Draw B bootstrap samples from the original data. Each bootstrap sample excludes, on average, 37% of the data, these data are called out-of-bag data (OOB data).
  (ii) Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p variables as candidates for the splitting. The node is split using the variable that maximizes the survival difference between the child nodes.
  (iii) Grow the tree under the constraint that each terminal node should contain at least one unique event of interest.
  (iv) Calculate a cumulative hazard function (CHF) for each tree. Obtain the ensemble cumulative hazard function by averaging across all trees.
  (v) Use the OOB data to compute the prediction error for the ensemble CHF.
- **CoxBoost**: CoxBoost is an extension of the Cox model that incorporates boosting algorithm introduced in [14]. Boosting is a popular, iterative ensemble method that builds a strong predictive model by sequentially combining multiple weak models. The boosting algorithm initially sets equal weights to all observations, then for each successive (m-th) iteration, where $m \in \{2, 3, \ldots, M\}$, the weights for each observation are modified. At each (m-th) step, the misclassified observations have their weights increased, whereas correctly classified observations have their weights decreased. As the algorithm proceeds, observations that are difficult to classify correctly gain greater influence [15].
- **Penalized Cox (Elastic Net)**: As the amount of collected data rapidly grows, supported by advancements in detection techniques, high-dimensional settings are becoming increasingly common across most domains. For classical cases, with significantly more observations than predictors, the Cox model tends to work well. However, in situations where the number of predictors is close to or even exceeds the number of observations, the Cox model tends to output degenerate behavior, as all parameters β_j are converging towards $\pm\infty$, [16].
  To address this challenge, Simon et al. [16] proposed an algorithm that incorporates the elastic net penalty as
  $$\lambda\left(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i^2\right), \quad \alpha \in [0,1], \tag{2}$$
  which is to be minimized by the algorithm and is a mixture of the L_1 (LASSO) and L_2 (ridge regression) penalties. LASSO tends to work well for sparsity problems, as it tends to choose only a few nonzero coefficients. On the other hand, in the presence of correlations between the predictors, ridge regression tends to perform better, but sets no coefficients to exactly zero. The elastic net combines the strength of both approaches and as α changes between 0 and 1, the approach is changing to be more ridge-like or more LASSO-like [16].

- **Conditional Inference Forests (CIF)**: Conditional inference trees were introduced in [17]. The authors argue that the majority of recursive partitioning algorithms, such as CART [18], have fundamental problems in overfitting and selection bias towards covariates with many possible splits [17].
  To address these concerns, conditional inference trees adopt a more statistically principled approach that takes into account the distributional properties of the data. The core of this approach lies in determining the association between covariates and the response, which allows for an unbiased selection of variables, irrespective of their scale or the number of categories they might have [17].
- **Oblique Random Survival Forests (ORSF)**: A potential modification of the random survival forests described above is extending the splits to allow a linear combination of the predictors instead of using only a single predictor. This extension is applied in oblique random survival forests, as described in [19]. Authors argue that the described approach might improve the random survival forests especially when the predictors are correlated and some are irrelevant to the survival outcome, which is often the case when working, for example, with large medical databases.

B. Model Evaluation

In order to assess model performance, there is a need for metrics that can quantify the prediction performance of the survival models. A commonly used measure is *Harrell's concordance index* [20], which measures how well the model ranks two random individuals in terms of survival time. It is defined as a ratio of concordant pairs (correctly ordered pairs) to all comparable pairs. In order for the random pair i and j to be comparable, the sample with lower observed time y had to experience an event, i.e., $y_j > y_i$ and $\delta_i = 1$, where δ_i is a binary event indicator. A pair is concordant if the estimated risk predicted by the survival model is higher for observations with lower survival time (in other words, if the survival time is in general low, the risk had to be high because it shortened the survival time). Otherwise the pair is called discordant. The concordance index is then computed as

$$C = \frac{\text{number of concordant pairs}}{\text{number of all comparable pairs}}. \tag{3}$$

C. Explainability

With the introduction of machine learning methods into survival analysis, there emerged a need to deepen our understanding of these algorithms. As stated in [21], the successful adoption of machine learning in healthcare critically depends on techniques that enhance our ability to interpret model outputs.

In response to this challenge, an entire subfield has developed, commonly known as Explainable AI (XAI). This subfield focuses on creating techniques that make machine learning and artificial intelligence solutions more transparent and provide better understanding of the process behind decision making of the model [22].

However, until recently, these techniques were not designed to handle censoring and provide explanations for survival models. To fill this gap, [23] proposed a `survex` framework implemented into R, with the goal of empowering stakeholders with model understanding and building trust in machine learning models. The `survex` framework is designed to be model agnostic and, hence, can be applied to any survival model that returns predictions in the form of survival or cumulative hazard function [23].

Furthermore, methods dedicated to explain survival models such as SurvSHAP(t) [24] or SurvLIME [25] are also incorporated in the `survex` package along with explanations tailored to incorporate the time dimension. Moreover, local explanations that refers to prediction as well as global explanations regarding the whole model are provided [23].

D. Real-world data

Real-world dataset chosen for the purpose of the practical application of survival analysis methods, was MIMIC-III [26]. MIMIC-III is a large relational database consisting of 26 tables containing information about 38,597 adult patients and 7,870 neonates that were admitted to critical care units (ICU) to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012. Information about patients' demographics, diagnoses, laboratory tests, physiological measurements, drug codes, and many more are collected. Different types of information are kept separately in different tables [26].

The dataset underwent extensive preprocessing, the details of which are beyond the scope of this paper but are reproducible and available upon request. The final dataset contained 11,435 unique patients out of which 5,036 experienced an event of interest (i.e., death resulting from a severe health condition). This equals to a censoring level of roughly 56 %.

E. Simulated data

The goal of the analysis in the simulated dataset was to further assess the qualities of the individual algorithms in simulation settings that represent different scenarios. Another objective was to examine how well the models were able to recognize variables that have a real impact on the outcome and which models are more likely to get confused by variables representing only random noise. This can be done only under the simulation study, where these variables are known ahead.

The simulated dataset was created using standard Weibull distribution, which incorporates time-dependent effects of the variables. For simplification, a formula of how such a generating model can look for one covariate is illustrated in equation (4). In this example, it is assumed that the time-dependent effect of the covariate is induced by interaction with log time [27],

$$h_i(t) = \gamma \lambda t^{\gamma-1} \exp(\beta_0 X_{i,j} + \beta_1 X_{i,j} \cdot \log(t)), \tag{4}$$

where

- $h_i(t)$ is the hazard function for the i-th individual at time t,
- γ is the shape parameter of the Weibull distribution,
- λ is the scale parameter of the Weibull distribution,
- $X_{i,j}$ is the j-th covariate value for i-th individual,
- β_0 is log hazard ratio for the j-th covariate,
- β_1 specifies the amount by which the log hazard ratio changes for the j-th covariate, for every unit increase in $\log(t)$.

To test model robustness in noisy environments, a dataset containing 32 variables was created, with only 8 of these variables having a real impact on the outcome. The values of each variable were simulated from common probability distributions such as normal, log-normal or Bernoulli.

IV. Results and Discussion

A. Real-world data

For the analysis of real-world data, 80/20 train/test split was conducted and patient's diagnosis and demographics were used as predictors.

On the test set, RSF performed slightly better initially, but Cox-based models (especially CoxBoost and Cox with elastic net regularization) showed better or equal performance over the long term. Overall, performance was very similar.

Fig. 1 display comparison of the performance, using concordance index, over all used models. We observe that performance stability, particularly at later time points (when only a smaller number of individuals who have not experienced the event of interest remain, making prediction more difficult), is achieved by the CoxBoost and Penalized Cox models, whereas the traditional Cox proportional hazards model demonstrates poorer predictive performance.

Traditional Cox model found age, gender (males having 11.5% higher risk), admission type (emergency/urgent increasing risk), and diagnosis (sepsis being most dangerous, increasing risk by 33.8 % vs. Altered Mental Status (AMS)) to be significant predictors.

Random Survival Forest identified age, diagnosis, and admission type as most important, but ranked ethnicity/insurance higher than gender, unlike Cox model.

For better understanding of the model's behaviour, we used the R package `survex` [23] and displayed the predicted survival function for a new patient for different values of the variable of our interest (diagnosis). In this case, we identified the most typical patient who experienced the event of interest (death) as a 73-year-old female who, in reality, was diagnosed with congestive heart failure and died 71 days after the first admission. The practical implications of this explainable finding are straightforward – a newly admitted patient who is a 73-year-old female should be treated with maximum care, as she is at the highest risk of death given this diagnosis.

This demonstrates how powerful the machine learning algorithms could be in understanding complex relationships, which are often present in real-world data.

As was mentioned in the methodology section, the Cox model is constrained by strong assumptions – one of which being the proportional hazard ratios. This constraint is evident in Fig. 2, where the survival curves for individual diagnoses do not intersect. In contrast, the Random Survival Forest (RSF) model imposes no such assumption, enabling more flexible modeling of time-dependent effects. This flexibility is illustrated in Fig. 3 – for example, pneumonia initially presents lower risk compared to seizure or intracranial hemorrhage, but its predicted risk eventually exceeds both over time.

B. Simulated data

The algorithms used in the simulation part were the same as the algorithms used in real-world data analysis. Model performance was evaluated by concordance index on the test set after splitting 80 % of the data for training and 20 % for testing, can be seen in Table I. The size of the dataset before splitting is denoted by n. The best performance under each setting is bolded.

Initially, a dataset with only 100 observations was created, leaving only about 80 for training and 20 observations for

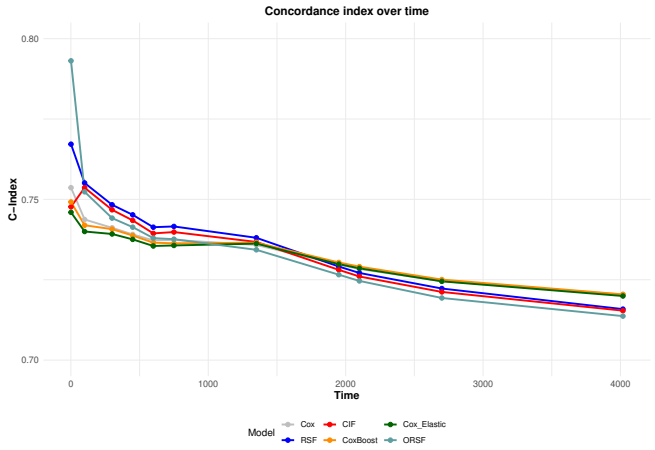


Fig. 1. Concordance index over time for all models.

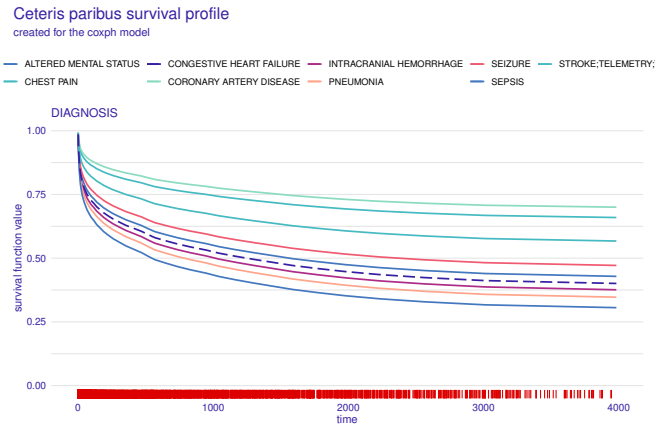


Fig. 2. Predictions for patient 16680 under different diagnoses – Cox PH model.

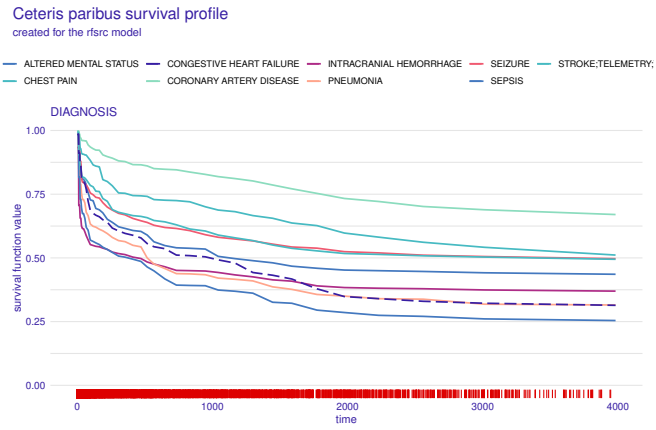


Fig. 3. Predictions for patient 16680 under different diagnoses – RSF.

testing. Under the Weibull distribution with scale 0.025 and shape 1.0, the Cox model with elastic net regularization performed the best on the test set. However, as the Weibull distribution was changed, presenting more risky environments with less censoring, the traditional Cox model surpassed the Cox model with regularization. Moreover, it is interesting to note that under the Weibull distribution with a parameter of scale 0.070 and shape 1.1, the conditional inference forests, the CoxBoost and the oblique random survival forest all had concordance index below 0.5, meaning that it was worse than random guessing (!).

Subsequently, the number of instances was increased to 2500 in order to see how a bigger dataset would influence the performances of each algorithm and also to see if the level of censoring would still be as influential as it was with the small dataset. First of all, from Table I, it is visible that once the size of the dataset is significantly increased, the influence of the censoring is decreased, and the performances of the algorithms do not tend to differ that much among different distribution settings. The best performing models, under this size of the dataset, were the CoxBoost and the Cox model with elastic net regularization.

TABLE I
CONCORDANCE INDEX FOR EACH OF THE ALGORITHM UNDER EACH SETTING

Weibull parameters	W(1.0, 0.025)	W(1.0, 0.060)	W(1.1, 0.070)
n=100			
Cox	0.788	**0.701**	**0.625**
RSF	0.765	0.591	0.554
CIF	0.671	0.468	0.391
CoxBoost	0.782	0.552	0.489
Cox_elastic	**0.882**	**0.701**	0.620
ORSF	0.647	0.565	0.467
n=2500			
Cox	0.619	0.641	0.647
RSF	0.623	0.641	0.635
CIF	0.608	0.628	0.632
CoxBoost	**0.633**	0.634	**0.654**
Cox_elastic	0.620	**0.642**	0.649
ORSF	0.621	0.638	0.640
n=5000			
Cox	0.638	0.627	0.626
RSF	0.629	0.618	0.614
CIF	0.636	0.625	0.619
CoxBoost	0.640	0.633	0.627
Cox_elastic	**0.641**	**0.634**	**0.630**
ORSF	0.628	0.627	0.622

Finally, the size of the dataset was even more increased to 5000 observations. This was done in order to see whether adding more instances would result in better model performance. In comparison with the setting, where only 2500 observations were generated, the performances improved only a little and, in some cases, got even a little bit worse. From that we can conclude that beyond a certain point, adding additional observations, which come from the same distribution, is bringing smaller and smaller value. The Penalized Cox model appears to be the best-performing model for large datasets, which aligns with our expectations, as penalization can effectively address the imbalance between the number of observations and covariates.

Additionally, we assessed each model's ability to distinguish between informative variables and random noise under different simulation settings. Under the settings of only 100 observations, algorithms like the RSF or the traditional Cox model tended to struggle in recognizing important variables and were only able to do better once the number of observations was significantly increased. On the other hand, the *Cox model with elastic net regularization was able to recognize the majority of important variables even under the setting of only 100 observations*. This likely contributed to its superior performance under that setting.

V. CONCLUSION AND FUTURE RESEARCH IDEAS

Overall, machine learning approaches demonstrated promising performance that was at least competitive with the traditional Cox model on both real-world and simulated data. Moreover, machine learning models offer valuable flexibility in capturing complex, non-linear relationships between covariates and survival outcomes. Applying techniques such as elastic net regularization combined with the Cox model under a penalized framework proved especially effective in scenarios with limited data availability or an imbalance between observations and covariates.

Future research could explore deep learning and multimodal data integration, which may enhance performance by providing richer patient information. Additionally, leveraging time-dependent covariates could reveal important patterns from measurements collected during a patient's stay. Moreover, future work should incorporate formal statistical testing (e.g., bootstrapped confidence intervals for the concordance index) to enable more rigorous model comparisons. While this study focused on a single ICU dataset, evaluating models across multiple real-world datasets would help assess generalizability across healthcare settings.

VI. Acknowledgment

This paper is supported by the grant IG410035 with no. F4/51/2025, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: http://www.jstor.org/stable/2985181

[2] P. Chapfuwa, C. Li, N. Mehta, L. Carin, and R. Henao, "Survival cluster analysis," Feb. 2020. doi: https://dx.doi.org/10.48550/ARXIV.2003.00355

[3] L. Štěpánek, F. Habarta, I. Malá, and L. Marek, "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, vol. 21. IEEE, Sep. 2020. ISSN 2300-5963 p. 453–460. [Online]. Available: http://dx.doi.org/10.15439/2020F198

[4] ——, "Non-parametric comparison of survival functions with censored data: A computational analysis of greedy and monte carlo approaches," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, vol. 39. Polish Information Processing Society, Oct. 2024. ISSN 2300-5963 p. 725–730. [Online]. Available: http://dx.doi.org/10.15439/2024F223

[5] V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel, "Support vector machines for survival analysis," 2007.

[6] C. Fouodo, I. König, C. Weihs, A. Ziegler, and M. Wright, "Support vector machines for survival analysis with r," *The R Journal*, vol. 10, no. 1, p. 412, 2018. doi: https://dx.doi.org/10.32614/rj-2018-005

[7] L. Štěpánek, F. Habarta, I. Malá, L. Marek, and F. Pazdírek, "A machine-learning approach to survival time-event predicting: Initial analyses using stomach cancer data," in *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, Oct. 2020. doi: https://dx.doi.org/10.1109/ehb50910.2020.9280301

[8] L. Štěpánek, F. Habarta, I. Malá, L. Štěpánek, M. Nakládalová, A. Boriková, and L. Marek, "Machine learning at the service of survival analysis: Predictions using time-to-event decomposition and classification applied to a decrease of blood antibodies against covid-19," *Mathematics*, vol. 11, no. 4, p. 819, Feb. 2023. doi: https://dx.doi.org/10.3390/math11040819

[9] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, Feb. 2018. doi: https://dx.doi.org/10.1186/s12874-018-0482-1

[10] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. doi: https://dx.doi.org/10.1609/aaai.v32i1.11842

[11] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, "Deep learning for survival analysis: a review," *Artificial Intelligence Review*, vol. 57, no. 3, Feb. 2024. doi: https://dx.doi.org/10.1007/s10462-023-10681-3

[12] F. E. J. Harell, *Regression Modeling Strategies*, 2nd ed., ser. Springer eBook Collection. Cham: Springer, 2015. ISBN 9783319194257

[13] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics 2008, Vol. 2, No. 3, 841-860*, vol. 2, no. 3, Sep. 2008. doi: https://dx.doi.org/10.1214/08-aoas169

[14] H. Binder and M. Schumacher, "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models," *BMC Bioinformatics*, vol. 10, no. 1, Jan. 2009. doi: https://dx.doi.org/10.1186/1471-2105-10-18

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. ISBN 978-0-387-84857-0

[16] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, 2011. doi: https://dx.doi.org/10.18637/jss.v039.i05

[17] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Sep. 2006. doi: https://dx.doi.org/10.1198/106186006x133933

[18] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen, *Classification and Regression Trees*, 1st ed. New York: Chapman and Hall/CRC, 1984. ISBN 0-534-98053-8

[19] B. C. Jaeger, S. Welden, K. Lenoir, J. L. Speiser, M. W. Segar, A. Pandey, and N. M. Pajewski, "Accelerated and interpretable oblique random survival forests," *Journal of Computational and Graphical Statistics*, vol. 33, no. 1, pp. 192–207, Aug. 2023. doi: https://dx.doi.org/10.1080/10618600.2023.2231048

[20] F. E. Harrell, "Evaluating the yield of medical tests," *JAMA: The Journal of the American Medical Association*, vol. 247, no. 18, p. 2543, May 1982. doi: https://dx.doi.org/10.1001/jama.1982.03320430047030

[21] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, Jun. 2018. doi: https://dx.doi.org/10.1109/ichi.2018.00095

[22] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*. Springer International Publishing, 2019, pp. 563–574. ISBN 9783030322366

[23] M. Spytek, M. Krzyziński, S. H. Langbein, H. Baniecki, M. N. Wright, and P. Biecek, "survex: an r package for explaining machine learning survival models," 2023.

[24] M. Krzyziński, M. Spytek, H. Baniecki, and P. Biecek, "Survshap(t): Time-dependent explanations of machine learning survival models," 2022. doi: https://dx.doi.org/10.48550/ARXIV.2208.11080

[25] M. S. Kovalev, L. V. Utkin, and E. M. Kasimov, "Survlime: A method for explaining machine learning survival models," 2020.

[26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, May 2016. doi: https://dx.doi.org/10.1038/sdata.2016.35

[27] S. L. Brilleman, R. Wolfe, M. Moreno-Betancur, and M. J. Crowther, "Simulating survival data using the simsurv r package," *Journal of Statistical Software*, vol. 97, no. 3, 2021. doi: https://dx.doi.org/10.18637/jss.v097.i03

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 653–658
DOI: 10.15439/2025F6262
ISSN 2300-5963 ACSIS, Vol. 43

Coverage of OpenStreetMap data at firm level. Assessing the suitability for social science representative firm samples and found

Christian Gerhards
0000-0002-8180-3248
Federal Institute for Vocational Education and Training
Friedrich-Ebert Allee 114-116,
53113 Bonn, Germany
Email: gerhards@bibb.de

***Abstract*— Establishment surveys are essential for gaining well-founded insights into company structures and developments, whereby representativeness ensures that the results can be reliably transferred to the population and enable bias-free conclusions. Company surveys in Germany are often based on the company file of the Federal Employment Agency (BA), which, however, only allows limited links to other data sources. The free online service OpenStreetMap (OSM) could serve as an alternative, publicly accessible data source. This article analyses the extent to which OSM is suitable for drawing a representative sample of companies. It is hypothesised that OSM covers a large proportion of businesses, especially in metropolitan areas, and that customer-oriented sectors are better covered than industrial or B2B businesses. To verify this, establishments are identified in OSM and compared with the BA establishment file, differentiated by region and sector. The results show a high but not complete coverage with regional differences in favour of metropolitan areas. There are limitations in terms of data quality and contactability of the establishments.**

***Index Terms*— OpenStreetMap (OSM), Company level data, Sectoral Coverage, Survey Methodology.**

I. Introduction: New approaches in Business Research: Computational Approaches with OpenStreetMap and Supplementary Digital Data Sources

DATA on companies is of central importance for social science research, as it provides insights into economic structures, labour markets, regional development and social dynamics. An establishment refers to a social workplace. Many analyses restrict the definition to establishments with at least one employee subject to social insurance contributions. This therefore excludes solo self-employed persons. Establishments in this sense are to be distinguished from companies, which may consist of several establishments.

Information on location, size, sector and company structures makes a significant contribution to understanding employment processes, innovation, education and social inequality. Both found data from sources such as OpenStreetMap (OSM) and data from systematic company surveys offer different, complementary perspectives: While found data can provide up-to-date and widely available information, targeted surveys enable a deeper analysis of specific company characteristics and contexts. The combination of both approaches significantly expands the potential for well-founded social science analyses and evidence-based policy advice.

The availability of large, freely accessible digital data sets such as OSM opens up new opportunities for interdisciplinary research in the social sciences, economics and humanities. As a first step, the data from OSM itself can be used to analyse companies in Germany, for example. There is also the potential to link additional data from company websites, job advertisements and company valuations. Finally, the company addresses available in OSM can be used to conduct new company surveys that are as representative as possible.

This article is positioned in the context of computational approaches that utilise digital methods for theoretical, methodological and applied questions. It looks at the extent to which OSM is a suitable data source for the representative mapping of businesses in Germany. On the one hand, this applies as a source for a complete survey (use of the OSM data itself) and as an applied approach as a basis for drawing representative establishment samples. In particular, the article examines the potential of OSM for the applied generation of representative firm samples and critically reflects on the quality, reliability and representativeness of such digital resources.

The article thus addresses the central challenges of linking heterogeneous data sources such as crowdsourcing data with regard to the methodological integration of quantitative and qualitative information (e.g. OSM data with big data analyses of job advertisements or company evaluations).

It will also discuss how computer-aided methods such as web mining can contribute to systematically accessing digital resources and critically reflecting on their use for social science, economic and humanities issues.

II. State Of Research: Quality of OpenStreetMap

Found data refers to data that was not collected specifically for research purposes but was originally generated for other purposes, while the opposite is primary data collected specifically as part of scientific studies. OpenStreetMap (OSM) can be regarded as "found data" in the context of firm data collection, as the information about firms is not systematically

obtained through structured surveys, but is collected through diverse, often uncoordinated contributions from different sources.

The mapping of businesses in OpenStreetMap (OSM) is based on a variety of data sources, which together determine the diversity and depth of the information available. A significant proportion of the entries come from volunteers who use their local knowledge to record points of interest (POIs) such as shops, service providers and offices directly on site and add them to the OSM database. These personal on-site surveys are complemented by the mapping of freely available aerial and satellite imagery, utilising high-resolution imagery from services such as Bing Maps to determine the exact location and structure of businesses. In addition, mass imports of open government and company data are carried out, which are integrated into the database in compliance with the OSM import rules. Such imports can include, for example, data from local authorities or other public bodies ([1]). Last but not least, organised mapping initiatives, for example by companies or non-governmental organisations, contribute to the expansion and updating of operational data in OSM through targeted collection campaigns.

Each of these sources has specific strengths and potential biases that must be taken into account when using the data for scientific analyses. The reliability of OSM in capturing and mapping firms in Germany has been the subject of numerous scientific studies (e.g. [2], [3]). These studies show that the data quality of OSM strongly depends on the geographical location: In urban areas, the completeness and accuracy of the data is generally higher than in rural regions. For example, a study by Zielstra and Zipf ([4]) found that in German cities, the coverage of OSM is comparable to proprietary geodata, while in rural areas there may be differences.

The completeness of the businesses recorded in OSM also varies. While larger retail chains and well-known establishments are often well documented, smaller or less publicised businesses are often missing. One of the reasons for this is that the collection of such data is heavily dependent on the local community and there is no systematic survey. A study by Neis et al. ([5]) emphasises the importance of active user participation for the data quality of OSM.

Various tools and methods have been developed to assess the data quality of OSM. For example, the Heidelberg Institute for Geoinformation Technology (HeiGIT) launched the "ohsome" platform in 2018, which enables a structured analysis of OSM data quality ([6]). This platform takes into account quality characteristics such as completeness, thematic accuracy and logical consistency in accordance with ISO 19157 standards.

III. Theoretical Background And Hypotheses

OpenStreetMap (OSM) is a collaborative project in which voluntary contributors collect and record geodata and make it available to the general public. These contributors - also known as mappers - are individuals or organisations who are active with different interests: from improving local map content to scientific purposes and commercial applications. While some specifically want to map their own environment in detail, others participate for altruistic or public interest motives. At the same time, there are also regions where few or no contributors are active, which can lead to uneven data coverage.

OSM is classic Volunteered Geographic Information (VGI): Data collection depends on the voluntary commitment of the population, but also on institutional imports and technical constraints. Goodchild's concept of "Citizens as Sensors" ([7]) describes precisely this area of tension between supply (data and contribution potential) and demand (benefits for user groups).

Agency districts of the Federal Employment Agency are regional areas of responsibility of the employment agencies that divide the German labour market into administratively delimited units. They are relevant for company analyses in Germany as they provide a geographical structure that can be used to systematically record company data, evaluate it in a regionally differentiated manner and link it to labour market indicators. The theoretical mechanisms originate from VGI research (participation and digital divide theories) and geodata governance. They provide hypotheses to systematically explain why some agency districts of the Federal Employment Agency in Germany (AAB) are significantly better covered with company POI in OSM than others. A combination of individual factors (digital competence, community size) and structural contexts (urbanity, open government data, economic profile) is likely to explain most of the variance.

The quality and completeness of company data in OSM varies considerably between different AAB. These differences can be explained by socio-demographic and economic structural characteristics of the regions. The following section presents a theory-based derivation of five hypotheses that address these differences. Empirical studies show that data density correlates strongly with the number of active mappers and that social and economic differences ("digital divide") limit regional participation. On this basis, the following hypotheses can be derived for the AAB:

Hypothesis 1: The more people live in an employment agency district, the better the coverage of businesses in OpenStreetMap.

The number of OSM contributions correlates positively with the population of a region. In regions with a higher population, there are more potential contributors, which leads to better data coverage. Herfort [3] show that the completeness of OSM data is higher in densely populated areas.

Hypothesis 2: The more businesses there are in the primary sector in an agency district, the poorer the coverage.

The primary sector includes businesses involved in the extraction of raw materials, the secondary sector in processing and the tertiary sector in services. Businesses in the primary sector, such as agriculture and forestry, are often located in rural areas, which tend to be mapped in less detail in OSM. This may be due to the fact that businesses with little customer contact are less likely to be the focus of mapping initiatives.

Hypothesis 3: The more businesses there are in the secondary sector in an agency district, the poorer the coverage.

Although the secondary sector is often present in urban areas, certain industrial establishments may be less well mapped in OSM due to access restrictions or low visibility to the public.

Hypothesis 4: The higher the population density in the agency district, the better the coverage.

High population density increases the probability that a location is known to mappers. This can facilitate mapping in OSM.

Hypothesis 5: The lower the average age in the agency district, the better the coverage.

Younger population groups tend to be more tech-savvy and more willing to participate in digital platforms such as OSM. A higher level of OSM contributor activity can therefore be expected in regions with a lower average age, which leads to better data coverage.

IV. Data Basis, Operationalisation and Model

A. Data basis

Empirically, a comparison has now been made of how many companies are found per AAB in the official statistics on the one hand and in OSM on the other. The Federal Employment Agency (BA) determines the number of companies with employees subject to social security contributions on the basis of employers' social security notifications. This information forms the BA's so-called company file. The notifications are required by law and are used to compile government employment statistics. It is therefore a complete survey with a high degree of reliability.

To analyse the data from OSM, it was downloaded as a file as of 2025-05-25T20:21:00Z (https://download.geofabrik.de/europe/germany.html). The query was programmed using Python by integrating libraries for processing OSM data (osmium), tabular data (pandas) and geographical data (geopandas).

In OSM, all companies with employees subject to social security contributions were also researched as far as possible. After a previous random search, all so-called nodes from OSM in Germany are selected for which one of the following tags is filled with content: 'shop', 'amenity' (maintenance, education, …), 'office', 'craft', 'industrial'. In OSM, a "node" refers to a single point, defined by its geographical coordinates (latitude and longitude).

The companies found with geo-coordinates were then assigned to the AAB. A so-called shape file from the Federal Employment Agency was used for this purpose. The BA file "Deutschland_Agenturbezirke.shp" contains the geographical boundaries and associated information of the 150 AAB in Germany. The file can be downloaded from the BA statistics website in SHP format. (https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Regionale-Gliederungen/BA-Gebietsstruktur-Nav.html) It was then compared which companies are located in which AAB. The number of establishments per AAB was determined and stored in a file..

B. Operationalisation

- To empirically examine the question of the extent to which OSM is suitable for drawing a representative sample of businesses, the registration rate of businesses in OSM is defined as the dependent variable. On the one hand, this rate results from the nodes or points of interest (POI) registered in OSM for businesses (see above). This number is set in relation to the number of official business census of the Federal Employment Agency (BA) statistics. In OpenStreetMap (OSM), points of interest (POIs) are specific nodes to which specific attributes are assigned in order to characterise locations of particular importance.
- Various predictors are used to explain this coverage rate. These include regional structural characteristics from the BA's regional data, including population figures and economic indicators. For the analyses, characteristics of the individual AAB were taken from tables provided by the BA on its website. These include the proportion of employees in the primary, secondary and tertiary sectors, population density, the average age of the population and the number of employees per company. Table 1 provides an overview of the key figures per AAB used in the following. Most employees work in the tertiary sector. The majority of people are over 50 years old. On average (self-calculated index from the age groups), people are 44 years old. The population density varies greatly between the AAB, ranging from 47 to 7339.5 inhabitants/km2. On average, 16 people are employed per firm.

C. Model

A linear regression model (OLS) is used for the analysis in order to quantify the influence of these predictors on the recording rate (log. number of firms according to osm per aab in Germany) in order to identify systematic differences in the recording quality between the AAB. M1 includes Log. Number of employees in official BA-statistics (1), Number of employees in primary sector, Number of employees in secondary sector and Pop. Density (2). M2 additionally includes the Index of Average age per AAB. M3 adds an interaction term between (1) and (2).

V. Results

A. Descriptive results

In order to assess the suitability of OpenStreetMap (OSM) as a data source for recording businesses, a quantitative comparison of the businesses recorded in OSM with the official register data of the Federal Employment Agency (BA) is first carried out, supplemented by an analysis of regional differences and a modelling of relevant influencing factors.

TABLE 1: KEY FIGURES FOR THE INDIVIDUAL AGENCY DISTRICTS IN GERMANY
SOURCE: OWN CREATION. BASED ON: BA STATISTICS ON THE INDIVIDUAL AGENCY DISTRICTS. N = 147 AGENCY DISTRICTS (BERLIN IS MISSING)

Variable	Mean	Std. dev.	Min	Max
% Companies prim. sector	1,0	0,8	0,0	3,6
% Secondary sector Sector	29,8	8,6	9,5	51,4
% Companies tert. Sector	69,2	8,6	48,4	90,4
Persons under 25 years of age	136361	65758	44806	476924
Persons aged 25 to under 50	177142	100968	55786	716246
Persons 50 years and older	250959	99757	99536	716990
Index Average age	43,7	1,1	40,6	46,4
Population density (inhabitants/km2)	588,6	936,5	47,2	7338,5
Employees per company	15,8	2,7	10,8	26,6

TABLE 2: COMPARISON OF THE NUMBER OF REGISTERED COMPANIES (PER DISTRICT) ACCORDING TO BA AND OSM IN GERMANY. SOURCE: OWN COMPILATION: OFFICIAL BA STATISTICS (CUT-OFF DATE 31/12/2023) AND OSM QUERY (CUT-OFF DATE 25/05/2025). N = 147 AGENCY DISTRICTS.

Variable	Mean	Std. dev.	Min	Max	Total
Number of companies according to official (BA) statistics	14443,94	10049,79	5350	99550	2.137.703
Number of businesses according to OSM	5503,13	2745,02	2143	20471	824.738

In total, approximately 39% of all firms in Germany can be identified in OpenStreetMap (OSM), assuming that the matched entries refer to the same entities. (see Table 2). This shows that, as expected, less establishments are found than are recorded in the official statistics.

However, it must be noted that this comparison is conducted at the macro level. It remains unclear whether there are false positives or false negatives in the assignments of firms at the employment agency district level. For example, establishments may be recorded that have no employees subject to social security contributions.

To assess the potential of using OSM data for social science research, a more detailed examination of the identified firms is undertaken. For follow-up surveys, the availability of identifying information – such as postal address, email address, or telephone number – is particularly important. In addition, company websites play a key role in enabling further data enrichment. The following table 3 provides an overview of the proportion of firms for which these attributes could be retrieved.

TABLE 3: % OF FIRMS IN OSM WITH. N = 824.738 FIRMS TOTAL = NUMBER OF REGISTERED COMPANIES IN GERMANY. SOURCE: OWN COMPILATION: OSM QUERY (CUT-OFF DATE 25/05/2025).

	% firms with…
Name	95
Address: city	56
Address: postcode	57
Address: street	59
Address: house number	59
Phone	27
Email	13
Fax	4
Website	29

The analysis of business entries in OpenStreetMap (OSM) reveals varying levels of detail across different attributes.

Nearly all firms (95%) are identified by name, indicating a high degree of basic recognition in the dataset. Address-related information is moderately well-represented, with 56–59% of firms including details such as city, postcode, street, and house number. In contrast, contact information is less frequently available: only 27% of firms list a phone number, 13% provide an email address, and a mere 4% include a fax number. Websites are linked in 29% of cases. These findings suggest that while core identification and location data are commonly present in OSM business entries, contact details are comparatively sparse.

In the next step, the number of firms recorded in OSM will be compared, at the level of employment agency districts, with the number of firms that should be present according to official statistics. If one looks at the correlation between the values from the official statistics and the data from OSM, you can see a clear linear correlation across the AAB (see Figure 1). The dispersion is moderate, which indicates a good mapping.

A graphical analysis shows that regions in west Germany, central Germany and north-east Germany are more likely to ikely have a higher coverage rate of firms in OSM. No data is available for Berlin due to a different data structure (see Figure 2). Some of these regions are particularly rural.

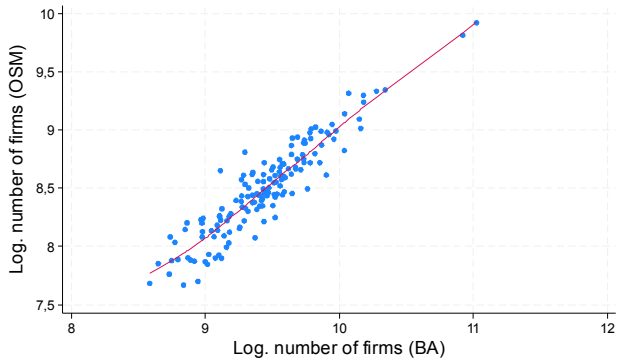


Fig. 1: Log. Number of firms per AAB according to OSM and BA in Germany, Source: Own compilation. Variables logarithmised due to skewness. n=147 agency districts.

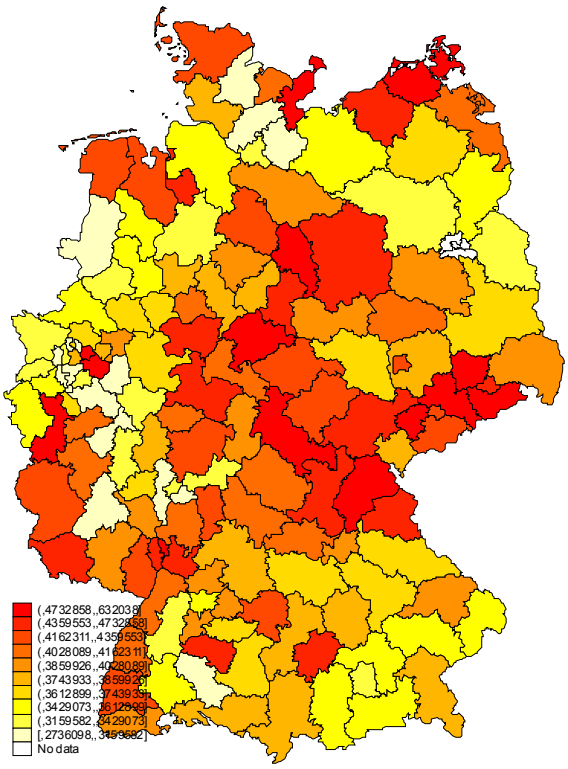


Fig. 2: higher (red) and lower coverage rate of firms (white) in OSM compared to BA in the agency districts in Germany.
Source: own compilation. The totals of the figures per agency district are compared. Berlin: no data.

B. Multivariate results

A regression model is calculated for further analysis. The dependent variable for each AAB is the logarithmised number of firms according to the OSM. The number of firms according to the BA is controlled for (Table 4).

Overall, the results show that both demographic and economic-structural characteristics have no significant influence on the completeness of voluntarily created geodata. The analysis shows no regional differences in the recording of businesses in OSM. Firstly (H1), a higher population size in the AAB does not improve the capture rate. Secondly (H2), a high proportion of firms in the primary sector, such as agriculture and forestry, does not correlate with the capture quality. Thirdly (H3), there is no correlation for the secondary sector. Fourthly (H4), a higher population density does not indicate a poorer coverage rate. Finally, there is no significant correlation between the average age of the population in the AAB and a better coverage rate.

VI. CONCLUSION

This study shows that OpenStreetMap (OSM) offers moderate coverage of businesses in Germany (about 39%). The total number of establishments recorded in OSM is not comparable to the official establishment file of the Federal Employment Agency (BA), although not necessarily the same establishments are recorded.

It was all the more important to examine whether the coverage is randomly distributed or follows a systematic pattern. The analysis shows that there are no significant regional and sectoral differences: Larger populations and lower average ages seems not to be associated with better coverage quality, while establishments in the primary and secondary sectors and regions with high population density seems not to tend to be better represented. These results illustrate both the potential and limitations of OSM for firm-related research.

Despite the limitations observed, OSM offers a fundamental advantage over traditional official registers: Because the data from the Federal Employment Agency's establishment file (BA establishment file) is based on mandatory social security data, it is generally subject to the requirements of the GDPR. Data on companies from OSM generally has no personal reference and can therefore be linked more easily with other digital resources. In addition, OSM opens up new opportunities for analysing "found data", for example by linking it to company websites, job advertisements or company valuations. These approaches enable deeper insights into company structures and dynamics that are difficult to capture with traditional surveys or register data.

For future research, an in-depth analysis of the validity and accessibility of the businesses recorded in OSM is recommended, for example by enriching the datasets with business email addresses, business websites, job adverts or telephone numbers - if these are missing. Companies could then be contacted for company surveys using contact information. Enrichment with qualitative information and text data can provide a new in-depth insight into the behaviour of companies. Improved procedures could help to systematically identify false-positive entries or the allocation of establishments as such and further increase data quality. In particular, approaches for the automated linking of OSM data with other open sources - such as knowledge graphs or web mining/web scraping of company websites - should be developed. At the same time, critical reflection is essential, particularly with regard to data protection, anonymisation and potential distortions, in order to ensure the validity and

TABLE 4: REGRESSION MODEL: LOG. NUMBER OF FIRMS ACCORDING TO OSM PER AAB IN GERMANY. SOURCE: OWN COMPILATION. N=147. MISSING AAB (MAX. 150): BERLIN (OTHER DATA COLLECTION) MODEL M2 ADDITIONALLY WITH AGE INDEX, MODEL M3 ADDITIONALLY WITH INTERACTION TERM.

	M1	M2	M3
Log. Number of employees in official BA-statistics (1)	0,931***	0,945***	1,031***
% of employees in primary sector	-0,029	-0,035	-0,034
% of employees in secondary sector	-0,003	-0,003	-0,003
Pop. Density (2)	-0,04	-0,03	0,005
Index average age		0,019	0,019
(1) x (2)			-0,014
Constant	6,469***	5,530***	5,336***
n	147	147	147
Adj. R^2	0,861	0,861	0,861

reliability of the results. In view of the increasing importance of flexible and up-to-date data sources in the social sciences, economics and humanities, OSM is therefore a promising tool that should be further utilised and developed in interdisciplinary research approaches.

VII. REFERENCES

[1] D. Zielstra, H. H. Hochmair and P. Neis, "Assessing the effect of data imports on the completeness of OpenStreetMap – A United States case study", Transactions in GIS, vol. 17, no. 3, pp. 315–334, 2013.

[2] J. E. Vargas-Munoz, S. Srivastava, D. Tuia and A. X. Falcão, "OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing", IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 1, pp. 184–199, March 2021, doi: 10.1109/MGRS.2020.2994107.

[3] B. Herfort, S. Lautenbach, J. P. de Albuquerque and A. Zipf, "A spatio-temporal analysis investigating completeness and inequalities of OpenStreetMap building data", Nature Communications, vol. 14, 39698, 2023, doi: 10.1038/s41467-023-39698-6.

[4] D. Zielstra and A. Zipf, "A comparative study of proprietary geodata and volunteered geographic information for Germany", in 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 2010, pp. 1–15.

[5] P. Neis, D. Zielstra, A. Zipf and A. Struck, "Empirical analyses of the data quality of OpenStreetMap – Experiences from two years of operating several OSM online services", in Applied Geoinformatics 2010, 2010.

[6] A. Klonner et al., "'Ohsome' OpenStreetMap Data Evaluation: Fitness of Field Papers for Participatory Mapping", in Proceedings of the Academic Track at the State of the Map 2019, M. Minghini, A. Y. Grinberger, L. Juhász, G. Yeboah and P. Mooney, Eds., pp. 35–36, Heidelberg, Germany, September 21–23, 2019. [Online]. Available: https://zenodo.org/communities/sotm-2019, doi: 10.5281/zenodo.3387725.

[7] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography", GeoJournal, vol. 69, pp. 211–221, 2007.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 659–664
DOI: 10.15439/2025F1886
ISSN 2300-5963 ACSIS, Vol. 43

A Multithreaded Java-Based Video Encoder for Multicore Systems

Maciej Okoń
0009-0005-7427-6784
Maria Curie-Skłodowska University (UMCS)
ul. Pl. Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland

Beata Bylina
0000-0002-1327-9747
Maria Curie-Skłodowska University (UMCS)
ul. Pl. Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland
Email: beata.bylina@mail.umcs.pl

***Abstract*—The growing demand for efficient video compression solutions underscores the increasing significance of research into coding optimization on multi-core systems.**

This paper presents the implementation and performance analysis of a multithreaded video encoder developed in Java and optimized for modern multi-core processors. The encoder performs intra-frame compression of raw YUV 4:2:0 video using a frame-based parallelism approach. This method maximizes CPU core utilization while minimizing thread management overhead.

Tests were conducted on five platforms equipped with multi-core processors from Intel and AMD. The application's execution time was measured for varying numbers of frames. The obtained results demonstrate effective scalability of the encoder as the number of processed frames increases, confirming that Java can be an effective tool for implementing parallel video compression on multi-core systems.

I. Introduction

THE INCREASING demand for efficient video processing continues to test the limits of contemporary computing platforms. Video compression is a computationally intensive task, particularly for high-resolution, high frame rate content, and achieving real-time or near-real-time performance is often essential in applications such as live streaming, video surveillance, or large-scale media processing pipelines. While dedicated hardware accelerators like GPUs are often used for such workloads [1], many real-world applications, especially in server environments, rely on general-purpose CPUs [2]. Numerous studies have explored video encoding algorithms and performance optimization techniques, providing valuable insights into how encoding efficiency can be improved through software and hardware-level enhancements [1], [3], [4].

This paper presents a performance-oriented evaluation of multithreaded video compression on modern server-grade processors using a custom-built encoder implemented in Java. Java was selected for its platform independence and robust support for multithreading, which facilitated the implementation of parallel processing using native concurrency constructs and stream-based data pipelines [5]. The encoder performs intra-frame compression on raw YUV 4:2:0 video and supports both single-threaded and multithreaded execution. Frame-based parallelism was used instead of block-based parallelism to more effectively leverage the high thread counts available on the test machines. This approach ensures sufficient workload distribution across all cores, avoiding granularity limitations that arise when the number of parallel tasks is too low. It also reduces overhead associated with thread management, allowing the encoder to scale efficiently with increasing core counts. Our evaluation measures the total encoding time, starting after the video data is fully loaded into memory and ending once the encoded file has been written to disk. As such, the performance results reflect not only the efficiency of parallel computation, but also task distribution overhead, thread synchronization, and serial operations such as output handling.

Five multicore platforms were tested, representing a range of CPU architectures and core counts. By comparing execution time and speedup across varying frame counts, we explore the strengths and limitations of each system in handling a realistic, mixed-parallelism workload. Rather than isolating idealized, parallel-only scenarios, our measurements reflect the practical end-to-end behavior of CPU-based encoding workflows.

This article is organized as follows. Section II begins with an introduction to video encoding and compression, describing the basic concepts and challenges of reducing video data while maintaining visual quality. It explains how video encoders convert raw video images into a compressed format through algorithmic steps. It also highlights the computational requirements of these processes and the importance of parallelization for performance. Section III presents the main video compression algorithms such as discrete cosine transform (DCT), quantization and motion estimation. It explains how these techniques reduce redundancy and achieve efficient compression. Section IV describes the architecture and internal structure of the encoder, detailing the modular design, block-based processing and the implementation of parallelization at the frame level using multithreading in Java. Section V describes the test environment and methodology, followed by an analysis of the performance results on different systems, comparing single-threaded and multithreaded execution. Section VI presents the conclusions from the experiments and possible directions for further research.

II. Video encoding and compression

Video encoding is a computationally demanding process that transforms raw, high-resolution video data into a compressed format suitable for storage and transmission [6]. This transformation typically involves a series of algorithmic

steps designed to reduce redundancy and preserve perceptual quality.

At the core of most intra-frame compression schemes lies a block-based pipeline consisting of the following operations [7]:

1) **Block partitioning:** the input frame is divided into fixed-size macroblocks (e.g., 16×16 pixels).
2) **Transformation:** each macroblock is converted into the frequency domain using the Discrete Cosine Transform (DCT), enabling more efficient compression by concentrating signal energy into fewer coefficients.
3) **Quantization:** DCT coefficients are scaled and rounded to reduce precision, effectively discarding less perceptually significant data.
4) **Entropy coding:** quantized coefficients are compressed using techniques such as run-length encoding (RLE) to eliminate statistical redundancy.

These stages are computationally expensive, particularly at high resolutions and frame rates. As a result, the performance of a video encoder depends not only on the efficiency of these individual algorithms, but also on how well they are integrated into a scalable execution model. Our approach focuses on frame-based parallelization, where individual frames are processed concurrently across multiple threads.

III. Transform coding of video data and reduction algorithms

TRANSFORM coding is one of the more computationally demanding steps in the encoding pipeline. This process converts visual data from the spatial domain — where it's represented as individual pixel values — into the frequency domain. It is used to isolate the components that carry the most visual weight from those that can be discarded with minimal impact on perceived quality [7]. After the transformation, we reduce the precision of the less important frequency components using quantization. This lossy compression step shrinks the file size drastically, without noticeably affecting how the video looks [8]. The specific technique we rely on for this transformation is the Discrete Cosine Transform (DCT).

A. Discrete Cosine Transform (DCT)

DCT represents a block of an image as a weighted sum of cosine waves oscillating at different frequencies. This shift from pixels to frequency components reveals a useful pattern: most of the important visual information tends to concentrate in just a few low-frequency coefficients, while the high-frequency components — often corresponding to fine detail or noise — can be compressed more aggressively or discarded entirely [9].

In a typical implementation, the process begins by dividing each video frame into smaller blocks, usually 8×8 pixels. Each of these blocks is then transformed using the 2D DCT. The result is a matrix of coefficients that describe the contribution of each frequency component within that block — the top-left corner of the matrix holds the DC (average) value, while the remaining entries describe increasingly fine detail as you move outward. This structure allows us to perform quantization more intelligently. Because human vision is more sensitive to low-frequency information, we can preserve those coefficients with higher precision and apply more aggressive compression to the less perceptible high frequencies.

Mathematically, the 2D DCT (type II, $O(N^4)$) for a block of size $N \times N$ is defined as follows:

$$Y(u,v) = \frac{1}{4} \cdot c(u) \cdot c(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cdot \cos\left(\frac{(2x+1)\cdot u \cdot \pi}{2N}\right) \cdot \cos\left(\frac{(2y+1)\cdot v \cdot \pi}{2N}\right) \tag{1}$$

where:

- $Y(u,v)$ is the resulting DCT coefficient at frequency indices (u,v),
- $f(x,y)$ is the input pixel value at position (x,y),
- $c(k) = \frac{1}{\sqrt{2}}$ when $k = 0$, and $c(k) = 1$ otherwise,
- N is the block size, typically 8.

One practical advantage of DCT is its separability. The 2D transform can be broken down into a 1D DCT applied first along the rows, and then along the columns (or vice versa). This significantly reduces the computational complexity ($O(N^3)$) [10].

B. DCT quantization

Once each image block has been transformed via DCT, we're left with a set of frequency-domain coefficients. While many of these coefficients — especially the ones representing higher frequencies — contribute little to the perceived visual quality, they still occupy space. Quantization reduces this burden by simplifying these values.

The process involves dividing each coefficient by a corresponding value from a quantization matrix and then rounding the result:

$$Q(x) = \text{round}\left(\frac{x}{Q_{i,j} \cdot S}\right) \tag{2}$$

Where:

- $Q(x)$ is the quantized coefficient,
- x is the original DCT coefficient,
- $Q_{i,j}$ is the value from the quantization matrix at position (i,j),
- S is a scaling factor that adjusts the compression strength.

The quantization matrix controls how much detail is kept in different parts of the image. Lower numbers in the matrix correspond to low-frequency components, which capture the general structure and smooth areas of the image—these are preserved more accurately. Higher numbers are assigned to high-frequency components, which represent fine details or noise; these are compressed more aggressively.

A commonly used quantization matrix from the JPEG standard is shown below.

$$\begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

C. Motion estimation and compensation

To further improve compression efficiency, modern video encoders utilize the similarity between consecutive frames through motion estimation and compensation - processes that analyze changes between frames and predict what future frames will look like based on the movement of visual elements.

Step 1: Motion estimation. This step works by comparing two consecutive frames to identify macroblocks that have shifted in position. This is usually done by searching within a defined area of the reference frame to find the best match for each block in the current frame. Metrics such as the Sum of Absolute Differences (SAD) or Mean Squared Error (MSE) are used to evaluate how well blocks match. The result of this process is a set of motion vectors, which describe how far and in which direction each block has moved.

Step 2: Motion compensation. This step then uses these vectors to shift corresponding blocks in the reference frame to approximate the content of the current frame. Instead of encoding the full frame, the encoder stores only the difference (residual) between the predicted and the actual frame. This residual typically contains less information than the original frame.

IV. Encoder architecture

The encoder was developed in Java with a modular, object-oriented structure, centered around the core classes `VideoObjectPlane`, `Macroblock`, `Bitstream`, and `VOPHeader`. Each video frame is divided into 16x16 pixel macroblocks, which are then processed using DCT, quantization, and run-length encoding (RLE).

The encoder supports two execution modes: sequential and multithreaded. In the multithreaded version, the `ExecutorService` from `java.util.concurrent` package is used to create a thread pool that matches the number of available CPU cores. This allows each video frame to be processed independently in a separate thread, significantly improving encoding performance. Synchronization is handled through `Future` objects, which act as placeholders for the results of asynchronous tasks and allow results to be collected in order. .

Two levels of parallelism were considered during implementation: frame-level and block-level. Frame-level parallelism was chosen and implemented, with each frame processed in a separate thread. This approach aligned well with the hardware characteristics of the testing environment. While block-level parallelism was explored as a finer-grained alternative, it proved to be significantly slower, and thus not integrated into the final encoder.

The encoder processes raw video data stored in a YUV 4:2:0 file with a resolution of 1920x1080 pixels. This format represents color using separate planes for luminance (Y) and chrominance (U and V), with chrominance subsampled by a factor of two in both horizontal and vertical dimensions. Video frames are loaded using the `loadYUVFrames` function, which reads the raw data from disk, partitions it into individual frames, and converts them to `BufferedImage` objects in RGB color space. Each frame is then encoded as an `I_VOP` object, that then segments the image into individual macroblocks and compresses each block.

Each `Macroblock` object stores pixel data in the same 4:2:0 structure, with luminance represented by 256 samples and each chrominance channel by 64 samples, corresponding to the 16×16 luminance block and its associated 8×8 chrominance blocks. These values are stored in separate one-dimensional arrays for Y, U, and V components and are extracted during macroblock construction from the input RGB image via internal color space conversion. Internally, the pixel data is converted from RGB to YUV, then further processed into 8x8 sub-blocks for DCT and quantization. The DCT implementation uses separable 2D transformation with precomputed cosine values for efficiency and relies on double-precision arithmetic to maintain numerical accuracy during computation.

The result of the encoding process is stored in a binary file, written using the custom `Bitstream` class. The output maintains the original block order thanks to the use of `Future.get()`. This ensures deterministic output even when block processing completes in a different order.

The entire encoder was developed using standard Java libraries, including `java.util.concurrent` for parallelism and `java.awt` for image handling. Due to its modular architecture, the system can be extended in the future with additional features, such as motion estimation for inter-frame coding or dynamic encoding strategies based on scene characteristics like `Video Objects`.

V. Performance analysis and test environment

To evaluate the performance of the implemented video encoder, a series of experiments were conducted on five different computing platforms, each equipped with a distinct server-grade processor, selected to cover a range of architectures, core counts, and generations:

- **C1 (2 x Intel Xeon E5-2670 v3)** – 12 cores / 24 threads, Haswell-EP architecture (22 nm), released in 2014. An older platform with limited performance compared to modern standards.
- **C2 (2 x Intel Xeon Gold 5218R)** – 20 cores / 40 threads, Cascade Lake Refresh (14 nm), released in 2020. A mid-range server CPU offering a good balance between core count and clock speed.

- **C3 (2 x Intel Xeon Gold 6342)** – 24 cores / 48 threads, Ice Lake-SP architecture (10 nm), released in 2021. With improved single-core performance and memory bandwidth support.
- **C4 (2 x Intel Xeon Platinum 8358)** – 32 cores / 64 threads, Ice Lake-SP architecture (10 nm), released in 2021. High core density suited for parallel workloads.
- **C5 (2 x AMD EPYC 9654)** – 96 cores / 192 threads, Zen 4 (Genoa) architecture (5 nm), released in 2022. A high-end processor optimized for massive parallelism and throughput.

Each system processed a standard YUV 4:2:0 1920x1080 video file, with frame counts ranging from 250 to 2000 for the multithreaded (MT) version[1] and from 250 to 1500 for the single-threaded (ST) version[2].Depending on the number of frames, the size of the test files ranged from approximately 741 MB (250 frames) to 5.79 GB (2000 frames). This setup allowed for the analysis of both raw execution time and scalability with increasing frame counts.

To quantify the performance gain achieved through parallelization, the speedup factor was calculated as the ratio of the single-threaded execution time to the multithreaded execution time:

$$S = \frac{T_{\mathrm{ST}}}{T_{\mathrm{MT}}}. \tag{3}$$

The performance data collected across all tested platforms is summarized in Table I, which lists the minimum measured processing times for both single-threaded and multithreaded executions, along with the computed speedup values. The results are presented for increasing frame counts, illustrating how processing time scales with input size. Speedup values were calculated using equation 3. Figures 1 and 2 visualize the speedup and multithreaded processing times across all systems, respectively, providing a comparative view of scalability and performance gains with increased threading and workload.

In general, all platforms demonstrate substantial performance improvements from multithreading, with speedup tending to increase as the number of processed frames grows. However, the magnitude and consistency of these gains vary across systems. Notably, some platforms—particularly `C2` `C3`, and `C4` —exhibit a plateau or even slight dip in speedup around mid-range workloads (750 to 1250 frames), indicating that the platforms either have, or are close to reaching their maximum effective scaling and additional workload does not significantly improve parallel efficiency.

Among the tested systems, the AMD-based `C5` platform initially shows comparatively modest speedup at lower frame counts, which can be attributed to the overhead of managing its high core count when the workload is too small to fully utilize the available resources. As the workload increases, `C5`'s scalability becomes clear: by the 1500-frame mark, it surpasses all other platforms in speedup and achieves the shortest absolute multithreaded processing time at 2000 frames.

Interestingly, `C1`, despite having the poorest absolute processing times, consistently achieves the second highest speedup among all platforms. This indicates that while its baseline single-threaded performance is limited, it benefits strongly from multithreading, scaling efficiently with increasing workload size. In contrast, `C2`, which also features fewer cores, shows more consistent but lower speedup values, plateauing around 12 to 13× speedup across moderate to larger frame counts, and is eventually outperformed by more scalable systems at higher workloads. `C3` and `C4` occupy a middle ground, maintaining solid and relatively stable speedup improvements that exceed those of `C2`, yet do not quite match the top-end performance and scalability of `C5` at larger frame counts.

VI. Conclusion

This work aimed to design, implement, and evaluate a multithreaded video encoder in Java, capable of leveraging the parallel processing capabilities of modern multicore architectures. In pursuit of this goal, the encoder was developed using a modular, object-oriented structure in Java, incorporating key compression techniques such as the Discrete Cosine Transform (DCT), quantization, and run-length encoding (RLE). The implementation included both single-threaded and multithreaded execution modes, enabling a comparative analysis of encoding performance across different hardware platforms and workloads. A series of experiments were conducted using five server-grade systems with varying core counts and architectures, focusing on the encoding of high-definition video data in YUV 4:2:0 format. The objective was to assess the scalability and performance benefits of multithreading for video encoding tasks.

The results demonstrate that all tested CPUs benefit substantially from multithreading, with speedups ranging from approximately 9x to over 24x depending on the system and workload. Among them, `C5` achieved the highest ultimate acceleration, peaking above 24× speedup at larger frame counts, demonstrating superior scalability on high concurrency workloads despite initially modest gains at lower frame counts. Notably, `C1`, despite having the slowest absolute processing times, consistently achieved the second highest speedup, indicating efficient scaling relative to its baseline single-threaded performance. `C3` and `C4` maintained solid and relatively stable speedup improvements, exceeding `C2`'s generally lower but consistent gains, which plateaued around 12 to 13× at moderate to large workloads.

Importantly, our measurements suggest that raw core count alone does not fully explain observed performance. Differences in scaling behavior—especially how speedup grows with frame count—highlight the role of thread orchestration overhead and single-threaded efficiency in managing output and coordination. For example, while `C5` initially showed lower speedup at small workloads, it ultimately surpassed

[1]A total of 2000 frames were used for testing. This value was chosen as the maximum that could be reliably loaded across all systems, given their varying available memory.

[2]Due to long processing times, the single-threaded encoder was only tested up to 1500 frames.

TABLE I
MINIMUM ST AND MT PROCESSING TIMES AND COMPUTED SPEEDUP ACROSS ALL MACHINES

Frame Count	C1			C2			C3			C4			C5		
	T_{ST} [s]	T_{MT} [s]	Speedup	T_{ST} [s]	T_{MT} [s]	Speedup	T_{ST} [s]	T_{MT} [s]	Speedup	T_{ST} [s]	T_{MT} [s]	Speedup	T_{ST} [s]	T_{MT} [s]	Speedup
250	63.702	5.029	12.67	44.123	4.302	10.26	41.383	3.446	12.01	42.752	4.044	10.57	63.874	7.043	9.07
500	141.167	8.460	16.69	87.956	7.202	12.21	83.395	6.052	13.78	85.934	6.769	12.70	131.713	10.498	12.55
750	202.898	11.845	17.13	133.306	10.776	12.37	126.079	8.991	14.02	129.259	9.562	13.52	192.968	11.030	17.49
1000	272.535	15.743	17.31	178.920	14.486	12.35	169.908	12.423	13.68	174.848	13.445	13.00	276.740	14.447	19.16
1250	352.270	18.463	19.08	225.432	18.109	12.45	213.186	15.155	14.07	219.300	16.112	13.61	357.784	18.686	19.15
1500	433.261	22.316	19.41	268.962	21.007	12.80	254.934	17.570	14.51	263.189	18.791	14.01	405.174	16.592	24.42
2000	–	29.113	–	–	29.340	–	–	23.829	–	–	25.065	–	–	23.697	–

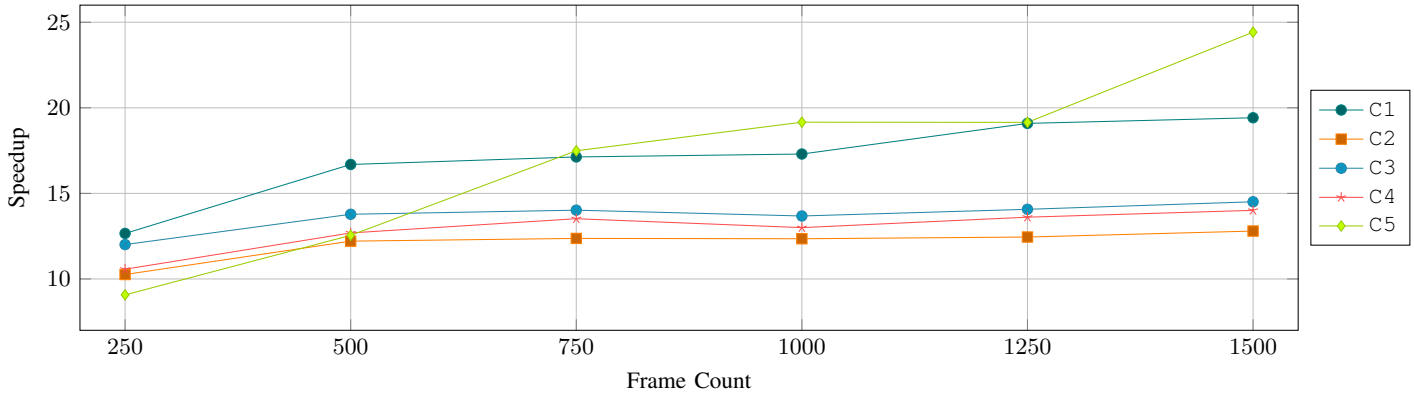


Fig. 1. Speedup of multithreaded processing across all machines

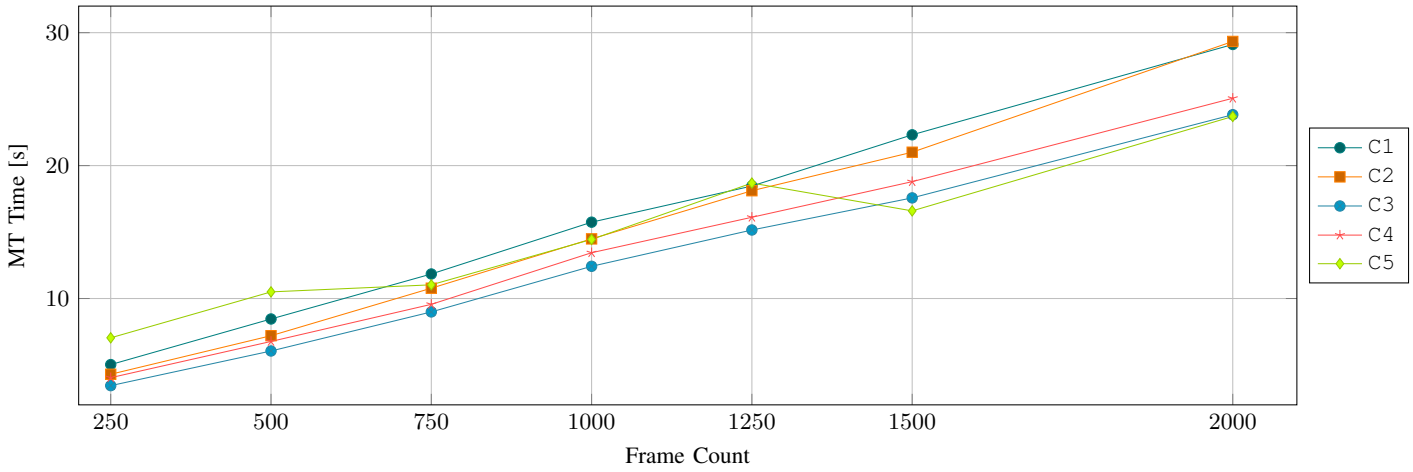


Fig. 2. Comparison of minimum multithreaded processing times for all tested systems

several peers as the workload increased, indicating superior multithreaded scalability at higher concurrency levels.

These findings validate the need to assess full-system performance for parallel workloads, not just isolated parallel execution paths. In practice, the total encoding time is shaped by how well a CPU balances parallel computation with effective scheduling and sequential operations. This is especially relevant in environments without GPU acceleration, where CPUs must take on end-to-end responsibility for video processing.

Future research may explore extending the encoding model to include inter-frame prediction, introducing more complex dependencies across frames, or benchmarking hybrid CPU-GPU configurations. Nevertheless, this work offers a grounded perspective on CPU capabilities in a realistic, high-throughput encoding pipeline, and serves as a foundation for further exploration of CPU-bound multimedia processing.

REFERENCES

[1] S. Katsigiannis, V. Dimitsas, and D. Maroulis, "A GPU vs CPU performance evaluation of an experimental video compression algorithm," in *Proceedings of the 7th International Workshop on Quality of Multimedia Experience (QoMEX)*, Pylos, Greece, May 2015, pp. 1–6. DOI: 10.1109/QoMEX.2015.7148134.

[2] S. Sankaraiah, L. H. Shuan, C. Eswaran, and J. Abdullah, "Performance Optimization of Video Coding Process on Multi-Core Platform Using GOP Level Parallelism," *International Journal of Parallel Programming*, vol. 42, no. 6, pp. 931–947, Dec. 2014. DOI: 10.1007/s10766-013-0267-4.

[3] Y. Zhang, C. Zhang, R. Fan, S. Ma, Z. Chen, and C.-C. J. Kuo, "Recent Advances on HEVC Inter-frame Coding: From Optimization to Implementation and Beyond," *arXiv preprint arXiv:1910.09770*, Oct. 2019. Available: https://arxiv.org/abs/1910.09770.

[4] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep Learning-Based Video Coding: A Review and A Case Study," *arXiv preprint arXiv:1904.12462*, Apr. 2019. Available: https://arxiv.org/abs/1904.12462.

[5] Oracle, "Concurrency Utilities," *Java Platform, Standard Edition 8 API Specification*, 2014. [Online]. Available: https://docs.oracle.com/javase/8/docs/api/java/util/concurrent/package-summary.html.

[6] A. J. Kane, "Encoding vs. Decoding," AVIXA, [Online]. Available: https://www.avixa.org/pro-av-trends/articles/encoding-vs-decoding. [Accessed: Dec. 2024].

[7] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia*. Chichester, U.K.: Wiley, 2003.

[8] D. Le Gall, *MPEG: A Video Compression Standard for Multimedia Applications*. IEEE, 1991.

[9] I.-M. Pao and M.-T. Sun, "Modeling DCT coefficients for fast video encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 608–616, 1999, doi: 10.1109/76.767126.

[10] S. A. Khayam, "The Discrete Cosine Transform (DCT): Theory and Application," *Course Notes, Department of Electrical & Computer Engineering*, 2003.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 665–673
DOI: 10.15439/2025F2608
ISSN 2300-5963 ACSIS, Vol. 43

reVISION: A Polish Benchmark for Evaluating Vision-Language Models on Multimodal National Exam Data

Michał Ciesiółka
0009-0009-8283-1118
Adam Mickiewicz University
Center for Artificial Intelligence AMU
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
Email: michal.ciesiolka@proton.me

Filip Graliński
0000-0001-8066-4533
Adam Mickiewicz University
Center for Artificial Intelligence AMU
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
Email: filipg@amu.edu.pl

***Abstract*—Vision-Language (VL) models have gained significant popularity in recent years for tasks involving data extraction and image recognition. In this paper, we introduce reVISION, a large-scale Polish benchmark for evaluating such models, comprising over 39k questions extracted from Polish national exams. We assess both the models' general knowledge and their ability to handle tasks that go beyond plain text, including tables and the recognition of specialized visual objects. Our study also examines the effects of image resizing and the inclusion of additional context via OCR-extracted text. To explore the correlation between human and model performance, we compare their results across multiple exam years.**

I. Introduction

FROM the very beginning, Large Language Models have been evaluated using exam-like tests based on multiple-choice questions, as exemplified by the MMLU benchmark created specifically for assessing capabilities of GPT-3[7]. MMLU was popular among LLM practitioners despite its inconsistencies and even outright errors [6]. A more principal way to build LLM benchmarks was proposed by [8], where the benchmark was created using Polish national exams.

Currently, many LLMs are actually not only about language, but they are multi-modal models, handling both language and vision. Again, a natural need to evaluate them in a systematic manner and an idea to use real-life exams remains valid, as actually a significant part of exam questions uses 2D layout and images (tables, graphs, diagrams, etc.). In this paper, we extend the ideas put forward by authors [8] to the bimodal (text and vision) domain, by preparing reVISION, a new comprehensive benchmark for measuring the quality of VL models.

Figure 1 shows an example of a difficult physics question extracted from the 2014 high school exam, which both Qwen2.5-VL-72B-Instruct and GPT-4.1 failed to answer correctly. In contrast, Figure 2 presents an easier question from the 2020 professional exam, which was correctly answered by both models.

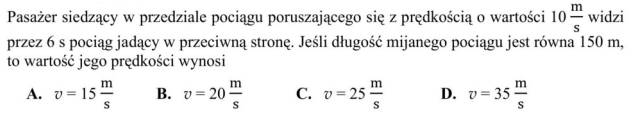


Fig. 1: Example of a difficult question from the 2013 high school physics exam. The question states: A passenger sitting in a train compartment moving at a speed of 10m/s sees, for 6 seconds, a train traveling in the opposite direction. If the length of the passing train is 150m, what is its speed?

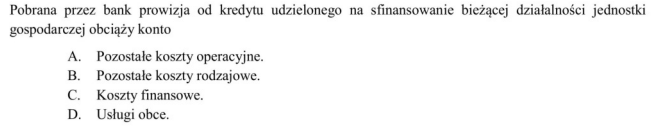


Fig. 2: Example of an easy question from the 2020 professional exam. The question states: The commission charged by the bank on a loan granted to finance the current operations of the business entity is recorded under the account: A. Other operating costs. B. Other cost types. C. Financial costs. D. External services.

Our contributions consist in stating and answering the following questions:

- How effectively do vision–language (VL) models answer Polish exam questions?
- How do VL models compare to human performance?
- How do visually grounded questions differ from text-based questions?
- Does resizing images to reduce computational cost affect model performance?
- Does providing additional context in the form of OCR-extracted text improve model accuracy?

II. Related work

Ever since the introduction of the first vision-language (VL) models [12], their performance has been continually challenged using a variety of benchmarks. Assessing their capabilities is

crucial for selecting the appropriate model for the task at hand. These evaluations range from testing object, attribute, and relation recognition [16] to performance on multi-disciplinary tasks [14]. Despite the diversity of tasks, all benchmarks share a common goal: to identify and compare models in order to determine which performs best in a given category.

LLMzSzŁ [8], a previous work that relied heavily on Polish exam questions, addressed the need for Polish-language benchmarks in the research landscape. It demonstrated the importance of academic datasets and benchmarks that capture the linguistic specificity of languages like Polish. By incorporating visual information, we aim to support the development and evaluation of multimodal models in the Polish-language context, extending the capabilities of previous text-only benchmarks.

Previous work such as EXAMS-V [4] emphasizes the need for benchmarks that assess curriculum-specific knowledge. The EXAMS-V dataset consists of thousands of multilingual exam questions. The authors highlight the dataset's difficulty, noting that even the most advanced models, such as GPT-4V and Gemini, struggle with it. While Polish professional exams are included in the benchmark, they account for just over 2.5k questions. In this paper, we focus on expanding the number and variety of questions while narrowing the language scope to a single language, while using clear criteria for inclusion and future extensions. Similarly, M3Exam [15] provides a multilevel benchmark composed of human exam questions. The study demonstrates that models face challenges when processing multilingual texts, particularly in low-resource languages.

III. Dataset

A. Data sources

Each year in Poland, three types of national exams are conducted to assess the knowledge of Polish students: the eighth-grade exam, the high school exam, and professional exams. The middle school exam, which was part of the Polish education system prior to recent educational reforms, was also considered when preparing the data sources. After each exam session, the exams are published by the Polish Central Examination Board (Pol. CKE, Centralna Komisja Egzaminacyjna). These archives provided us with access to both questions and answers from a reliable source, ensuring the high quality of our dataset.

B. Data selection

The primary constraint considered when selecting documents for dataset preparation was the number of closed-ended questions with a single correct answer. This approach eliminated the need for a larger model or human annotators as judges, enabling us to focus solely on extracting one type of question while ensuring objectivity by removing emotional bias and personal preferences. Since many types of exams were heavily focused on open-ended answers, we decided to exclude them from further processing. This strict elimination process resulted in the following categories: math, natural sciences, biology, physics, and the Polish language for pre-high school questions, and arts, mechanics (including mining and metallurgy), and agriculture (including forestry) for professional exams.

C. Dataset preparation

The preparation of the dataset began with downloading the exams that were made public over the years. We automated this stage using basic web scraping methods and Python scripts. All exams were published in PDF format, with their answers in separate files, making it crucial to track which question file corresponds to which answer file.

To extract all the questions from a single PDF file, we first had to determine the accurate X-axis positions of each question. We took a single page of the file in order to test the techniques for position extraction. Our initial approach was to prompt a multimodal model with an image and have it extract the bounding boxes of entire questions. However, this proved ineffective as multimodal models, although good for OCR text extraction, cannot accurately pinpoint and return text positions even on small images. Having this in mind, we decided to use an OCR engine, specifically PaddleOCR[1], that can extract text, precisely draw its bounding boxes, and return their positions. In this way, we identified the text indicating the beginning of a question. For most exams, the keyword was 'Task' (Pol. *Zadanie*). If the word appeared above the actual task, we used the lower X-axis of 'Zadanie' as the starting position of the task; otherwise, we used the upper X-axis. The upper X-axis of the next question was used as the ending position. The images were then cropped along those axes.

When initially testing on entire exam documents, we attempted to convert the entire PDF file into a single long image and then used an OCR engine to detect text. However, this approach proved ineffective, as the engine was unable to detect any text in the image. The issue arises primarily due to the extreme aspect ratio of the images. OCR engines are typically optimized for standard-sized documents and conventional aspect ratios. A more effective approach was based on the fact that each question had to be contained on a single page, making it possible to split the document into individual pages, convert them into images, and then detect tasks within them. After fine-tuning the detection and splitting process to match each exam type, we were able to extract almost 40,000 images, each containing exactly one question. This figure reflects the results after a cleaning process, which involved removing approximately 30% of the original questions, primarily due to duplication, with a smaller portion removed for issues such as poor cropping. The cleaning process was fully automated, involving the removal of duplicate images by comparing the OCR-extracted text from each image. Images containing no text or lacking answer options labeled A, B, C, and D were also removed. To ensure the accuracy of the cleaning process, random samples were extracted from the dataset for manual verification.

The second-to-last step in the preparation of the dataset was to match the questions with their corresponding answers. We

[1]https://github.com/PaddlePaddle/PaddleOCR

used a simple PDF text extraction tool and cleaning scripts to extract answers from Professional Exams files, as the format of the answer key has remained consistent over the years. For other exam types, the answers were manually extracted.

Once the dataset was nearly complete, we decided to add more metadata, specifically indicating whether a task requires additional data, such as an image or table, to answer the question, or if it is entirely text-based. To automate this task, we utilized the Qwen2-VL-7B-Instruct[2] model to analyze and classify the questions. We used the system and user prompt shown below.

```
System prompt:
    You will see a picture of a Polish
    task.If there is more than just a
    text in the image, like an image,
    table, etc, return true. Otherwise
    return false. Respond ONLY with
    'true' or 'false'.

User prompt:
    Does this task need an image
    for context?
```

Table I presents the hardest professional exams categories. A more detailed analysis of these categories is provided in Section V-C.

D. Documents cropping as a tool for building text-based datasets

Dividing a set of questions into individual items can be useful not only for creating image-based datasets but also for constructing textual datasets. When using an OCR engine to extract text from questions that do not require visual context for understanding, the post-processing effort can be significantly reduced compared to standard text extraction techniques, as the data is already segmented into more manageable chunks.

The growing popularity of using multimodal models as a means for data extraction, particularly with the use of Structured Outputs[3], opens up the possibility of generating structured datasets that require minimal or no post-processing. Feeding data to a model in smaller units, rather than entire document pages, can improve the accuracy of data extraction and analysis.

E. Dataset summary

To better understand the finalized dataset, a numerical representation of its potential size and actual results will be given now. Starting with the lowest extraction rate, the 8th-Grade Exams achieved a 27.5% extraction rate. This was followed by the High School Exams at 42.9%, the Middle School Exams at 61.7%, and the Professional Exams with the highest rate of 67.7%. Until 2012, Mathematics and Nature were combined in the middle school exams, resulting in a joint discipline. Table III presents the exact number of questions for each exam and discipline, while Table IV presents the distribution of answer options in the dataset.

F. Dataset availability

The dataset will be publicly available on the Hugging Face page.

IV. Evaluation

A. Evaluation harness

The tool we used to assist with model evaluation was lmms-eval [5] [10], a fork of LM Evaluation Harness[6] [5] framework. Each model was prompted with the following task (in English) in addition to being shown the image:

```
Answer the Polish exam question from
the image. Answer with the good answer
letter only. Possible answers are
A or B or C or D.
```

The evaluation was carried out using two different approaches. The first approach involved resizing the images so that their longest side was 512 pixels, while preserving the aspect ratio and readability. This resizing was performed not only to save memory and reduce computational overhead, but also to evaluate the consequences of image resizing, which are discussed later in the paper. In the second approach, models were evaluated using full-size images to assess their full capabilities. All analyses presented later in the paper are conducted using the outputs from full-sized images.

A common approach for selecting a model's answer involves extracting the log-likelihoods—or probabilities—of each possible option. However, this functionality was not implemented for all of the models evaluated. Therefore, we generated the answers in text format and processed them post hoc by stripping any punctuation or extraneous content beyond the answer letter.

Questions with answers other than A, B, C, or D were excluded from the evaluation. This prompting strategy also simulates a more realistic usage scenario, reflecting how such models are often used in everyday applications.

B. Evaluated models

For the general evaluation on resized images (the 1st approach), we selected 18 different open-weight models across 7 model families, with sizes ranging from 2 billion to 72 billion parameters. We also included one closed-weight GPT model in the evaluation.

GPT 4.1: gpt-4.1-2025-04-14 [7]

Qwen: Qwen2.5-VL (3B-Instruct, 7B-Instruct and 72B-Instruct) [2], Qwen2-VL (2B, 2B-Instruct, 7B and 7B-Instruct) [13]

InternVL: InternVL2_5-8B [3]

[2]https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[3]https://platform.openai.com/docs/guides/structured-outputs?api-mode=responses

[4]Percentages are approximate and may not sum to exactly 100% due to rounding.

[5]https://github.com/EvolvingLMMs-Lab/lmms-eval

[6]https://github.com/EleutherAI/lm-evaluation-harness

[7]https://openai.com/index/gpt-4-1/

TABLE I: Top Hardest Professional Exam Categories based on top 10 models performance.

Code	Models	Visual tasks (%)	Domain	Category
R.13	8	24.0	Forest Technician	Protection and management of forest resources
MG.11	8	46.3	Underground Miner / Mining Technician	Exploitation of underground deposits
R.15	7	36.7	Inland Fish Farmer Technician	Organization of fish farming operations in aquaculture
MG.30	7	44.2	Optician Technician	Production and repair of visual aids
R.14	6	36.0	Forest Technician	Utilization of forest resources
M.21	6	55.3	Blacksmith	Production and repair of blacksmith products
R.02	5	55.3	Farmer / Agribusiness / Agricultural Technician	Agricultural production
RL.06	5	40.7	Horse Trainer	Horse riding and training

TABLE III: Distribution of Questions Across Exams and Disciplines

Exam	Discipline	Original Questions	Questions
8th-Grade Exam	Polish Language	101	9
8th-Grade Exam	Mathematics	99	46
Middle School Exam	Mathematics and Nature	346	152
Middle School Exam	Mathematics	184	96
Middle School Exam	Nature	192	79
High School Exam	Biology	169	21
High School Exam	Physics	399	154
High School Exam	Mathematics	492	280
Professional Exam	Arts	3480	2547
Professional Exam	Mechanical, Mining and Metallurgical	30240	21057
Professional Exam	Agriculture and Forestry	23200	14905

TABLE IV: Distribution of answer options in the dataset.[4]

Answer Option	Count	Percentage (%)
A	9,564	24.3
B	10,235	26.0
C	10,075	25.6
D	9,457	24.0
Other	15	0.04
Total	39,346	100.0

Llava-HF: 1.5 (7B and 13B), v1.6 (Mistral-7B and 34B), NeXT (Llama3-8B and 72B) [11]

Phi: Phi-4 [1]

Idefics2: Idefics2 (8B and 8B–chatty) [9]

InstructBLIP: Vicuna-7B

C. Cropped images results

The general accuracy of the evaluated models on resized images is summarized in Table V.

As shown in the table, the Qwen models dominate this part of the evaluation, despite being trained primarily on English and Chinese data. All other models tend to hover around the 25% random-guess baseline. The Qwen2-2B model began generating random tokens instead of answering the questions, which explains its low accuracy score.

The LLaVA models, even in their largest 72B variant, demonstrate that model size alone is insufficient; the quality and relevance of the training data are the primary determinants of performance. These models, trained on English data, achieve at most 28.06% accuracy.

TABLE V: Performance of models on images resized to 512px

Model	Size	Acc. (%)
Qwen/Qwen2.5-VL-72B-Instruct	72B	60.56
Qwen/Qwen2.5-VL-7B-Instruct	7B	44.82
Qwen/Qwen2.5-VL-3B-Instruct	3B	37.31
Qwen/Qwen2-VL-7B-Instruct	7B	34.16
Qwen/Qwen2-VL-7B	7B	29.04
OpenGVLab/InternVL2/5-8B	8B	28.69
llava-hf/llava-next-72b-hf	72B	28.06
microsoft/Phi-4-multimodal-instruct	14.7	27.65
Qwen/Qwen2-VL-2B-Instruct	2B	26.02
llava-hf/llava-v1.6-34b-hf	34B	25.91
HuggingFaceM4/idefics2-8b-chatty	8B	25.76
HuggingFaceM4/idefics2-8b	8B	25.09
random guessing	-	*25.00*
llava-hf/llava-1.5-7b-hf	7B	24.34
Salesforce/instructblip-vicuna-7b	7B	24.32
llava-hf/llava-1.5-13b-hf	13B	24.31
llava-hf/llava-v1.6-mistral-7b-hf	7B	24.2
llava-hf/llama3-llava-next-8b-hf	8B	24.16
Qwen/Qwen2-VL-2B	2B	20.18

TABLE VI: Performance of models on full-size images

Model	Size	Acc. (%)	No resize increase
GPT-4.1	-	69.77	-
Qwen/Qwen2.5-VL-72B-Instruct	72B	65.25	+7.44%
Qwen/Qwen2.5-VL-7B-Instruct	7B	48.95	+9.21%
Qwen/Qwen2.5-VL-3B-Instruct	3B	41.16	+10.32%
Qwen/Qwen2-VL-7B-Instruct	7B	46.69	+36.68%
Qwen/Qwen2-VL-7B	7B	40.58	**+39.74%**
OpenGVLab/InternVL2/5-8B	8B	28.56	-0.45%
llava-hf/llava-next-72b-hf	72B	28.24	+0.64%
microsoft/Phi-4-multimodal-instruct	14.7	28.77	+4.05%
Qwen/Qwen2-VL-2B-Instruct	2B	31.73	+21.94%
llava-hf/llava-v1.6-34b-hf	34B	26.01	+0.39%
HuggingFaceM4/idefics2-8b-chatty	8B	25.25	-1.98%
HuggingFaceM4/idefics2-8b	8B	25.18	+0.36%
llava-hf/llava-1.5-7b-hf	7B	24.32	-0.08%
Salesforce/instructblip-vicuna-7b	7B	24.32	0%
llava-hf/llava-1.5-13b-hf	13B	24.34	+0.12%
llava-hf/llava-v1.6-mistral-7b-hf	7B	24.32	+0.5%
llava-hf/llama3-llava-next-8b-hf	8B	24.19	+0.12%
Qwen/Qwen2-VL-2B	2B	20.05	-0.64%

D. Full image results

As seen in Table VI, model accuracy approaches practical usefulness only for larger models, such as GPT-4.1 and Qwen2.5-VL-72B-Instruct. Other models, even those that performed above the guess rate of 25%, still yield unsatisfactory results. These findings highlight the need for benchmarks like this one to rigorously evaluate model performance on real-world tasks. Moreover, they underscore the substantial room for improvement in smaller or mid-sized vision-language models.

E. To resize or not to resize

Depending on the model version, the performance gains from using high-resolution images vary. The Qwen2.5 models show modest improvements of just over 10%, while the Qwen2 models demonstrate increases approaching 40%. Notably, models that perform poorly on resized images tend to see limited improvement even when image quality is not altered.

These findings suggest that when image downscaling is necessary, newer model versions should be preferred to mitigate the negative effects of reduced image quality.

V. Results Analysis

A. Confusion analysis

To better understand the models' "thinking" processes—that is, the sequence of statistical and text-processing steps they use to generate answers—a deeper analysis of their answer choices is required. Each confusion matrix in Figure 3 illustrates the relationship between the true answers (y-axis) and the answers predicted by the models (x-axis). This enables a granular examination of how frequently the top three performing models confuse specific answer options.

One of the most notable insights from these matrices is the visual representation of the models' uncertain guesses. While

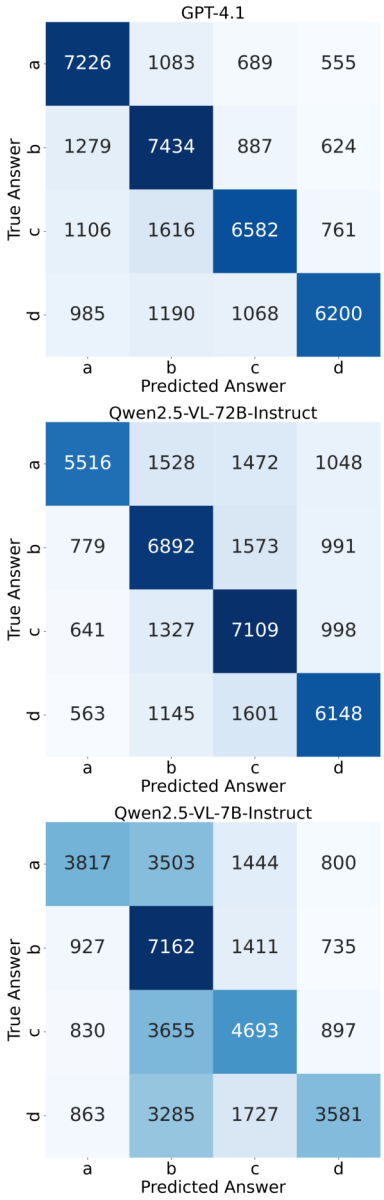


Fig. 3: Confusion matrices for different models.

the GPT model demonstrates the best overall performance across most answer classes, it is slightly outperformed by the 72B Qwen model on questions with "c" as the true answer. When uncertain, GPT exhibits a bias toward selecting earlier answer options, with the likelihood of selection decreasing progressively across the choices. Notably, it appears to almost avoid choosing the furthest option, "d", under uncertainty.

The Qwen models tend to favor the middle answer choices when in doubt. The larger 72B model most frequently guesses "b" or "c" under uncertainty, reflecting a moderate central bias. In contrast, the smaller 7B model exhibits a strong preference for answer "b," even to the extent that it accidentally outperforms the larger model in predicting this specific answer.

TABLE VII: Performance of vision-language models on visual- and text-based tasks.

Model	Visual Acc. (%)*	Text Acc. (%)*
GPT-4.1	65.47	73.07
Qwen/Qwen2.5-VL-72B-Instruct	59.09	69.98
Qwen/Qwen2.5-VL-7B-Instruct	44.72	52.19
Qwen/Qwen2-VL-7B-Instruct	42.74	49.71
Qwen/Qwen2.5-VL-3B-Instruct	38.35	43.32
Qwen/Qwen2-VL-7B	36.21	43.92
Qwen/Qwen2-VL-2B-Instruct	30.55	32.64
microsoft/Phi-4-multimodal-instruct	28.33	29.11
OpenGVLab/InternVL2_5-8B	28.37	28.71
llava-hf/llava-next-72b-hf	27.14	29.09
llava-hf/llava-v1.6-34b-hf	24.87	26.88
HuggingFaceM4/idefics2-8b-chatty	25.37	25.16
HuggingFaceM4/idefics2-8b	24.73	25.53
llava-hf/llava-1.5-13b-hf	23.67	24.86
llava-hf/llava-1.5-7b-hf	23.62	24.86
Salesforce/instructblip-vicuna-7b	23.53	24.92
llava-hf/llava-v1.6-mistral-7b-hf	23.72	24.71
llava-hf/llama3-llava-next-8b-hf	23.21	24.95
Qwen/Qwen2-VL-2B	23.42	17.47

*Task classification is based on automated labeling with an estimated 12.2% error rate. See Section V-B for details.

This pronounced bias suggests a reliance on heuristics or a limited ability to disambiguate between options in uncertain scenarios.

B. Visual vs. Text-Based Tasks

To confirm the accuracy of Qwen's classification of images that require additional context, such as a picture or table, a random sample of 1,000 images was selected for manual verification. The verification revealed that 122 of the 1,000 images were misclassified, with the vast majority being false positives (120 out of 122). This disproportion may indicate that the prompt did not sufficiently guide the model to identify tasks requiring additional context. All subsequent tables that report model results based on the image-context grouping should be interpreted in light of an estimated 12.2% labeling error, with a 95% confidence interval ranging from approximately 10.2% to 14.2%. While this level of error is acceptable for exploratory analysis and broad grouping, it is not sufficiently accurate for training downstream models. We therefore emphasize that findings based on this stratification are preliminary and should not be overinterpreted without more precise annotation.

Out of 39,331 questions considered when calculating the accuracy of the models, 17,058 were marked as visual-based and 22,273 were marked as text-based. Table VII shows results of each model both on visual- and text-based tasks.

With few exceptions, most of the tested models exhibit higher accuracy on text-based questions compared to those requiring visual context. This discrepancy stems from the fact that vision-language models are still primarily trained on textual data and are less focused on the deep analysis of figures, tables, or the recognition of specific and highly niche objects. Visual tasks often demand fine-grained perception, complex reasoning, and the integration of visual features with language understanding. Even the most advanced models evaluated remain far from achieving human-like reasoning and multidimensional understanding. This performance gap underscores the limitations of current multimodal training approaches and highlights the need for more diverse, richly annotated datasets that better capture the complexities of visual reasoning.

C. Professional Exams categories analysis

This analysis was conducted on the top 10 best-performing models, as it required a level of domain knowledge sufficient to demonstrate informed responses rather than mere guessing. There are 150 unique category codes, and for the easiest and hardest analyses, the codes that appeared most frequently among the top 20 easiest and hardest categories for the selected models were identified. The proportion of visually-based tasks was also considered when assessing question difficulty, as text-based tasks were generally easier for the models to answer as shown in Section V-B. This led to the exclusion of categories with more than 70% visual tasks among the hardest categories, and less than 30% visual tasks among the easiest categories. These thresholds were selected to create a clear contrast between predominantly visual and predominantly text-based categories, while also preserving a sufficient number of categories to support a robust and meaningful analysis. These thresholds were selected to create a clear contrast between predominantly visual and predominantly text-based categories, while also preserving a sufficient number of categories to support a robust and meaningful analysis.

The most challenging categories were closely tied to the limited availability of widely accessible resources for model training, particularly those that are highly niche and vocational. Eight categories fell into this group, the most difficult of which were *Protection and Management of Forest Resources* and *Exploitation of Underground Deposits*.

The knowledge required to become one of these specialists is heavily underrepresented in training corpora and typically demands extensive hands-on experience and practical training.

In contrast, most of the easiest categories belonged to the broader domain of mechanics. Notably, Shipbuilding works—a subcategory within this domain—appeared among the top 20 easiest categories in nine models. With approximately 1.64 billion cars in the world, mechanical data is widely available and has certainly been included in model training. The second easiest domain was *Environmental Protection*, a highly discussed topic, particularly in today's context of increasing global awareness and concern.

VI. Comparison against humans

Comparing the accuracy of Multimodal Language Models to exact human performance is not feasible, as human results are published only for entire exams rather than individual questions for all but the High School exams. All exams, except for the Professional exams, contain a significant number of open-ended questions, which were not included in our dataset.

Additionally, some closed-ended questions may have been excluded from the dataset due to the postprocessing described earlier. Data on human performance in professional exams were extracted from the WaszaEdukacja website[8], which is not affiliated with the Central Examination Board (CKE). Therefore, the exam content presented on the site may not fully correspond to the official questions used in the models' evaluation. The purpose of this section is not to compare human knowledge to model performance directly, but rather to evaluate whether the difficulty reflected in human performance for a given year aligns with the general performance of the models—even if the exact questions differ. This type of analysis applied to textual exams dataset has been proposed and shown to correlate well with human performance in [8]. For the visualized models performance, the top 3 performing models were chosen: GPT-4.1, Qwen2.5-VL-72B-Instruct and Qwen2.5-VL-7B-Instruct.

A. Eighth-grade exam

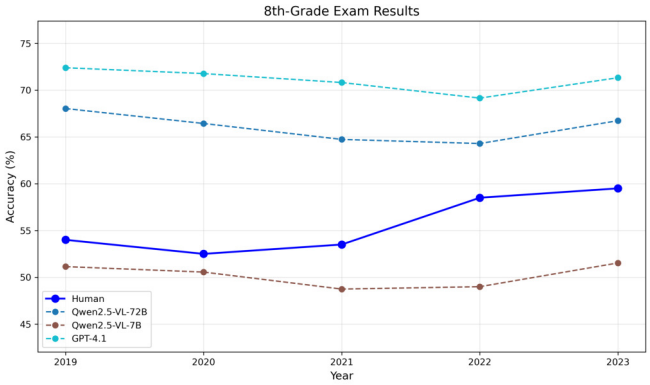


Fig. 4: 8-th grade selected models results compared to average human results.

In the combined results from all 8th-grade question categories—grouped due to the limited number of questions available for individual analysis—the GPT model emerges as the top-performing model. All models exhibit similar year-to-year fluctuations in accuracy, differing primarily in their overall performance levels. GPT achieves an average accuracy of 69.87%, while the Qwen models follow with 64.48% and 48.99%, respectively. A notable divergence between human and model performance occurred between 2020 and 2022, where the average human accuracy improves compared to the previous years, all the evaluated models experience a decline in performance.

[8]https://waszaedukacja.pl/egzaminy/zawodowy

B. Middle school exam

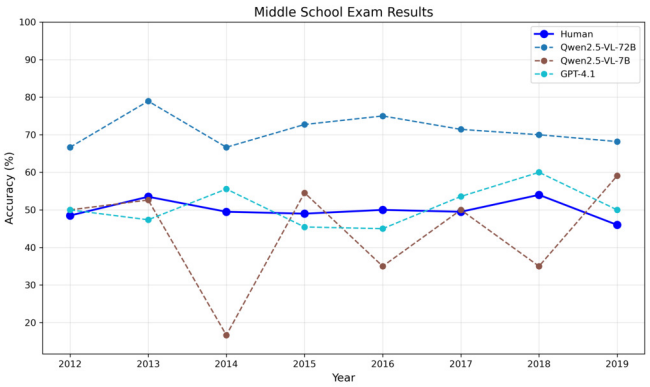


Fig. 5: Middle school selected models results compared to average human results.

Qwen2.5-VL-72B performs exceptionally well on the Middle School exams, achieving an impressive average accuracy of 77.14%, clearly leading among the evaluated models. In contrast, GPT-4.1 achieves a more modest average accuracy of 58.02%, aligning more closely with the smaller Qwen2.5-VL-7B model, which scores 51.98%. Interestingly, the performance of the 72B Qwen model mirrors the trend of human results across the years, with the exception of 2018, where a noticeable deviation occurs. GPT-4.1 maintains accuracy levels generally comparable to human performance, while the Qwen 7B model exhibits high variability—ranging from below 20% to as high as 60%—indicating inconsistency in its performance across exam years.

C. High school exam

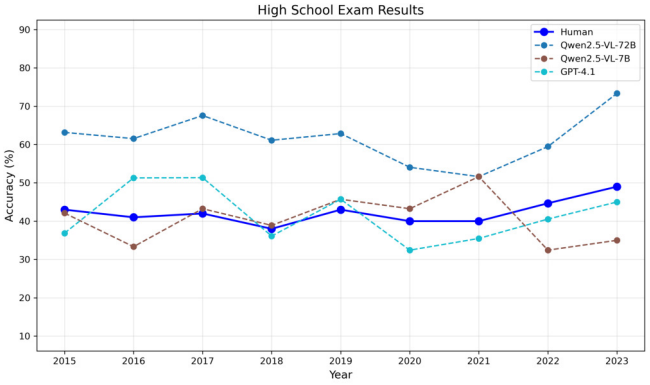


Fig. 6: High school subjects selected models results compared to average human results.

For the high school exams, the Qwen2.5-VL-72B model's performance mirrors the overall trend of human performance over the years, with rises and declines occurring in parallel, though at a different absolute level of accuracy. In contrast, the other two models hover around the human performance baseline, with GPT-4.1 achieving an average accuracy of 53% and Qwen2.5-VL-7B scoring 47.34%.

D. Professional exam

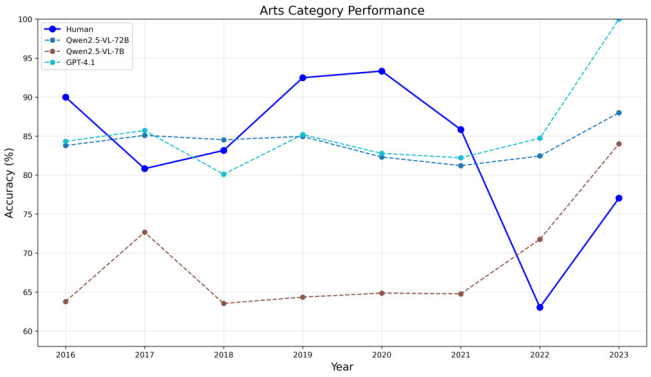


Fig. 7: Arts selected models results compared to average human results.

1) Arts: In contrast to the models' performance, fluctuations in human accuracy are particularly noticeable in the arts professional exams. The two best-performing models consistently maintain accuracy above 80%, while human performance occasionally drops below 65%, highlighting a significant performance gap in this category.

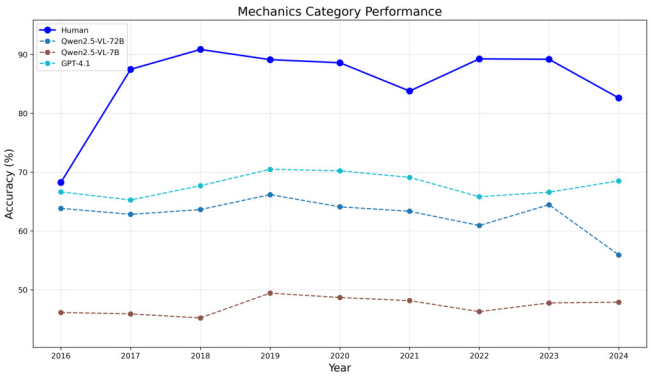


Fig. 8: Mechanics (including mining and metallurgy) selected models results compared to average human results.

2) Mechanical, Mining and Metallurgical: In the mechanical category, human performance significantly surpasses that of the models, indicating that in highly specialized domains, models still underperform and have considerable room for improvement. All models exhibit similar trends of rising and falling performance over time, though at different levels of average accuracy. This consistent fluctuation pattern across models suggests a correlation between model size (i.e., number of parameters) and overall performance, with larger models achieving higher accuracy while following the same performance trajectory.

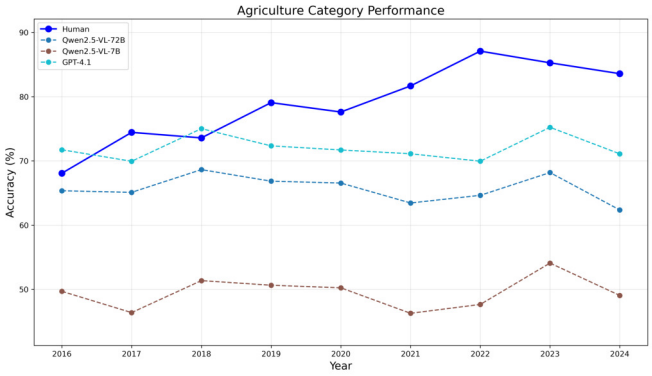


Fig. 9: Agriculture (including forestry) selected models results compared to average human results.

3) Agriculture and Forestry: The accuracy trends for models and humans on agricultural exam questions appear to move in opposite directions: while human performance steadily improves over the years, model performance consistently declines. This contrast highlights the limitations of using vision-language models as a benchmark for assessing exam difficulty. It suggests that what may seem harder for a model may, in fact, be easier for a human—and vice versa—underscoring the fundamental differences in reasoning strategies and domain familiarity. This emphasizes the need for caution when relying on model-based metrics in educational evaluation or exam design.

VII. Additional OCR context

In this section we will try to answer the question if prompting the model with additional context from the image—in the form of Optical Character Recognition (OCR)-extracted text—can improve its accuracy and overall performance. Including OCR-extracted text from the image in the prompt itself can enhance model performance by providing additional context that may not be captured through visual input alone, bridging the gap between modalities. To encourage the model to utilize this additional information when answering the question, we appended the following text to the prompt:

```
Use the OCR text as well as the image
to answer the question. OCR text:
```

The following models were selected to evaluate the benefits of additional context: **Qwen2.5-VL-7B-Instruct**, **Qwen2-VL-7B**, **Qwen2-VL-7B-Instruct**, **Llama3-llava-next-8b-hf**, **Phi-4**.

For the OCR engine, EasyOCR [9], was used, as it handles Polish characters more effectively than the previously used PaddleOCR. The text indicating the beginning of each question did not contain any Polish letters, so the earlier use of PaddleOCR did not result in processing errors. The extracted text was provided in the prompt verbatim, without any post-processing. The OCR was performed on the original, pre-resize image.

[9]https://github.com/JaidedAI/EasyOCR

A. Results

Across the tested models, only the Phi-4 model exhibited notable improvements when prompted with additional OCR context. This outcome may be attributed to its modular architecture, which employs separate adapters (LoRAs) for vision, speech, and text. Phi-4 processes each modality independently and integrates them later, enabling more effective fusion of external textual input with visual features.

Interestingly, the Qwen2-VL-7B-Instruct model showed a decrease in performance with the additional OCR context. This suggests that the injected text may have conflicted with the model's internal visual interpretation, potentially leading to confusion or interference during reasoning. Unlike Phi-4, Qwen models use a single vision encoder that processes visual inputs and aligns them with textual representations via rotary embeddings before feeding the combined data into a unified Transformer stream.

The overall results suggest that, while promising in theory, incorporating OCR context when processing images of questions does not consistently improve model performance and may, in some cases, degrade it. Nevertheless, the observed gains in Phi-4 underscore that, when properly integrated, OCR context can enhance performance. Therefore, it is important to continue exploring modality-specific architectures and training strategies.

TABLE VIII: General performance of selected models with additional OCR context

Model	Acc.	OCR Acc.	Increase
Qwen2.5-VL-7B-Instruct	48.95	50.24	+2.64%
Qwen2-VL-7B	40.58	40.84	+0.64%
Qwen2-VL-7B-Instruct	46.69	46.16	-1.14%
Llama3-llava-next-8b-hf	24.19	25.07	+3.64%
Phi-4	28.77	31.46	**+9.35%**

VIII. Limitations

While this benchmark is a necessary step toward assessing a model's knowledge, it evaluates only the model's final answers to closed-ended questions and does not account for the reasoning process behind those answers. Although the provided responses may be sufficient for tasks such as assisting a student in finding a correct answer, they fall short in educational contexts where the underlying reasoning is also evaluated. In such scenarios, a model might provide the correct answer but fail to justify it coherently, potentially misleading users about its understanding. Additionally, the dataset used in this benchmark represents idealized conditions in which all questions are fully visible and neatly typewritten. As a result, the benchmark assesses visual recognition only within these controlled parameters and does not reflect the broader variability encountered in real-world visual inputs, such as handwritten notes, partially obscured text, or classroom environments.

One of the previously mentioned issues is the error rate in Qwen's visual-context task recognition. Although the classification was sufficient for simple comparisons and provided some insight into the increased difficulty compared to text-only tasks, the error rate may render the data unsuitable for other applications.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang. Phi-4 technical report, 2024.

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[4] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, and P. Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024.

[5] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, 07 2024.

[6] A. P. Gema, J. O. J. Leang, G. Hong, A. Devoto, A. C. M. Mancino, R. Saxena, X. He, Y. Zhao, X. Du, M. R. G. Madani, et al. Are we done with MMLU? *arXiv preprint arXiv:2406.04127*, 2024.

[7] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[8] K. Jassem, M. Ciesiółka, F. Graliński, P. Jabłoński, J. Pokrywka, M. Kubis, M. Jabłońska, and R. Staruch. LLMzSzŁ: a comprehensive LLM benchmark for Polish, 2025.

[9] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models?, 2024.

[10] B. Li, P. Zhang, K. Zhang, F. Pu, X. Du, Y. Dong, H. Liu, Y. Zhang, G. Zhang, C. Li, and Z. Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024.

[11] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015.

[13] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[14] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI, 2024.

[15] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, and L. Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023.

[16] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 675–680
DOI: 10.15439/2025F8073
ISSN 2300-5963 ACSIS, Vol. 43

Applications and Challenges of Artificial Intelligence in Educational Course Design and Delivery

Daniel Dan
School of Applied Data Science, Modul University Vienna, Vienna, Austria
ORCID: 0000-0002-7251-7899
Email: daniel.dan@modul.ac.at

Anna Wróblewska, Bartosz Grabek, Michał Taczała
Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
ORCIDs: A: 0000-0002-3407-7570, B: 0009-0008-4917-8017, M: 0009-0004-4407-2361
Emails: anna.wroblewsk1, bartosz.grabek.stud, michal.taczala.stud@pw.edu.pl

Minoru Nakayama
School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo, Japan
ORCID: 0000-0001-5563-6901
Email: nakayama@meiji.ac.jp

***Abstract*—This survey reviews AI goals and tools for: (1) preparing educational materials, (2) interacting with teachers and students, and (3) assessing the results and providing feedback with (semi-)automatic methods. As a summary, we provide the crucial challenges to be tackled and discuss the associated ethical concerns.**

***Index Terms*—Artificial Intelligence, Automated Assessment, Higher Education, Large Language Models, Learning Analytics**

I. Introduction

RECENT advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP), particularly the rise of Large Language Models (LLMs), have transformed human-computer interaction across domains. Significant changes are evident in education, where artificial intelligence supports students in their everyday studies and helps teachers deliver material more engagingly and effectively [1].

A recent survey indicates that teachers now use AI more often than students [2]. Our complementary survey reviews NLP methods for course design, delivery, and personalised assessment, including technical details and their implementation. We discuss how these solutions are designed, integrated into higher-education workflows, and evaluate their effectiveness for both students and teachers. The listed solutions form a strong basis for fully integrated, technology-aided education, and we present an example AI-driven engineering course.

Our research questions are:

- What current approaches apply NLP to course design and delivery?
- How do these methods fit educational settings?
- How can they be integrated into different course aspects?

The survey follows the class flow. Section II reviews NLP for course-content creation; Section III examines technology that enhances live teaching. Section V covers automated assessment, proctoring, and anti-cheat tools, and Section VI, explores personalised learning systems. Finally, we outline implementation challenges and ethics.

II. Course Content Creation & Modification

Designing a university course involves transforming ideas into course materials that are engaging, academically sound, and easy-to-follow [3], [4]. AI now supports instructors in this process. Designing teaching material must consider principles that support learning [5], [6], and how students learn and are assessed [7]. In practice, a course is split into several lectures – seminar, frontal, or interactive. While face-to-face teaching remains dominant, innovative educational technologies reduce student passivity, enabling them to interact and actively learn the class material [8]. Student assessment remains delicate and error-prone. Modern methods emphasize continuous homework and assignments, a mid-term checkpoint, and a final written, oral, or project-based exam [9]. At course launch, instructors draft a syllabus, a "contract" that defines objectives, lecture sequence, assessment, and permitted tools [10]. Some universities now ban student use of generative AI for homework [11]. Rather than prohibiting evolving tools, we advocate channeling them, e.g., letting chatbots handle low-stakes assessment to save instructor time. AI already cuts preparation time: a 2024 survey of 144 instructional designers found that 67% reported moderate-to-significant efficiency gains, and 64% let ChatGPT draft objectives or quizzes first [12].

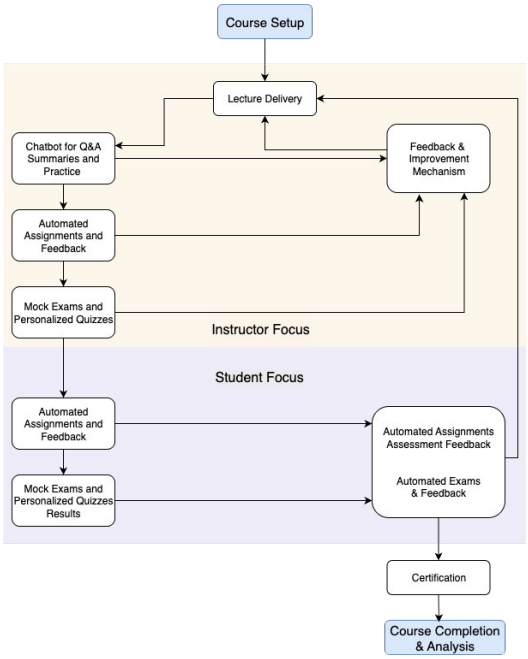


Fig. 1. Flowchart of subsequent steps in educational academic course delivery and students' assessment, up to the final certification. Note: light yellow are boxes more in the focus of the instructor, and grey ones are more for students

Figure 1 summarises the course flow. In an AI-enhanced version, instructors become supervisors and facilitators, while AI acts as a coach, evaluator, and feedback provider [13]. We will touch especially on the important ethical aspects and challenges while working on the creation of educational tools. The crucial questions here are:

1) How can AI/LLMs generate lecture materials and exercises?
2) How can technology structure the course and each session?
3) How can automated tools improve grading and feedback?
4) How do we evaluate the quality of AI-prepared material?
5) What are the pros and cons of teacher–AI collaboration?

A. How can AI/LLMs generate lecture materials and exercises?

AI tools can now draft lesson plans aligned with standards [14], [15], [16]. By specifying duration, lesson count, learner profile, and desired outcomes, instructors obtain varied analogies, examples, and explanations. Prompt engineering is continually refined [17], and multiple versions can be generated to suit individual students or fresh perspectives [18].

B. How can we use AI for detailed lesson planning and time allocation?

ChatGPT-style assistants quickly produce weekly and daily plans – topics, activities, assignments, assessments [19]. They can also recommend external videos and readings and link them to objectives, ensuring self-directed study complements the course [20]. Beyond text, AI can seed graphic-novel tasks, game-based projects, and debate topics to deepen engagement [21], [22], [23].

C. Grading and feedback generation

Timely, actionable feedback is hard at scale [24]. AI can create individualized quizzes and instant comments, drawing on prior performance, while human oversight preserves fairness and limits bias [25], [26].

D. Materials quality evaluation

AI-generated content still requires a documented two-pass human review: a subject-matter expert checks accuracy, then an instructional designer ensures pedagogy and accessibility [27], [28], [29]. Automated fact-checkers, readability analyzers, and rubric-based scores (e.g., AIGDER $\geq$ 85/100) flag issues for revision [30], [31], [32], [33]. Student "muddiest-point" feedback pinpoints residual confusion for rapid remediation [34], [35], [36], [37], [38].

E. Interaction between instructors and AI

Instructors work in cycles—prompt, review, refine, and embed pedagogy—when collaborating with AI. Attitudes range from scepticism about bias and deskilling [39] to enthusiasm for efficiency and creativity [40]. Empirical studies show that training and guided use shift beliefs and improve outcomes [41], [42].

III. Teaching

Generative AI tools can create diversified and personalised teaching materials and even automate routine "paper work" such as emails, deadline checks and consistency control [43]. Yet quality is essential; recent studies stress rigorous evaluation, from re-using credible sources [46] to iterative assessment protocols [44], [45].

Beyond content creation, AI supports real-time observation of teaching and learning. A design study [47] foregrounds learning-analytics dashboards -single screens aggregating key indicators- for ongoing improvement. Surveys classify multiple dashboard types and confirm they mainly serve teacher supervision [48]

A. Interactive content generation

Automatic content or question generation hinges on carefully crafted LLM prompts that spell out context, audience, and difficulty. Retrieval-Augmented Generation can instead feed the model verified knowledge before output [38].

B. Engaging lecture delivery

During lectures, AI can act as a creative co-teacher—suggesting analogies or examples [49], powering avatar instructors that field questions [50], and turning text into images, videos or podcast teasers for multimedia slides [51], [52]. Students can critique AI answers to hone critical thinking and visualise experiments via dynamic images.

C. Real-time learning process supervision

Real-time supervision captures how, not just what, students learn. AI dashboards visualise engagement and performance from digital environments [53]; open-learner models update each learner's knowledge profile [54]. Dashboards must stay interpretable to be actionable [55], [56], [57]. Predictive analytics within these dashboards can flag at-risk or low-engagement students for prompt support [58], [59], [60], [61].

D. AI chatbot Q&A summaries and practice

LLM-powered chatbots have spread rapidly across higher education. Surveys find enthusiasm tempered by doubts: ease of use is clear, but cognitive gains are still debated [62], [63]. When offered as optional study partners, chatbots can scaffold distance or dyslexic learners through personalised encouragement, summaries and practice prompts [64], [65].

IV. Aid in Academic Activities

A. Coding assistance

One of the core competencies acquired by engineering and computer-science students is coding. Programming is indispensable for many assessed projects. Requirements vary—Mathematical Computation students code in R, Computer-Science majors in Python or Java, Robotics students in low-level hardware languages—but all must solve complex problems. AI-powered coding assistants (ChatGPT, Gemini, Claude, Grok) now provide instant code generation, refactoring, explanations and debugging via LLMs plus web search.

Recent work compared ChatGPT-assisted learners with self-directed peers: final scores were similar, but ChatGPT users debugged more, refined AI code and, citing ease-of-use, planned to keep using the tool [66].

These findings show how AI tools scaffold programming at every level, offering step-by-step guidance and fast debugging. Effectiveness, however, hinges on good prompts and critical evaluation; instructor-led prompt training markedly boosts outcomes [67]. LLMs still struggle with large code-bases, may hallucinate and often default to generic patterns. Over-reliance can erode academic integrity, so institutions need guardrails that balance tool use with independent problem-solving.

B. Academic writing aids

A review [68] groups modern AI writing tools into literature management, drafting, plagiarism checks, data analysis and other features. They accelerate brainstorming by mining vast corpora and adding predictive analytics throughout the research workflow.

V. Assessment

A. Assessment generation

Automated assessment is now a popular NLP topic because it saves teachers time and money. Its goal is to achieve scores similar to human graders with far less effort and cost. Many methods exist (e.g., semantic similarity), but large language models (LLMs) currently dominate. Their chief flaw is hallucination—inventing incorrect information, which undermines reliability. One mitigation is retrieval-augmented generation (RAG): a standard LLM linked to a vetted knowledge base. Retrieving evidence relevant to each question reduces hallucination and equalises scores [69], though irrelevant hits can still occur. LLMs further support AI-assisted grading [70]; yet their human-like prose raises doubts about assessment validity. Human oversight, therefore, remains essential [71].

LLM-generated assessments can still be inconsistent. The "LLM-as-a-judge" pattern—a second model auditing the first—adds a control layer that flags misgraded answers and outliers with high accuracy [72].

B. Automated final exam / exam proctoring

AI-driven NLP already handles full exams—multiple-choice and essay tasks. Studies confirm feasibility [73] but show score distributions can narrow, and domain-knowledge gaps (e.g., biomedical items) cut accuracy [74]. Essay scoring benefits from LLM text processing [75], yet quality hinges on prompt tuning: focusing on different rubric points shifts grades [76]. Performance also depends on language coverage [77], and research now extends LLM scoring to oral exams [78]. Item-response theory (IRT) frameworks could anchor AI scoring, replacing today's opaque heuristics [**?**].

C. Anti-cheat solutions

Anti-cheat tools remain vital [79]. Traditional detectors such as Turnitin catch plagiarism [80], while stylometric authorship checks compare writing style [81]. These methods falter on short answers. GPT-based classifiers now spot AI-generated text but still produce false positives. The most reliable—though intrusive—approaches monitor cursor or keystrokes during exams [82], [83].

VI. Personalized Learning

Personalized learning represents a fundamental shift in the educational process, aiming to replace the singular, standardized educational experience with a customised, learner-centred one delivered via intelligent tutoring systems, content recommendations or adaptive assessments.

A. Adaptive assessments

This protocol first creates a large RAG-based question bank, each item labelled for difficulty and topic. The instructor sets the time window, duration, item count and proctoring; the system then randomises—or adaptively selects—questions for every student. Learners log in while camera, safe-browser and image-recognition tools safeguard integrity. Built-in timers show time left and auto-submit; real-time alerts flag suspicious activity. After all submissions, RAG compares answers with a rubric, awards partial credit and exports grades to the LMS. Plagiarism scans and on-the-fly follow-up questions protect originality, while analytics refine subsequent tests.

Considerably less research views these tests from the learner's side. Co-designing AI-based assessments with students can boost critical thinking and creativity [**?**], and meta-awareness that items are machine-generated maps to Bloom's highest levels—analysis, synthesis and evaluation [85].

B. Personalized feedback

Providing reasonable feedback is a crucial part of the whole AI-supported exam-conducting ecosystem. It helps students fill gaps in knowledge, focus on the right areas, and understand the material faster. Research shows that the best feedback is timely, specific, and actionable [86]. Equally important is wording it pedagogically so it truly motivates learning. The most common solution today is a data-driven approach: transformer models (e.g., BERT) can evaluate many aspects of an answer through their understanding of context [87]. Yet merging expert knowledge with AI feedback is challenging; LLM comments are often generic and miss key details [86]. LLMs may also hallucinate—studies report that about 15% of feedback can be totally incorrect [86]. Prompt design, therefore, matters: telling a model to "behave as a teacher," giving step-by-step instructions, and asking it to analyse first and comment second all raise quality.

Despite these gains, the model is still far from perfect, so retrieval-augmented generation (RAG) can further improve performance. RAG combines an LLM with local data sources such as presentations or notes, ensuring feedback stays strictly aligned with course material [88]. A complementary method splits the task among several specialised agents, each handling a smaller part; this helps catch mistakes and adjust tone [88].

Feedback during learning is tentative diagnostic information, so a relationship of mutual trust between learner and instructor is essential. When automation intervenes, its effect on that relationship must be monitored; simple chatbots can bore users and discourage sustained dialogue.

LLM-based feedback systems have now been piloted in real courses, and learner perceptions are being evaluated [89]. Careful messaging can promote reflection on one's learning situation [90], and design choices—topic coverage, timing, tone—must be optimised for each activity [91].

VII. Challenges & Future Directions

In recent years, many studies have explored the possibilities, challenges and threats of AI tools in education. Yet most merely list what AI can do; few run carefully designed experiments to test usability and reliability. The key task now is to conduct comparative studies—contrasting AI-based recommendations, testing, grading and feedback with long-standing practices. For instance, we should compare LLM resources with earlier crowd-sourcing platforms used by students before chatbots.

We must also continuously measure how well up-to-date, fine-tuned and well-prompted LLMs acquire and refresh knowledge.

Moreover, ethical challenges loom large; automated tools must remain especially sensitive to educational contexts. As AI is introduced, we must be aware of data and algorithmic biases. Models trained on large corpora can mirror societal, cultural or institutional prejudices, prioritising students by past performance or learning style and producing unfair outcomes tied to race, language, socioeconomic status or disability. Thus transparency and explainability are needed for every recommendation or assessment, yet remain elusive because LLMs are inherently opaque.

Student privacy is also at stake, particularly under learning surveillance. Over-reliance on automation, coupled with inadequate informed consent, can erode professional judgement on both sides of the classroom. Unequal digital access—whether through language or technology—may widen existing educational gaps.

Mitigation demands careful design: human-in-the-loop oversight, informed student choice, and channels for learner voice. Done well, AI frees teachers to spend more quality, personalised time with each student.

Acknowledgments. This work was supported by the EU Horizon Europe OMINO (Overcoming Multilevel Information Overload) grant No. 101086321 and by the Polish Ministry of Education and Science (International Projects Co-Financed program). The views expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Research Executive Agency, which are not responsible for them.

References

[1] Y. Walter, "Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education," Int. J. Educ. Technol. High. Educ., vol. 21, no. 1, p. 15, Feb. 2024, doi:10.1186/s41239-024-00448-3. https://doi.org/10.1186/s41239-024-00448-3

[2] Quizlet, Inc., "State of AI in Education," 2023. [Online]. Available: https://www.prnewswire.com/news-releases/quizlets-state-of-ai-in-education-survey-reveals-teachers-are-surprise-ai-champions-301884427.html

[3] J. Mason and S. Goodenough, "Course creation," in *Distance Teaching for Higher and Adult Education*, Routledge, 2018, pp. 100-120. https://doi.org/10.4324/9780429430930-8

[4] E. Dickey and A. Bejarano, "GAIDE: A Framework for Using Generative AI to Assist in Course Content Development," in 2024 IEEE Frontiers in Education Conf. (FIE), Oct. 2024, pp. 1-9. https://doi.org/10.1109/FIE61694.2024.10893132

[5] J. Singer, R. W. Marx, J. Krajcik, and J. C. Chambers, "Constructing extended inquiry projects: Curriculum materials for science education reform," Educ. Psychol., vol. 35, no. 3, pp. 165-178, 2000.https://doi.org/10.1207/S15326985EP3503_3

[6] R. A. Reiser, "A history of instructional design and technology: Part II: A history of instructional design," Educ. Technol. Res. Dev., vol. 49, no. 2, pp. 57-67, 2001. https://doi.org/10.1007/BF02504928

[7] J. W. Pellegrino, J. D. Bransford, and M. S. Donovan, *How People Learn: Bridging Research and Practice*, Washington, DC: National Academies Press, 1999.

[8] N. Friesen, *The Textbook and the Lecture: Education in the Age of New Media*, Baltimore, MD: JHU Press, 2017.

[9] A. J. Nitko, "Conceptual frameworks to accommodate the validation of rapidly changing requirements for assessments," Curriculum Assess., vol. 1, no. 2, pp. 143-163, 2001.

[10] J. L. Hess and M. S. Whittington, "Developing an effective course syllabus," Nacta J., pp. 23-27, 2003.

[11] K. de Fine Licht, "Generative artificial intelligence in higher education: Why the 'banning approach' to student use is sometimes morally justified," Philos. Technol., vol. 37, no. 3, p. 113, 2024. https://doi.org/10.1007/s13347-024-00799-9

[12] L. McNeill, "Automation or innovation? A generative AI and instructional design snapshot," in The IAFOR Int. Conf. on Education-Hawaii 2024 Official Conf. Proc., 2024, pp. 187-194. https://doi.org/10.22492/issn.2189-1036.2024.17

[13] R. R. Hansen et al., "The effects of an AI feedback coach on students' peer feedback quality, composition, and feedback experience," Læring og Medier, vol. 31, pp. 1-31, 2025. https://doi.org/10.7146/lom.v17i31.148831

[14] C. Lammert, S. DeJulio, S. Grote-Garcia, and L. M. Fraga, "Better than Nothing? An Analysis of AI-Generated Lesson Plans Using the Universal Design for Learning & Transition Frameworks," The Clearing House: A Journal of Educational Strategies, Issues and Ideas, vol. 97, no. 5, pp. 168-175, 2024. https://doi.org/10.1080/00098655.2024.2427332

[15] A. Bucchiarone et al., "Designing and Generating Lesson Plans combining Open Educational Content and Generative AI," in Proc. 27th Int. Conf. Model Driven Eng. Lang. Syst., ACM/IEEE, 2024, pp. 78-86. https://doi.org/10.1145/3652620.3687773

[16] A. Walter, "Utilizing Language-Generating Artificial Intelligence in Educational Planning: A Case Study," J. Interdiscip. Teacher Leadersh., vol. 8, no. 1, pp. 29-59, 2024.

[17] A. Przegalinska and T. Triantoro, *Converging Minds: The Creative Potential of Collaborative AI*. Boca Raton, FL: CRC Press, 2024. https://doi.org/10.1201/9781032656618

[18] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," IEEE Access, vol. 8, pp. 75264-75278, 2020. https://doi.org/10.1109/ACCESS.2020.2988510

[19] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system," BenchCouncil Trans. Benchmarks, Stand. Eval., vol. 3, no. 2, Art. 100115, 2023. https://doi.org/10.1016/j.tbench.2023.100115

[20] A. Plasencia and N. Navas, "MOOCs, the flipped classroom, and Khan Academy practices: The implications of augmented learning," in *Innovation and Teaching Technologies: New Directions in Research, Practice and Policy*, Springer, 2014, pp. 1-10. https://doi.org/10.1007/978-3-319-04825-3_1

[21] F. Holguin, "Exploring AI in Graphic Novels," 2024.

[22] K. Guo, Y. Zhong, D. Li, and S. K. W. Chu, "Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation," Comput. Educ., vol. 203, Art. 104862, 2023. https://doi.org/10.1016/j.compedu.2023.104862

[23] U. Lee et al., "Can ChatGPT be a debate partner? Developing ChatGPT-based application "DEBO" for debate education, findings and limitations," Educ. Technol. Soc., vol. 27, no. 2, pp. 321-346, 2024.

[24] D. D. Stevens, *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*, 2nd ed., New York, NY: Routledge, 2023.

[25] J. Gnanaprakasam and R. Lourdusamy, "The role of AI in automating grading: Enhancing feedback and efficiency," in *Artificial Intelligence and Education-Shaping the Future of Learning*, IntechOpen, 2024, pp. 153-167. https://doi.org/10.5772/intechopen.1005025

[26] I. T. Awidi, "Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool," Comput. Educ. Artif. Intell., vol. 6, Art. 100226, 2024. https://doi.org/10.1016/j.caeai.2024.100226

[27] M. Bazluki, K. Gyabak, and B. Uderman, "Instructor feedback on a formal online course quality assurance review process," Online J. Distance Learn. Admin., vol. 21, no. 2, pp. 1-9, 2018.

[28] W. W. Lee, "Subject Matter Experts and Instructional Designers: Making Distinctions," Perform. Instr., vol. 33, no. 8, pp. 23-25, 1994. https://doi.org/10.1002/pfi.4160330807

[29] R. Mancilla and B. Frey, "Instructional design staffing for online programs," Online J. Distance Learn. Admin., vol. 26, no. 2, Art. n2, 2023.

[30] M. Smith, J. Turner, E. Sanford-Moore, and H. H. Koons, "The Lexile framework for reading: An introduction to what it is and how to use it," in Pacific Rim Objective Measurement Symp. 2015 Conf. Proc., 2016, pp. 411-424. https://doi.org/10.1007/978-981-10-1687-5_27

[31] E. Hoes, S. Altay, and J. Bermeo, "Leveraging ChatGPT for efficient fact-checking," PsyArXiv, Apr. 2023. https://doi.org/10.31234/osf.io/qnjkf

[32] Q. Huang, C. Lv, L. Lu, and S. Tu, "Evaluating the Quality of AI-Generated Digital Educational Resources for University Teaching and Learning," Systems, vol. 13, no. 3, Art. 174, 2025. https://doi.org/10.3390/systems13030174

[33] K. Shattuck, "Focusing research on quality matters," Am. J. Distance Educ., vol. 29, no. 3, pp. 155-158, 2015. https://doi.org/10.1080/08923647.2015.1061809

[34] H. A. Seneviratne and M. Thenabadu, "Evaluation of Student Perceptions on "Muddiest Point" Classroom Assessment Technique Implemented As A Formative Assessment Method," Int. J. Humanit. Art Soc. Stud., vol. 6, no. 1, 2021.

[35] N. Scaria, S. Dharani, and D. Subramani, "Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation," in *Int. Conf. on Artificial Intelligence in Education*, Cham: Springer, 2024, pp. 165-179. https://doi.org/10.1007/978-3-031-64299-9_12

[36] S. Moore, E. Costello, H. A. Nguyen, and J. Stamper, "An Automatic Question Usability Evaluation Toolkit," in *Artificial Intelligence in Education*, A. M. Olney et al., Eds., Cham: Springer Nature Switzerland, 2024, pp. 31-46. https://doi.org/10.1007/978-3-031-64299-9_3

[37] E. Costello, J. C. Holland, and C. Kirwan, "Evaluation of MCQs from MOOCs for common item writing flaws," BMC Res. Notes, vol. 11, Art. 849, Dec. 2018. [Online]. Available: https://doi.org/10.1186/s13104-018-3959-4. https://doi.org/10.1186/s13104-018-3959-4

[38] R. Meissner et al., "LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation," Front. Educ., vol. 9, Art. 1427502, Oct. 2024. https://doi.org/10.3389/feduc.2024.1427502

[39] S.-V. Oprea, I. Nica, A. Bâra, and I.-A. Georgescu, "Are skepticism and moderation dominating attitudes toward AI-based technologies?" Am. J. Econ. Sociol., vol. 83, no. 3, pp. 567-607, 2024. https://doi.org/10.1111/ajes.12565

[40] A. Kulal, H. U. Rahiman, H. Suvarna, A. Abhishek, and S. Dinesh, "Enhancing public service delivery efficiency: Exploring the impact of AI," J. Open Innov. Technol. Mark. Complex., vol. 10, no. 3, Art. 100329, 2024. https://doi.org/10.1016/j.joitmc.2024.100329

[41] S. K. Amemasor et al., "A systematic review on the impact of teacher professional development on digital instructional integration and teaching practices," Front. Educ., vol. 10, Art. 1541031, 2025. https://doi.org/10.3389/feduc.2025.1541031

[42] M. Diliberti et al., *Using Artificial Intelligence Tools in K-12 Classrooms*, RAND, 2024.

[43] S. Triberti, R. Di Fuccio, C. Scuotto, E. Marsico, and P. Limone, ""Better than my professor?" How to develop AI tools for higher education," Front. Artif. Intell., vol. 7, Art. 1329605, Apr. 2024. https://doi.org/10.3389/frai.2024.1329605

[44] G. W. Soad, N. F. Duarte Filho, and E. F. Barbosa, "Quality evaluation of mobile learning applications," in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1-8. https://doi.org/10.1109/FIE.2016.7757540

[45] X. Zeng, X. Peng, and C. Lu, "Survey on the quality assessment factors of educational APP," in *2017 Int. Symp. on Educational Technology (ISET)*, 2017, pp. 196-200. https://doi.org/10.1109/ISET.2017.52

[46] R. Topal and F. Shargh, "Teaching students how to find and identify reliable online sources: A series of exercises," J. Polit. Sci. Educ., vol. 19, no. 3, pp. 475-484, 2023. https://doi.org/10.1080/15512169.2022.2163899

[47] E. Mazzullo, O. Bulut, T. Wongvorachan, and B. Tan, "Learning Analytics in the Era of Large Language Models," Analytics, vol. 2, no. 4, pp. 877-898, Nov. 2023. https://doi.org/10.3390/analytics2040046

[48] B. A. Schwendimann et al., "Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research," IEEE Trans. Learn. Technol., vol. 10, no. 1, pp. 30-41, 2017. https://doi.org/10.1109/TLT.2016.2599522

[49] Z. Wang, J. Valdez, D. B. Mallick, and R. G. Baraniuk, "Towards Human-Like Educational Question Generation with Large Language Models," in *Artificial Intelligence in Education*, M. M. Rodrigo et al., Eds., Cham: Springer Int'l Publishing, 2022, pp. 153-166. https://doi.org/10.1007/978-3-031-11644-5_13

[50] M. C. Fink, S. A. Robinson, and B. Ertl, "AI-based avatars are changing the way we learn and teach: benefits and challenges," Front. Educ., vol. 9, Art. 1416307, 2024. https://doi.org/10.3389/feduc.2024.1416307

[51] A. Kokala, "Revolutionizing Content Creation: Leveraging AI-Driven Podcast Generation with NotebookLM and Personalized Insights," Int. Res. J. Mod. Eng. Technol. Sci., vol. 6, 2024.

[52] S. Wang, Z. Ning, A. Truong, M. Dontcheva, D. Li, and L. B. Chilton, "PodReels: Human-AI Co-Creation of Video Podcast Teasers," in Proc. 2024 ACM Designing Interactive Systems Conf., 2024, pp. 958-974. https://doi.org/10.1145/3643834.3661591

[53] P. Kannan and D. Zapata-Rivera, "Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: informing the design and development of learning analytic dashboards," in *Frontiers in Education*, vol. 7, 2022, Art. 913594. https://doi.org/10.3389/feduc.2022.913594

[54] R. Bodily, J. Kay, V. Aleven, I. Jivet, D. Davis, F. Xhakaj, and K. Verbert, "Open learner models and learning analytics dashboards: a systematic

review," in *Proc. 8th Int. Conf. on Learning Analytics and Knowledge*, 2018, pp. 41-50. https://doi.org/10.1145/3170358.3170409

[55] I. Molenaar and C. A. N. Knoop-van Campen, "How teachers make dashboard information actionable," IEEE Trans. Learn. Technol., vol. 12, no. 3, pp. 347-355, 2018. https://doi.org/10.1109/TLT.2018.2851585

[56] M. Sahin and D. Ifenthaler, "Examining of Learners' Dashboard Interaction in Computer Classification Testing Environment," in *2022 Int. Conf. on Advanced Learning Technologies (ICALT)*, 2022, pp. 152-154. https://doi.org/10.1109/ICALT55010.2022.00052

[57] N. Valle, P. Antonenko, K. Dawson, and A. C. Huggins-Manley, "Staying on target: A systematic literature review on learner-facing learning analytics dashboards," Br. J. Educ. Technol., vol. 52, no. 4, pp. 1724-1748, 2021. https://doi.org/10.1111/bjet.13089

[58] B. Gras, A. Brun, and A. Boyer, "For and by student dashboards design to address dropout," in Companion Proc. 10th Int. Conf. Learn. Analytics & Knowledge, Workshop on Addressing Drop-Out Rates in Higher Education (ADORE '20), 2020.

[59] E. M. Queiroga et al., "A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course," Appl. Sci., vol. 10, no. 11, Art. 3998, 2020. https://doi.org/10.3390/app10113998

[60] J. Smink and F. P. Schargel, *Helping Students Graduate: A Strategic Approach to Dropout Prevention*, Eye on Education, 2004.

[61] G. Akçapınar and M. N. Hasnine, "Discovering the effects of learning analytics dashboard on students' behavioral patterns using differential sequence mining," Procedia Comput. Sci., vol. 207, pp. 3818-3825, 2022. https://doi.org/10.1016/j.procs.2022.09.443

[62] C. McGrath, A. Farazouli, and T. Cerratto-Pargaman, "Generative AI chatbots in higher education: a review of a merging research area," Higher Education, 2024, pp. 1-17. https://doi.org/10.1007/s10734-024-01288-w

[63] D. Ravšelj et al., "Higher education students' perceptions of ChatGPT: A global study of early reactions," PLOS ONE, vol. 20, no. 2, Art. e0315011, Feb. 2025.

[64] S. D'Urso, F. Sciarrone, and M. Temperini, "Boulez: a chatbot-based federated learning system for distance learning," in Proc. Int. Conf. Information Visualisation (IV), IEEE, 2023, pp. 210-215. https://doi.org/10.1109/IV60283.2023.00045

[65] V. De Marco, F. Sciarrone, and M. Temperini, "TutorChat: a chatbot for the support to dyslexic learner's activity through Generative AI," in Proc. Int. Conf. Advanced Learning Technologies (ICALT), IEEE, 2024, pp. 155-157. https://doi.org/10.1109/ICALT61570.2024.00051

[66] D. Sun, A. Boudouaia, C. Zhu, and Y. Li, "Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study," Int. J. Educ. Technol. High. Educ., vol. 21, Art. 14, Feb. 2024. https://doi.org/10.1186/s41239-024-00446-5

[67] H. Güner and E. Er, "AI in the classroom: Exploring students' interaction with ChatGPT in programming learning," Educ. Inf. Technol., vol. 1, 2025, doi:10.1007/s10639-025-13337-7. https://doi.org/10.1007/s10639-025-13337-7

[68] M. Khalifa and M. Albadawy, "Using artificial intelligence in academic writing and research: An essential productivity tool," Comput. Methods Programs Biomed. Update, vol. 5, Art. 100145, Jan. 2024, doi:10.1016/j.cmpbup.2024.100145. https://doi.org/10.1016/j.cmpbup.2024.100145

[69] E. Cano-Pleite et al., "Exploring the Integration of an AI-Assisted Midterm Exam in a Thermal Engines Course," in Proc. 17th Annu. Int. Conf. Educ. Res. Innovation (ICERI2024), Seville, Spain, IATED, 2024, pp. 6877-6881. https://doi.org/10.21125/iceri.2024.1659

[70] G. Kortemeyer and J. Nöhl, "Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study," Phys. Rev. Phys. Educ. Res., vol. 21, no. 1, p. 010136, 2025. https://doi.org/10.1103/PhysRevPhysEducRes.21.010136

[71] S. Dhara, S. Chatterjee, R. Chaudhuri, A. Goswami, and S. K. Ghosh, "Artificial intelligence in assessment of students' performance," in *Artificial Intelligence in Higher Education*, CRC Press, 2022, pp. 153-167. https://doi.org/10.1201/9781003184157-8

[72] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," arXiv:2306.05685, 2023.

[73] J. Floden, "A comparison between human and AI grading of exams in higher education using ChatGPT," Br. Educ. Res. J., vol. 51, pp. 201-224, 2025. https://doi.org/10.1002/berj.4069

[74] W. Hersh and K. F. Hollis, "Results and implications for generative AI in a large introductory biomedical and health informatics course," npj Digital Medicine, vol. 7, Art. 247, 2024. https://doi.org/10.1038/s41746-024-01251-0

[75] A. Pack, A. Barret, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: insights into validity and reliability," Comput. Educ. Artif. Intell., vol. 6, pp. 1-9, 2024. https://doi.org/10.1016/j.caeai.2024.100234

[76] M. Stahi, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring LLM prompting strategies for joint essay scoring and feedback generation," in Proc. Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2024, pp. 1-16.

[77] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," Humanities Soc. Sci. Commun., vol. 11, Art. 723, 2024. https://doi.org/10.1057/s41599-024-03209-9

[78] A. Nitze, "Future-proofing Education: a prototype for simulating oral examinations using large language models," in Proc. STEMEIC '23 Conf., 2023, pp. 1-6.

[79] S. I. Salim, R. Y. Yang, A. Cooper, S. Ray, S. K. Debray, and S. Rahaman, "Impeding LLM-assisted Cheating in Introductory Programming Assignments via Adversarial Perturbation," in Proc. EMNLP 2024, 2024, pp. 445-463. https://doi.org/10.18653/v1/2024.emnlp-main.27

[80] M. El-Muwalla and A. Badran, "Turnitin: Building Academic Integrity Against Plagiarism to Underpin Innovation," in *Higher Education in the Arab World*, Springer, 2020, pp. 261-268. https://doi.org/10.1007/978-3-030-37834-9_11

[81] R. Crockett and K. Best, "Stylometric Comparison of Professionally Ghost-Written and Student-Written Assignments," in *Integrity in Education for Future Happiness*, European Network for Academic Integrity, 2020, pp. 35-49. https://doi.org/10.11118/978-80-7509-772-9-0035

[82] M. Psyridou, F. Prezja, M. Torppa, M.-K. Lerkkanen, A.-M. Poikkeus, and K. Vasalampi, "Machine Learning Predicts Upper Secondary Education Dropout as Early as the End of Primary School," Sci. Rep., vol. 14, Art. 1-12, 2024. https://doi.org/10.1038/s41598-024-63629-0

[83] N. U. R. Junejo, M. W. Nawaz, Q. Huang, X. Dong, C. Wang, and G. Zheng, "Accurate Multi-Category Student Performance Forecasting at Early Stages of Online Education Using Neural Networks," arXiv:2412.05938, 2024. https://doi.org/10.1038/s41598-025-00256-3

[84] A. Smolansky, A. Cram, C. Raduescu, S. Zeivots, E. Huber, and R. F. Kizilcec, "Educator and student perspectives on the impact of generative AI on assessments in higher education," in *Proc. 10th ACM Conf. Learn. @ Scale (L@S '23)*, Copenhagen, Denmark, 2023, pp. 378-382. https://doi.org/10.1145/3573051.3596191

[85] T. H. Bers, "Assessing Critical Thinking in Community Colleges," New Dir. Community Colleges, vol. 2005, pp. 15-25, 2005. https://doi.org/10.1002/cc.192

[86] E. Mazzullo and O. Bulut, "Automated Feedback Generation for Open-Ended Questions: Insights from Fine-Tuned LLMs," in Proc. FM-EduAssess at NeurIPS 2024 Workshop, PMLR, Edmonton, AB, Canada, 2024, pp. 1-18.

[87] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, and E. Gehringer, "Automated Feedback Generation for Student Project Reports: A Data-Driven Approach," in Proc. 54th ACM Tech. Symp. Comput. Sci. Educ. (SIGCSE 2023), Toronto, ON, 2023, pp. 916-922.

[88] S. C. E. Fung, M. F. Wong, and C. W. Tan, "Automatic Feedback Generation on K-12 Students' Data Science Education by Prompting Cloud-Based LLMs," in Proc. 18th Int. Conf. Educ. Data Mining (EDM 2024), 2024. https://doi.org/10.1145/3657604.3664673

[89] T. Ruwe and E. M.-Paus, "Embracing LLM feedback: the role of feedback providers and provider information for feedback effectiveness," Front. Educ., vol. 9, Art. 1461362, 2024. https://doi.org/10.3389/feduc.2024.1461362

[90] A. Kinder, F. J. Briese, M. Jacobs, N. Dern, N. Glodny, S. Jacobs, and S. Leßmann, "Effects of adaptive feedback generated by a large language model: A case study in teacher education," Comput. Educ. Artif. Intell., vol. 8, Art. 100349, 2025. https://doi.org/10.1016/j.caeai.2024.100349

[91] S. Riazi and P. Rooshenas, "LLM-Driven feedback for enhancing conceptual design learning in database systems courses," in *Proc. SICSE TS*, 2025, pp. 1-7. https://doi.org/10.1145/3641554.3701940

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 681–686
DOI: 10.15439/2025F4081
ISSN 2300-5963 ACSIS, Vol. 43

Towards the automated classification of German job titles according to KldB

Ralf Dorau*, Kristine Hein*
* Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
Email: dorau@bibb.de, kristine.hein@bibb.de

***Abstract*—The automated classification of occupational titles is a pivotal component of labor market analysis, survey research, and administrative data processing. This paper explores the viability of mapping German job titles to the German Classification of Occupations (KldB) by employing conventional machine learning methodologies to examine the challenges and limitations inherent in the data itself. To this end, the present study leverages two complementary datasets—manually annotated survey data and a dataset of occupational synonyms—to assess the performance of established classifiers under varying levels of taxonomic granularity. The methodological challenges inherent to this study include class imbalance, semantic ambiguity, and linguistic variability, which are all characteristics of German job title expressions. The findings of the study suggest that while coarse-level classifications can be addressed with relatively simple models and text representations, finer-grained distinctions remain challenging to resolve using title-based features alone. The findings indicate that more expressive models and richer contextual information may be necessary for high-resolution occupational coding.**

I. Introduction

THE automated classification of job titles is an important topic in research, but also for practitioners in the field of labor market research. For example, in questionars people may answer the question about their occupation. However, this needs to be matched to a certain occupation, for example the German Classification of Occupations (Kldb) or the International Standard Classification of Occupations (ISCO). Other tasks include the classification of online job advertisements (OJAs) or the matching of other occupations, e.g., from online platforms like Kununu [1]. Other researchers try to tackle the question of identifiying occupations in textual data, for example in parliamentary debates [2]. The German language has several challenges, as occupational titles are not only single nouns but may also be a combination of nouns and may also contain additional data like the examining institution (e.g., 'IHK'). Next, different forms for male, female, gendered or neutral titles may exist. For generic texts, another challenge is that surnames may originate from (partly historical) professions, for example 'Bäcker' (baker).

Linking a variety of different sources on labor market is usually considered a very challenging task [3], however, in this paper, we limit the question to the mapping and automated classification of job titles in German language. Here, we find dictionary-based approaches which are widely used, but also other ML-based approaches. Training data was widely collected from survey data or job-titles from KldB, ESCO or other synonym data.

The main contribution of our paper is a large collection of novel training data and a systematical analysis of challenges to train classificators for German job titles according to KldB and in particular different levels of occupations and their level of performance.

This paper is divided into five sections. The first section provides an introduction, the second section gives the the state of the art, and related work. The third section describes the data and methodological background. The fourth section is dedicated to experimental results and evaluation. Our conclusions are given in the last section.

II. Related Work

A multitude of classification categories are recognized for occupations. The International Standard Classification of Occupations (ISCO) was developed by the International Labor Organization (ILO) and published in 1958, 1968, 1988, and most recently in 2008)[1]. The ISCO 2008 has also been utilized within the European Union (EU), with certain German-speaking countries (Germany, Austria, and Switzerland) developing a customized version of the classification. The International Standard Classification of Occupations (ISCO) is structured at a skill level and linked to the "European Skills, Competences, Qualifications and Occupations" (ESCO) ontology, which adds another hierarchy level to the data. In Germany, the Classification of Occupations (KldB) serves as the reference classification for the Federal Employment Agency (BA) and its research institute (IAB)[2]. In this organization, occupations are structured at a task level. The most recent version is the 2020 revision of the KldB 2010, which has undergone a comprehensive redesign, thereby rendering the previous versions from 1988 and 1992 obsolete. The development of this system was undertaken with the objective of ensuring compatibility with the ISCO-08 standard. The study of job titles and taxonomies has a long history, extending even before the advent of computer technology [4].

A portion of the research has focused on the classification of OJAs according to the O*NET framework [5]. This has included the application of normalization approaches [6] and

[1] See https://www.ilo.org/public/english/bureau/stat/isco/isco08/.
[2] See https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/KldB2010-Fassung2020-Nav.html.

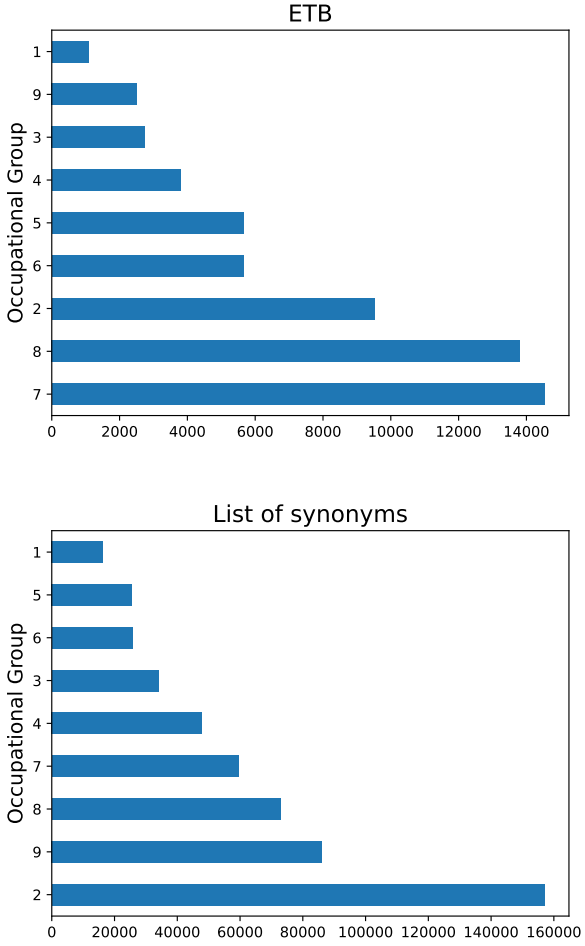


Fig. 1: Distribution of training data according to occupational groups

similarity-based methods [7]. The classification of job titles is also employed in the context of online job recruitment [8].

A limited number of publications have been published on the subject of German job titles, with a particular emphasis on the German KldB. For instance, a technical report based on OJAs [9] with challenges on level 4, but with promising results on level 1. Malte Schierholz's 2018 publication [10] introduced the concept of auxiliary classifications in the field of occupational coding. For further research on the subject of occupational coding in surveys, we refer to [11]. A master's thesis endeavors to predict KldB 5-digit job titles from survey data, thereby highlighting the persistent challenges associated with this endeavor, see [12]. In a similar vein, a scholarly article was published that compared the classification of survey data using BERT and GPT-3, see [13]. However, the absence of a standardized reporting methodology precludes the direct comparison of their results. Nevertheless, they evince analogous challenges to those observed in other studies. Consequently, both the classification of occupational areas and, in particular, the level of performance (5th digit) persist as arduous tasks.

III. Data and Methods

A. Data

The dataset aggregates 60,022 manually annotated occupational titles from the 2012, 2018, and forthcoming 2024 waves of the Erwerbstätigenbefragung (ETB; German Employment Survey)[3], each coded to KldB 2010. These representative surveys – conducted by BIBB and BAuA – target Germany's core workforce (employed individuals aged 15+ working $\geq$10 hours/week). While historical predecessor surveys (1979-1999 BIBB/IAB studies[4]) exist, their integration was precluded by fundamental incompatibilities between KldB versions, despite partial crosswalk feasibility.

[3]Main datasets: 10.7803/501.12.1.1.60 (2012), 10.7803/501.18.1.1.10 (2018). Supplementary variables: 2012: 10.7803/501.12.1.4.10 (full-text), 10.7803/501.12.1.3.20 (regional identifiers), 10.7803/501.12.1.5.30 (special variables); 2018: 10.7803/501.18.1.4.10 (full-text), 10.7803/501.18.1.3.10 (regional identifiers), 10.7803/501.18.1.5.10 (special variables).

[4]GESIS SUF: 10.4232/1.1243, 10.4232/1.12563, 10.4232/1.2565, 10.4232/1.12247

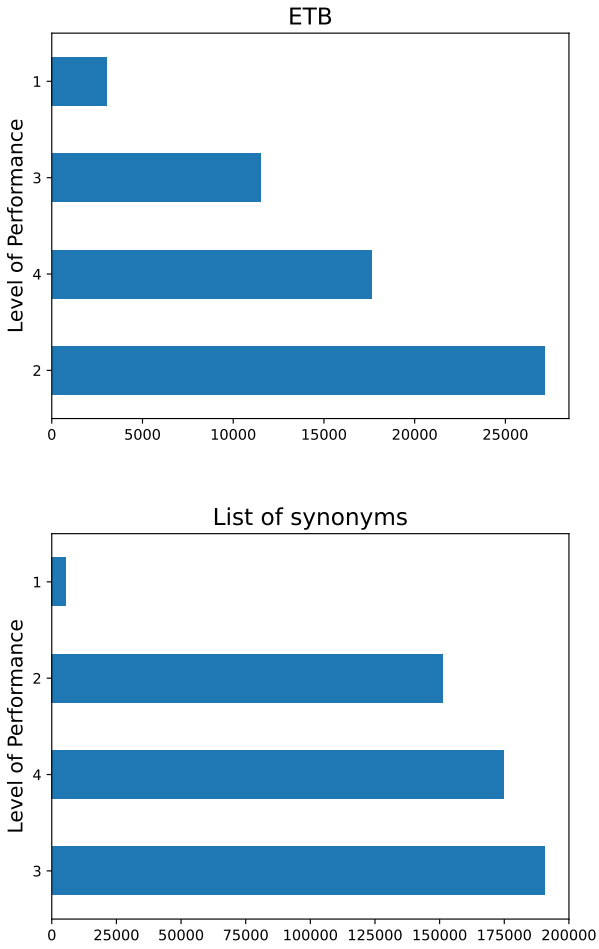


Fig. 2: Distribution of training data according to level of performance (5th-digit)

We use a second dataset comprising 526,535 synonyms and variants of male, female, and neutral job titles, provided by the German Federal Employment Agency (BA)[5]. Both datasets have their own biases, as can be seen in Figures 1 and 2: While in ETB dataset the majority of occupations are located in groups 7 and 8, they are in groups 2 and 9 for the list of synonyms. We find a similar bias with respect to the level of performance, coded in the 5th digit of KldB.

Data cleaning is an existentiell step. We remove occupational are 0 from the dataset as we omit military occupations. Second, we clean all data with errors, for example too large or small KldB numbers. In Dataset 1, 59,535 and in dataset 2, 522,197 entries remain.

B. Method

In this paper, we adopt the classification approach to labor market data presented in [14], [15]. The baseline model was trained using three standard classifiers with default values from the scikit-learn library [16] with logistic regression ($C = 10$), naive Bayes, and random forest. We split the data data so that 20% was used as test data and 80% as training data. As the main question is about specific challenges, we will apply traditional methods to identify the route for further research in this area.

However, another open question is the choice of text representation. As shown by this can significantly change the performance of classifiers. We can use the combination of term frequency and inverse document frequency as TF.IDF (term frequency–inverse document frequency). Another approach is the Word to Vector (W2V) which applies unsupervised learning to represent textual data and produces rather low-dimensional data. As we have seen, the later produces rather poor results and we will only showcase the results for Logistic Regression.

[5]Available at https://www.arbeitsagentur.de/institutionen/dkz-downloadportal.

TABLE I: Results for 1-digit on dataset 1

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	0.88 / **0.86**	**0.79 / 0.84**	**0.83 / 0.84**
Logistic Regression (W2V)	0.45 / 0.47	0.26 / 0.46	0.26 / 0.38
Naive Bayes	**0.90** / 0.84	0.68 / 0.79	0.75 / 0.79
Random Forest	0.84 / **0.86**	0.76 / 0.79	0.78 / 0.81

TABLE II: Results for 1-digit on datasets 1+2

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	**0.92 / 0.91**	**0.89 / 0.91**	**0.90 / 0.91**
Logistic Regression (W2V)	0.61 / 0.57	0.35 / 0.51	0.38 / 0.47
Naive Bayes	0.91 / 0.89	0.84 / 0.88	0.87 / 0.88
Random Forest	0.91 / **0.91**	0.88 / 0.90	0.89 / 0.90

IV. Experimental Results

A. Occupational areas (1-digit)

Tables I and I present the outcomes for occupational categories with a one-digit classification, contrasting the utilization of Dataset 1 exclusively or Datasets 1 and 2 collectively. The results of this study are largely consistent with those reported in the extant literature. As previously discussed, the W2V model demonstrated substandard performance. A comparative analysis reveals that, while all other models demonstrate comparable performance, Logistic Regression consistently exhibits superior performance in various applications.

The findings of this study demonstrate that extending the dataset has a substantial impact on the quality of the results. For instance, the F_1 score increased from 0.84 to 0.91, a notable enhancement.

B. Occupational group (3-digit)

As illustrated in Tables III and III, the results for occupational group (3-digit) are presented. A direct comparison with occupational areas reveals that the performance is suboptimal. However, the employment of both datasets results in a substantial enhancement of the outcomes. However, a more thorough examination reveals that this approach is particularly ineffective for groups with limited data. A cursory examination of dataset 1 reveals that there is an absence of data points corresponding to area 114 (occupations in fishing), 213 (occupations in industrial glass-making and -processing), and 214 (occupations in industrial ceramic-making and -processing). For future research endeavors, there is a necessity for not only improved training data that is more balanced, but also for the evaluation of cross-validation approaches.

C. Level of Performance (5th digit)

The results of the classification of the level of performance (5th digit), as demonstrated in Tables V and VI, also exhibit a number of noteworthy outcomes. First, it is noteworthy that the extended dataset does not yield substantial improvements in the results. Secondly, we observe that different levels yield different F_1 scores. For Logistic Regression, a score of 0.86 is observed for both level 2 and level 4, while the level 1 (0.73) and level 3 (0.71) demonstrate inferior performance. A similar outcome is observed for all other methods, although level 1 and 3 demonstrate an even poorer performance. Therefore, it appears that the fifth digit is a significant factor in the classification of job titles. The available data from a job title alone is insufficient for effective classification.

V. Discussion and Outlook

The classification of German job titles according to the KldB taxonomy remains a complex task, particularly at the more granular levels such as the 3-digit occupational group and the 5th-digit level of performance. The findings of this study underscore the strengths and weaknesses of conventional machine learning methodologies in this context. While the utilization of TF-IDF features in conjunction with Logistic Regression yielded robust outcomes, particularly for the 1-digit classification, performance exhibited a substantial decline for finer-grained levels. This phenomenon underscores the inherent intricacy of occupational data, wherein minor linguistic and contextual variations can wield substantial consequences for classification.

The present study demonstrates the significance of high-quality, balanced training data. The incorporation of the synonym dataset led to a substantial enhancement in performance at more extensive classification levels. However, this augmentation was accompanied by the introduction of biases stemming from the unequal distribution across occupational categories. This phenomenon is especially evident among underrepresented categories. Furthermore, our analysis indicates that classical models rapidly reach their limits when tasked with inferring the level of performance from job titles alone. In this context, the available information frequently lacks the requisite depth to accurately differentiate between skill levels. This finding underscores the necessity for more comprehensive or even external input data, such as job descriptions, qualifications, or contextual metadata, to capture the subtleties that may be obscured by simplistic job titles.

TABLE III: Results for 3-digit on dataset 1

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	**0.81** / 0.80	**0.60 / 0.74**	**0.67 / 0.75**
Logistic Regression (W2v)	0.19 / 0.36	0.09 / 0.28	0.10 / 0.24
Naive Bayes	0.49 / 0.74	0.28 / 0.60	0.33 / 0.61
Random Forest	0.79 / **0.81**	0.57 / 0.71	0.64 / 0.74

TABLE IV: Results for 3-digit on datasets 1+2

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	**0.87 / 0.85**	**0.80 / 0.83**	**0.83 / 0.84**
Logistic Regression (W2v)	0.49 / 0.44	0.25 / 0.35	0.29 / 0.33
Naive Bayes	0.85 / 0.80	0.55 / 0.70	0.63 / 0.71
Random Forest	0.86 / **0.85**	0.79 / 0.82	0.82 / 0.83

TABLE V: Results for Level of Performance (5th digit) on dataset 1

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	0.83 / **0.83**	**0.76 / 0.83**	**0.79 / 0.83**
Logistic Regression (W2v)	0.60 / 0.52	0.35 / 0.53	0.33 / 0.46
Naive Bayes	**0.87 / 0.83**	0.68 / 0.81	0.73 / 0.80
Random Forest	0.80 / 0.82	0.75 / 0.80	0.77 / 0.80

TABLE VI: Results for Level of Performance (5th digit) on datasets 1+1

Approach	Metrics (Macro/Weighted)		
	Precision	Recall	F_1 Score
Logistic Regression (Tf-idf)	**0.85 / 0.85**	**0.78 / 0.85**	**0.81 / 0.85**
Logistic Regression (W2v)	0.71 / 0.65	0.50 / 0.63	0.50 / 0.62
Naive Bayes	**0.85** / 0.82	0.67 / 0.81	0.71 / 0.81
Random Forest	0.84 / 0.84	0.77 / **0.85**	**0.81 / 0.85**

In terms of future research, several avenues emerge as potentially fruitful directions. Initially, the integration of extensive and heterogeneous datasets, encompassing historical KldB mappings or annotated survey data, could assist in mitigating class imbalance and enhancing generalizability. Secondly, the implementation of advanced language models, such as BERT, domain-specific transformers, or large language models (LLMs), has the potential to substantially improve classification accuracy, especially when fine-tuned on labor market data. Finally, methodological improvements, such as stratified cross-validation or ensemble learning approaches, have the potential to offer more robust evaluation and enhanced performance. While the present approach is both effective and stable, these insights serve as a foundation for more sophisticated future work in the automatic classification of job titles.

REFERENCES

[1] K. Hein and J. Dörpinghaus, "What is said about vet on social media in germany? trends, demands, and opinions." in *NORDYRK BOOK OF ABSTRACTS*, 2024, p. 109.

[2] J. Binnewitt, "Recognising occupational titles in german parliamentary debates," in *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, 2024, pp. 221–230.

[3] A. Fischer and J. Dörpinghaus, "Web mining of online resources for german labor market research and education: Finding the ground truth?" *Knowledge*, vol. 4, no. 1, pp. 51–67, 2024.

[4] N. R. Council, D. of Behavioral, S. Sciences, C. on Occupational Classification, and Analysis, "Work, jobs, and occupations: A critical review of the dictionary of occupational titles," 1980.

[5] F. Javed, M. McNair, F. Jacob, and M. Zhao, "Towards a job title classification system," *arXiv preprint arXiv:1606.00917*, 2016.

[6] Y. Zhu, F. Javed, and O. Ozturk, "Document embedding strategies for job title classification." in *FLAIRS*, 2017, pp. 221–226.

[7] I. Rahhal, K. M. Carley, I. Kassou, and M. Ghogho, "Two stage job title identification system for online job advertisements," *IEEE Access*, vol. 11, pp. 19 073–19 092, 2023.

[8] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, and T. S. Kang, "Carotene: A job title classification system for the online recruitment domain," in *2015 IEEE First International Conference on Big Data Computing Service and Applications*. IEEE, 2015, pp. 286–293.

[9] R. Baskaran and J. Müller, "Classification of german job titles in online job postings using the kldb-2010 taxonomy," 2023.

[10] M. Schierholz, "An auxiliary classification with work activity descriptions for occupation coding," *AStA Wirtschafts-und Sozialstatistisches Archiv*, vol. 12, pp. 285–298, 2018.

[11] A. Müller, "The implementation of the german classification of occupations 2010 in the iab job vacancy survey: documentation of the implementation process," IAB-Forschungsbericht, Tech. Rep., 2014.

[12] V. P. V. Karanam, "Occupation coding using a pretrained language model by integrating domain knowledge," 2022.

[13] P. Safikhani, H. Avetisyan, D. Föste-Eggers, and D. Broneske, "Automated occupation coding with hierarchical features: A data-centric

approach to classification with pre-trained language models," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 6, 2023.

[14] R. Fechner, J. Dörpinghaus, and A. Firll, "Classifying industrial sectors from german textual data with a domain adapted transformer," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023, pp. 463–470.

[15] R. Fechner and J. Dörpinghaus, "No train, no pain? assessing the ability of llms for text classification with no finetuning," in *Proceedings of the Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia*, 2024, pp. 8–11.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 687–692
DOI: 10.15439/2025F2839
ISSN 2300-5963 ACSIS, Vol. 43

Do LLMs dream of antique hermeneutics? Critical remarks on automated text interpretation

Jens Dörpinghaus*†‡ , Michael Tiemann*†
* Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
† University of Koblenz, Germany
‡ University of Pretoria,
Email: jens.doerpinghaus@bibb.de, tiemann@bibb.de

***Abstract*—This study investigates the potential of large language models (LLMs) to apply hermeneutical methods rooted in philosophy, theology, sociology and literary studies in a meaningful manner. Utilising a comparative experimental design, four LLMs were prompted to interpret a variety of texts, encompassing religious, philosophical, poetic, and conversational material. The findings indicate considerable variability, an absence of reproducibility, and a substantial reliance on prompt design, model type, and language. This suggests that LLMs do not employ coherent hermeneutical strategies. Despite the automation of formal features, limitations in context sensitivity, interpretive intentionality, and epistemic grounding render LLMs ill-suited to authentic hermeneutics. The study concludes that current LLMs are incapable of replicating the depth of human interpretive practice, and calls for further interdisciplinary research to define evaluation standards for machine-assisted exegesis.**

I. Introduction

IT IS evident that all perception and self-perception are founded on acts of interpretation, extending beyond the mere reception of auditory and written statements. The term 'hermeneutics' generally encompasses theories that address the possibility, necessity, conditions, challenges, and objectives of processes involving understanding and interpretation. Thus, it can also be seen as a 'theory of understanding' or 'theory of interpretation', see [1]. Despite the numerous endeavours to define hermeneutics, a comprehensive consensus remains elusive. Maier has highlighted two predominant tendencies in the understanding and definition of hermeneutics, see [2]. Firstly, a rather mechanical tendency, and secondly, a definition of understanding and comprehension that is more common in German-speaking countries. In theology, this is an essential basis for exegesis, i.e. the interpretation of texts.

In antiquity and the Middle Ages, hermeneutics served as the science and art of interpreting fundamental texts, especially the Bible and laws. In modern times, its field of application has expanded, developing into a general doctrine of the prerequisites and methods of proper interpretation and a philosophy of understanding, see [3], [2]. Hermeneutics are also widely discussed and applied in areas like study of literature, psychology, social sciences, history and medical research.

It is important to note that hermeneutical concerns are ubiquitous. In the analysis of data (exegesis), a certain form of data hermeneutics is essential, and communication is pervasive. Over the past several years, a new generation of generative AI tools has emerged. In Section II, we will demonstrate that researchers are concerned about which hermeneutics should be applied to AI-generated texts, or which tasks of hermeneutics can be carried out by LLMs. However, there is a paucity of research in this area, with only a small number of studies examining the hermeneutical capabilities of LLMs and generative AI in general.

The central research question guiding our study is as follows: Can contemporary LLMs apply hermeneutical approaches from Philosophy, Sociology, study of literature and Theology to textual data? Our research is particularly interested in methodological and technical reflections, while also raising novel questions that will serve as a foundation for a more robust theoretical framework in this field.

II. Literature Review

Some researchers argue to use hermeneutics on AI-generated texts [4], others propose hermeneutics as tool to optimize prompts [5]. However, there is a clear lack of hermeneutics for AI-generated texts as recognized in research [6], [7], [8]. While some researchers are generally positive on applying existing theory, e.g., by Heidegger or Lacan, to LLMs, see for example [9], there is no consent on this issue [10]. In general, researchers from non-technical domains seem to overestimate the features and capabilities of LLMs and AI in general, see for example [11]. Other researchers highlight the need for domain knowledge, for example in Bible teaching in confessional religious education using LLMs [12]. However, as our main research question is mainly concerned with whether LLMs can perform hermeneutical tasks, we will now turn to literature on this.

According to our best knowledge, there is only one preprint paper addressing this question directly [13] and being critical about hermeneutical and understanding capacities of LLms. Other research focuses more on technical limitations or the correct place of LLMs, for example by seeing LLMs as "AI-interns" as [14] proposes. Another paper by [15] focuses on the theological and ethical bias of LLMs for interpreting texts from the Hebrew Bible and finds a bias which aligns with "contemporary socio-political and environmental concerns". Some scholars work on hermeneutics within the medical domain. For example, Hans-Georg Gadamer [16] argues that

LLMs cannot take over patient interaction as it cannot take over moral practices. In general, we see the aspect of human interaction and experience as one of the key challenges in using LLMs for hermeneutical aspects. Although the aspect of machine relationships complementing human-human interactions was suggested [17] this does not solve any hermeneutical questions.

III. Methodological Approach

We will now discuss several methodological approaches from different domains. They all share a common foundation. For example, Körtner describes the reader as a central aspect of Biblical hermeneutics [18]. However, he also points at the connection between narrative approaches from literature and Biblical exegesis. Oeming describes some confusion about this approach and its naming. In general, we find this method as "New (Literary) Criticism, Literary Approach/Criticism, Holistic Approach, Rhetorical Criticism, Close Reading, Synchronical Approach, Narratology und Narrative Criticism" [19]. This approach tries to determine how biblical texts can be interpreted as literature, see [20]. In summary, "the literary critic looks at the text for what *it* says in itself by means of the patterning or shaping—the informing—of its content." [21] Thus, we will – after a philosophical approach – focus on a selection of theological approaches. This selection must be somewhat arbitrary, but we tried to bridge between the reformators Zwingli and Calvin and a rather contemporary approach, liberation theology, as a contextual theology. However, as there is a clear link to narrative analysis, we will also discuss narrative approaches [1].

There is also some discussion between hermeneutics and sciences. Wyss, for example, discussed about the experience of time and self [22]. Other scholars discuss the world of living ('Lebenswelt') as paradigm for hermeneutics in the world of qualitative and quantitative reserach [23]. Thus we include a hermeneutics from the domain of sociology, namely objective hermeneutics.

Objective Hermeneutics, sometimes also labelled structural hermeneutics, derives from cultural anthropology. [24] discusses its connections to theories by Berger and Luckmann, Mead, and Chomsky. The idea is that how entities see the world they live in is reflected in the speech. On an individual level this would relate to an expression of one persons perception of reality, its ordering and the connections between entities. Within sociological theory there are also broader levels of groups of persons or organisations (meso level) up to the level of states (macro level). The structures that can be found by applying objective hermeneutics can be communicated intersubjectively and are answers to objective problems of action [24, p. 12]. For every structure there are numerous situations that pose problems, or crises. In fact, every single action is at first a situation of crises. The structure devises some solution and tests it. The way to re-construct structures within objective hermeneutics is thus interesting in two ways in view of its application to LLMs. The first interesting aspect is a "layer of normality" ("Normalitätsfolie", [25, p. 136]). This is one of the starting points for reconstructing a structure and refers to ideal types of normal behaviour. The second interesting aspect refers to the logic of cognition. In objective hermeneutics researchers explicitely use abductive reasoning. This can be seen as an inference to the best explanation while at the same time only having assumptions about the laws and rules from which to infer. This is seen as a creative process and might as such be particularly challenging for LLMs. For the inclusion of objective hermeneutics in this experiment this leads to some assumptions: If the LLMs are capable of applying an objective hermeneutics approach, they should all come to the same conclusion. Even though generally structures are reconstructed, these all refer to specific situations that find expressions in text. Thus, characters, places, times, etc. of these situations can also be reconstructed and should be comparable between different analysts.

The interest in narrative and 'story' began with Richard Niebuhr in 1941, see [26]. After that, it was introduced in the 1970s[1]. While some scholars focused on 'text-oriented approaches', 'reader-oriented approaches' were also widely used. What they all have in common, however, is the usage of methods of criticism studies for the interpretation of narrative texts [29]. Different schools exists which are not only committed to different literary theories, but are also connected to different geographical areas, see [19] and in particular [30]. In narrative exegesis, the first step is to apply methods of text analysis: "The 'what' of a text (its content) and the 'how' of a text (its rhetoric and structure) are analyzed as a complete tapestry, an organic whole." [20] For this purpose, the text is examined with respect to its content layer and effect layer to find out which structure, order and content the text wants to convey. In order to refine the exegesis, different elements of the text are examined: time, locations, characters, perspective, plot or plot line, narrator, listeners, and readers, key words, repetitions and other structural elements and stylistic devices.

Narrative approaches can be applied to any narrative texts – religious texts, for example Biblical texts, are widely considered to be suitable. Bailey et al. argue: "A Gospel is a narrative, fashioned out of selected traditions, that focuses on the activity and speech of Jesus as a way to reveal his character and develops a dramatic plot that culminates in the stories of his passion and resurrection." [31] Other scholars emphasize the need to work with historical-critical methods, see for example [32] and [33]. Narrative approaches are also widely applied to biblical and theological texts [34], [35], [36], so the quintessential next step is to also consider hermeneutics in Theology, for example liberation theology [37] or the Swiss reformers Ulrich Zwingli and John Calvin [38], [39], [40], [41].

Despite having elected to examine only a limited number of approaches, it has become evident that, upon thorough discussion of the various hermeneutical approaches that have been posited, it can be concluded that all of them exhibit features that cannot be directly substituted with concepts that are quantifiable. However, our study reveals two fundamental

[1]See also [27] and the overview by [28].

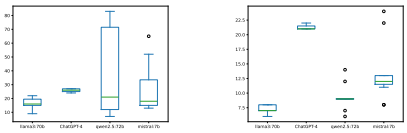


Fig. 1: Number of sections detected (direct) in Text 1 (left) and 2 (right)

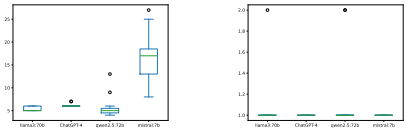


Fig. 2: Number of sections detected (direct) in Text 3 (left) and 4 (right)

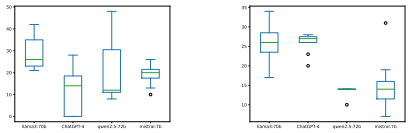


Fig. 3: Number of actors detected (direct) in Text 1 (left) and 2 (right)

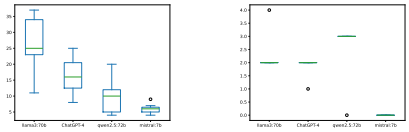


Fig. 4: Number of actors detected (direct) in Text 3 (left) and 4 (right)

aspects that all hermeneutics share: first, an understanding of human nature, and second, an understanding of language. While the question of whether AI can comprehend human nature remains a subject of ongoing discourse, the focus will be on the understanding of language. Danner summarizes: Hermeneutics is the study of interpretation, with a focus on comprehending meaning. The concept or phenomenon that is being understood is inherently present within the linguistic medium or can be articulated through it. The fundamental question, therefore, is: What is the origin of language? The fundamental question, therefore, is: what does precede language? The fundamental nature of language is predicated on its capacity to articulate elements of reality, that is to say, to "speak" of a reality that is already "out there" in the worldLLMs are unable to substitute the following aspects due to their inability to comprehend the context of reader and writer in a broader sense: (a) The reader and his perception or the listeners (unless we define LLMs as reader) (b) The context of a reader (unless we define LLMS as reader and define its context) (c) The author or narrator (d) With respect to both: Creator and created. Or the influence of the Holy Spirit (per Definition, as Luther and Zwingli described it), or Christ as hermeneutical principle (Calvin). In summary, it is evident that LLMs can only be utilized for the automation of fundamental principles of language understanding. However, as evidenced by our study of various hermeneutical approaches, different hermeneutics emphasize distinct aspects of language. Nevertheless, it is possible to identify several formalized structures that are, or may be, studied by hermeneutics. These include: (a) Plot as a broader concept, e.g., by selecting sections or units of meaning (b) Characters, e.g., the directly or indirectly acting persons (c) Space, while no single definition can be given. For example, according to Lefebvre and Soja [42] space can be defined as follows: "(Social) space is a (social) product." [43]. They differ between the *firstspace*, which includes physical space, the *secondspace* as mental space, and the *thirdspace* as social space. This, in turn, redirects the focus back to the question of which social actor a LLM is. (d) The concept of perspectives or points of view is also crucial in this analysis. However, it is crucial to acknowledge that an individual's understanding of points of view is significantly influenced by the context of the reader. This reiterates the necessity for a comprehensive examination of the social actor in question.

Consequently, an initial approach to assess the capabilities of LLMs in applying hermeneutics entails a concentration on plot and characters, as this method facilitates the introduction of objectivity to the inquiry. It is imperative to establish the following assumptions: The assumption is made that a single LLM is capable of interpreting a text employing a specific hermeneutical approach. Conducting a repeated experiment should, under these conditions, yield a consistent result or at least a reasonably proximate outcome. Secondly, it is hypothesized that different models will yield analogous results. The number of actors in a text is a subject that merits discussion; however, especially in narrative texts, this number should be nearly fixed. Furthermore, if two models yield divergent units of meaning or sections, they are likely to employ different perspectives and, consequently, divergent hermeneutical methods.

Our approach is as follows: We will study four different texts from various backgrounds and repeatedly query a LLM ($n = 10$) to select sections or units of meaning in a text and to return the number of actors in a text. For a detailed exploration of different prompting strategies, we refer to Appendix A.

IV. Experimental Results

To that end, the present study will examine four different texts. Initially, we examine a renowned German poem entitled "An die Freude" ("Ode to Joy", 1785) by Friedrich Schiller. Subsequently, we examine a biblical text comprising three narratives (Acts 8:1-25), having both a narrative and theological dimensions. Finally, we turn to a section of the philosophical text by the German philosopher Immanuel Kant, titled "Beantwortung der Frage: Was ist Aufklärung?" (Answering the Question: What Is Enlightenment?, 1784). The final text is a concise sentence: "Guten Abend, meine Damen und Herren" (Good evening, ladies and gentlemen).

Reconstructing a structure from one sentence with objective hermeneutics is challenging, but entirely possible. The least

TABLE I: Details of detected sections with direct approach on Texts 1-4 with different languages (G=German, E=English)

L	Text	Model	Min	Q_1	Median	Q_3	Max
G	Text 1	llama3:70b	9.0	15.0	16.0	19.5	22.0
		ChatGPT-4	24.0	25.0	26.0	27.0	27.0
		qwen2.5:72b	7.0	12.0	21.0	71.5	83.0
		mistral:7b	13.0	15.0	18.0	33.0	65.0
G	Text 2	llama3:70b	6.0	7.0	7.0	8.0	8.0
		ChatGPT-4	21.0	21.0	21.0	21.5	22.0
		qwen2.5:72b	6.0	9.0	9.0	9.0	14.0
		mistral:7b	8.0	11.5	12.0	13.0	24.0
E	Text 2	llama3:70b	8.0	8.0	10.0	18.5	24.0
		ChatGPT-4	7.0	8.5	10.0	10.0	10.0
		qwen2.5:72b	5.0	5.0	6.0	6.0	6.0
		mistral:7b	18.0	18.0	18.0	18.5	19.0
G	Text 3	llama3:70b	5.0	5.0	5.0	6.0	6.0
		ChatGPT-4	6.0	6.0	6.0	6.0	7.0
		qwen2.5:72b	4.0	4.5	5.0	5.5	13.0
		mistral:7b	8.0	13.0	17.0	18.5	27.0
E	Text 3	llama3:70b	4.0	4.0	4.0	7.0	28.0
		ChatGPT-4	4.0	4.0	4.0	4.0	5.0
		qwen2.5:72b	4.0	4.0	4.0	5.0	5.0
		mistral:7b	17.0	17.0	17.0	20.0	20.0
G	Text 4	llama3:70b	1.0	1.0	1.0	1.0	2.0
		ChatGPT-4	1.0	1.0	1.0	1.0	1.0
		qwen2.5:72b	1.0	1.0	1.0	1.0	2.0
		mistral:7b	1.0	1.0	1.0	1.0	1.0

TABLE II: Details of detected sections with SE approach on Texts 2-3 with different languages (G=German, E=English)

L	Text	Model	Min	Q_1	Median	Q_3	Max
G	Text 2	llama3:70b	7.0	7.0	7.0	7.0	10.0
		ChatGPT-4	5.0	5.0	5.0	9.3	25.0
		qwen2.5:72b	4.0	4.3	6.0	10.0	23.0
		mistral:7b	5.0	11.0	12.0	13.0	19.0
E	Text 2	llama3:70b	5.0	5.0	6.0	6.0	6.0
		ChatGPT-4	5.0	5.0	5.0	5.0	6.0
		qwen2.5:72b	3.0	3.0	3.0	4.0	4.0
		mistral:7b	9.0	12.0	12.0	13.0	21.0
G	Text 3	llama3:70b	3.0	4.5	5.0	6.5	21.0
		ChatGPT-4	5.0	7.0	7.0	17.0	23.0
		qwen2.5:72b	3.0	4.0	4.0	6.0	14.0
		mistral:7b	6.0	9.0	12.0	15.5	23.0
E	Text 3	llama3:70b	3.0	7.0	8.0	9.0	52.0
		ChatGPT-4	3.0	4.0	4.0	4.0	5.0
		qwen2.5:72b	8.0	9.0	9.0	9.5	22.0
		mistral:7b	12.0	15.0	16.0	33.5	44.0

a model should arrive at is to reconstruct personen (one speaking, more than one being spoken to), possible places (a theater, a lecture room, a TV setting) and times (evening). There are at least two sequences within this sentence, on referring to people being adressed, the other to a time. There is a published example of such an analysis by Oevermann himself [44]. All texts can be found online and there are likely references to them being analysed with one or another of the included hermeneutics. Selecting LLMs to test is a another challenge. We followed [45] by choosing the most widely used ChatGPT-4 and llama3:70b which are widely compared and discussed, see e.g. [46], [47], [48]. In addition, we choose qwen2.5:72b as another large model and, to compare with a rather small model, mistral:7b.

A. Plot

As illustrated in Figures 1 and 2, the boxplots depict the range of detected sections within the text corpus. For a more detailed analysis, please refer to Table tab:sections. It is noteworthy that the ChatGPT 4 and llama3 models demonstrate the greatest consistency in their output. However, for text 2, which comprises three stories, the output is particularly

TABLE III: Details of found actors with direct approach on Texts 1-4 with different languages (G=German, E=English)

L	Text	Model	Min	Q_1	Median	Q_3	Max
G	Text 1	llama3:70b	21.0	23.0	26.0	35.0	42.0
		ChatGPT-4	0.0	0.0	14.0	18.5	28.0
		qwen2.5:72b	8.0	11.0	12.0	30.5	48.0
		mistral:7b	10.0	17.5	20.0	21.5	26.0
G	Text 2	llama3:70b	17.0	23.5	26.0	28.5	34.0
		ChatGPT-4	20.0	26.0	27.0	27.5	28.0
		qwen2.5:72b	10.0	14.0	14.0	14.0	14.0
		mistral:7b	7.0	11.5	14.0	16.0	31.0
E	Text 2	llama3:70b	14.0	15.0	17.0	23.0	26.0
		ChatGPT-4	9.0	10.5	13.0	18.0	21.0
		qwen2.5:72b	9.0	9.0	9.0	9.0	9.0
		mistral:7b	13.0	13.0	13.0	13.0	13.0
G	Text 3	llama3:70b	11.0	23.0	25.0	34.0	37.0
		ChatGPT-4	8.0	12.5	16.0	20.5	25.0
		qwen2.5:72b	4.0	5.0	10.0	12.0	20.0
		mistral:7b	4.0	5.0	6.0	6.5	9.0
E	Text 3	llama3:70b	2.0	3.5	5.0	5.5	7.0
		ChatGPT-4	7.0	9.0	9.0	10.5	12.0
		qwen2.5:72b	2.0	3.5	4.0	6.0	12.0
		mistral:7b	1.0	4.0	6.0	6.5	7.0
G	Text 4	llama3:70b	2.0	2.0	2.0	2.0	4.0
		ChatGPT-4	1.0	2.0	2.0	2.0	2.0
		qwen2.5:72b	0.0	3.0	3.0	3.0	3.0
		mistral:7b	0.0	0.0	0.0	0.0	0.0

problematic, though it should be noted that no model partitions every verse (25). This phenomenon is also evident in Text 4, which essentially comprises a single sentence, particularly in instances where two sections have been identified. In such instances, the LLM has demonstrated an inability to distinctly delineate the boundaries of the two sections. Consequently, the outcomes for the poem "An die Freude" and Kant's text appear somewhat arbitrary. The discrepancy in the analysis of different languages, as demonstrated in Text 2 and Text 3, is equally perplexing. These alterations, far from being mere nuances, introduce divergent interpretations.

The inconsistency of models in their own perspectives is evident, and the inability to compare different models is apparent. The application of a hermeneutical perspective appears to be both unclear and erratic. This discrepancy persists when employing an alternative prompting strategy, as evidenced in Table II. However, the number of suggested sections does undergo a substantial change. However, the value ranges do not offer any possible explanations within the text itself. In summary, three factors appear to influence the output: the model, the prompt, and the language.

B. Actors

As demonstrated in Figures 3 and 4, the boxplots illustrate the range of detected actors within the text corpus. For a more detailed analysis, please refer to Table [tab:personen]. The ensuing results proffer illuminating insights from a variety of perspectives. For instance, it is noteworthy that Texts 3 and 4 do not explicitly mention any actors, yet they do implicitly refer to actors through the use of language such as "Ladies and Gentlemen." However, the models detect a multitude of actors. The detection of more than 35 actors in a philosophical text is substantial, and the identification of four actors in a single sentence of greeting appears implausible. It is noteworthy that models can exhibit disagreement in their estimation of the number of actors mentioned in a narrative text or poem (texts

TABLE IV: Details of found actors with SE approach on Texts 2-3 with different languages (G=German, E=English)

L	Text	Model	Min	Q_1	Median	Q_3	Max
G	Text 2	llama3:70b	22.0	29.0	37.0	43.0	53.0
		ChatGPT-4	35.0	26.0	26.0	36.0	36.0
		qwen2.5:72b	14.0	18.5	19.0	19.0	34.0
		mistral:7b	8.0	16.0	18.0	25.0	34.0
E	Text 2	llama3:70b	24.0	27.5	31.0	34.5	47.0
		ChatGPT-4	23.0	27.5	32.0	35.0	36.0
		qwen2.5:72b	36.0	36.0	36.0	36.0	36.0
		mistral:7b	27.0	28.5	30.0	34.0	34.0
G	Text 3	llama3:70b	14.0	23.5	25.0	34.0	44.0
		ChatGPT-4	5.0	7.5	9.0	9.5	21.0
		qwen2.5:72b	3.0	8.0	20.0	27.5	35.0
		mistral:7b	6.0	15.0	18.0	19.0	23.0
E	Text 3	llama3:70b	1.0	1.0	1.0	2.0	26.0
		ChatGPT-4	8.0	9.5	10.0	11.5	12.0
		qwen2.5:72b	10.0	10.0	10.0	10.0	10.0
		mistral:7b	1.0	1.0	1.0	2.5	3.0

1 and 2). It is important to note that modifying the input language can substantially alter the output.

The lack of comparability and inconsistency of these models to other LLMs, as previously discussed in the context of plot, further complicates the analysis. It is evident that modifying the prompting strategy leads to alterations in the output, as illustrated in Table IV. It is important to note that none of these numbers are directly connected to the text itself. LLMs exhibit deficiencies in addressing fundamental hermeneutical elements that are essential for text exegesis.

V. Conclusions and Outlook

This study critically examined the extent to which large language models (LLMs) are capable of engaging in hermeneutical analysis, drawing from traditions in philosophy, theology, sociology, and literary theory. The present study evaluated the consistency, interpretive depth and methodological coherence of LLM outputs by testing multiple models on a range of textual genres, including poetic, theological, philosophical and conversational texts. The findings indicate that current LLMs exhibit significant variability and lack reproducibility in the application of hermeneutical principles. Across a range of models and prompting strategies, the interpretation of plot structures and identification of actors demonstrated inconsistency, both within and across iterations. This variability suggests that LLMs are not applying coherent hermeneutical frameworks, but rather relying on surface-level linguistic patterns that diverge significantly depending on language, prompt design, and model architecture.

Nevertheless, this is at odds with some of the recent literature that has been published, which has asserted that LLMs "can analyze nearly any textual statement." [49] The experimental results obtained in this study appear to be in alignment with the arguments set forth by researchers who posit that LLMs are deficient in their capacity to perform even rudimentary logic-based tasks, such as counting and identifying general substructures in graphs [50], [51]. Therefore, it is debatable whether LLMs could offer any technical assistance with textual analysis, against, for example [52].

It has been demonstrated that there is a discernible discrepancy between certain components of LLM output and hermeneutical procedures. These components include, but are not limited to, text segmentation and the identification of narrative elements. The models demonstrate an inability to comprehend text, see [53], and their apparent lack of understanding of context, intentionality, and reader-writer dynamics. These components are fundamental to traditional hermeneutics. In other words: "Should our interpretation capabilities be engaged? If yes, under what conditions? The rules of the language game should be spelled out; they should not be passed over in silence." [54] Moreover, the validity of conventional hermeneutical principles is also open to question. For instance, the assumption that hermeneutics necessitates the interpretation of a human being with consciousness, engaging with interpretive dimensions involving principles, historical consciousness, praxis, or – in the context of theological hermeneutics – the influence of the Holy Spirit, is not necessarily indisputable. The assumption that an LLM could act as an exeget is not, in itself, sufficient to ensure that it will be able to identify all the important aspects of context necessary for the application of hermeneutics. It appears improbable that LLMs will ever be capable of this [55].

It is recommended that future work concentrate on the development of evaluation criteria for the purpose of assessing interpretive coherence in the context of machine-generated text analysis. This is a common debate also in the field of AI-generated text detection, see [56]. Furthermore, it is recommended that interdisciplinary frameworks be explored, with a view to more accurately reflecting the limitations of computational models within the domain of humanities or theological understanding. Moreover, the integration of insights from phenomenology and semiotics has the potential to provide supplementary conceptual foundations for the assessment of interpretive capacities of AI systems.

Appendix A
Prompts

In the process of designing the prompts, adherence was observed to the general principle for success when engineering a prompt with clear and concise, step-by-step reasoning, as outlined in [45]. TThe direct approach posed a question that was explicit in its formulation.The 'Single Explanation' (SE) strategy is a rhetorical device that functions to provide an illustration of a concept or idea.

References

[1] Danner, Helmut, *Hermeneutik: Zugänge, Perspektiven, Positionen*, 1. Aufl ed. Freiburg: wbg Academic, 2021.

[2] G. Maier, *Biblische Hermeneutik*. SCM, 2014.

[3] M. Oeming, *Biblische Hermeneutik*. WBG, 2013.

[4] R. Maciag, "The hermeneutics of artificial text," in *Computer Sciences & Mathematics Forum*, vol. 8, no. 1. MDPI, 2024, p. 94.

[5] L. Henrickson and A. Meroño-Peñuela, "Prompting meaning: a hermeneutic approach to optimising prompt engineering with chatgpt," *AI & SOCIETY*, pp. 1–16, 2023.

[6] L. Tiantian and M. Carl, "Toward practical hermeneutics of fourth paradigm ai for science," *«Technology and language»*, vol. 5, no. 1, pp. 89–105, 2024.

[7] K. Petković, "The truth and ai: Towards a new hermeneutic of the academic subject?" *Sociologija*, vol. 66, no. 4, pp. 581–611, 2024.

[8] P. Pinell, "Does artificial intelligence speak our language?: A gadamerian assessment of generative language models," *Political Research Quarterly*, vol. 77, no. 3, pp. 713–728, 2024.

[9] M. Heimann and A.-F. Hübener, "Circling the void: Using heidegger and lacan to think about large language models," *Cognitive Systems Research*, p. 101349, 2025.

[10] A. Diržytė, "Large language models and the enhancement of human cognition: Some theoretical insights," *Filosofija. Sociologija*, vol. 36, no. 1, 2025.

[11] M. Barnard and W. Otte, "Is the machine surpassing humans? large language models, structuralism, and liturgical ritual: A position paper," *International Journal of Practical Theology*, vol. 28, no. 2, pp. 289–306, 2024.

[12] M. Chrostowski and A. J. Najda, "Chatgpt as a modern tool for bible teaching in confessional religious education: a german view," *Journal of Religious Education*, pp. 1–20, 2024.

[13] R. Demichelis, "The hermeneutic turn of ai: Is the machine capable of interpreting?" *arXiv preprint arXiv:2411.12517*, 2024.

[14] M. Gonzalez, "Beyond replacement: A hermeneutic phenomenological approach to human-ai collaboration," in *2024 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2024, pp. 1–4.

[15] A. Elrod, "Uncovering theological and ethical biases in llms: An integrated hermeneutical approach employing texts from the hebrew bible," *HIPHIL Novum*, vol. 9, no. 1, pp. 2–45, 2024.

[16] K. Karches, "Hermeneutics as impediment to ai in medicine," *Theoretical Medicine and Bioethics*, pp. 1–19, 2025.

[17] E. Krapp, R. Neuhaus, M. Hassenzahl, and M. Laschke, "In a quasi-social relationship with chatgpt. an autoethnography on engaging with prompt-engineered llm personas," in *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, 2024, pp. 1–16.

[18] Körtner, Ulrich H. J., *Der inspirierte Leser: zentrale Aspekte biblischer Hermeneutik*. Göttingen [München]: Vandenhoeck und Ruprecht [Bayerische Staatsbibliothek], 1994.

[19] M. Oeming, *Biblische Hermeneutik: Eine Einführung*. Darmstadt: WBG, 2007.

[20] J. Resseguie, *Narrative Criticism of the New Testament: An Introduction*. Grand Rapids: Baker Publishing Group, 2005.

[21] N. R. Petersen, *Literary Criticism for New Testament Critics*. Minneapolis: Fortress Press, 1978.

[22] Wyss, Dieter, *Zwischen Logos und Antilogos: Untersuchungen zur Vermittlung von Hermeneutik und Naturwissenschaft*. Göttingen [München]: Vandenhoeck & Ruprecht, 1980.

[23] Schröer, Norbert and Bidlo, Oliver, "Die Entdeckung des Neuen: Qualitative Sozialforschung als Hermeneutische Wissenssoziologie." Wiesbaden: VS Verlag für Sozialwissenschaften / Springer Fachmedien Wiesbaden GmbH, Wiesbaden, 2011, conference.

[24] U. Oevermann, "Zur analyse der struktur von soialen deutungsmustern," 1973. [Online]. Available: https://publikationen.ub.uni-frankfurt.de/opus4/frontdoor/deliver/index/docId/4951/file/Struktur-von-Deutungsmuster-1973.pdf

[25] J. Reichertz, "Die objektive hermeneutik – darstellung und kritik," in *Qualitative Forschung*, E. König and P. Zedler, Eds. Beltz Verlag, 2002, pp. 123–156.

[26] H. Niebuhr and D. Ottati, *The Meaning of Revelation*. Westminster: John Knox Press, 2006.

[27] D. Ritschl and H. O. Jones, *Story als Rohmaterial der Theologie*. München: Kaiser, 1976.

[28] B. Wacker, *Narrative Theologie?* München: Kösel, 1977.

[29] G. Genette, *Figures III*. Paris: Éditions du Seuil, 2014.

[30] A. Cornils, *Vom Geist Gottes erzählen: Analysen zur Apostelgeschichte*. Tübingen: Francke, 2006.

[31] J. Bailey and L. Vander Broek, *Literary Forms in the New Testament: A Handbook*. Westminster: John Knox, 1992.

[32] N. Perrin, "Historical criticism, literary criticism and hermeneutics: The interpretation of the parables of jesus and the gospel of mark today," in *Hermeneutical Inquiry*, D. E. Klemm, Ed. Atlanta, Georgie: Scholars Press, 1986, pp. 253–266.

[33] J. A. Darr, "'Watch How you Listen' (Lk. 8.18): Jesus and the Rhetoric of Perception in Luke–Acts," in *The New Literary Criticism and the New Testament*, ser. JSNTS, E. S. Malbon and E. V. McKnight, Eds. Sheffield: Academic Press, 1994, vol. 109, pp. 87–107.

[34] B. Dörpinghaus, "Die narratologische exegese alttestamentlicher erzähltexte: Neue perspektiven auf die jakobserzählung," *European journal of theology*, vol. 31, no. 2, pp. 186–214, 2022.

[35] B. Dörpinghaus and H.-G. Wünch, "Relationships and forms in the social network of the jacob narrative: A narratological operspective," *Old Testament Essays*, vol. 36, no. 2, pp. 347–367, 2023.

[36] B. Dörpinghaus, "Esau und jakob: Familien und figuren im wandel?" *European Journal of Theology*, vol. 32, no. 2, pp. 180–201, 2023.

[37] B. Kern, *Theologie der Befreiung*. UTB, 2013.

[38] G. W. Locher, *Zwingli und die schweizer Reformation*, ser. Die Kirche in ihrer Geschichte. Vandenhoeck & Ruprecht, 1982.

[39] V. Leppin, "Zwingli, Ulrich," in *Theologische Realenzyklopädie*, H. Balz, G. Krause, and G. Müller, Eds. Berlin: de Gruyter, 2004, no. Bd. 36.

[40] W. van't Spijker, *Calvin*, ser. Die Kirche in ihrer Geschichte. Vandenhoeck & Ruprecht, 2001.

[41] W. Nijenhuis, "Calvin, Johannes," in *Theologische Realenzyklopädie*, H. Balz, G. Krause, and G. Müller, Eds. Berlin: de Gruyter, 1981, no. Bd. 07.

[42] E. Soja, *Thirdspace: Journeys to Los Angeles and Other Real-and-Imagined Places*. Wiley, 1996.

[43] H. Lefebvre and D. Nicholson-Smith, *The production of space*. Oxford: Blackwell, 1991, vol. 142.

[44] U. Oevermann, "Zur sache. die bedeutung von adornos methodologischem selbstverständnis für die begründung einer materialen soziologischen strukturanalyse." in *Adorno-Konferenz*, L. von Friedeburg and J. Habermas, Eds., 1983, pp. 234–289.

[45] R. Fechner and J. Dörpinghaus, "No train, no pain? assessing the ability of llms for text classification with no finetuning," in *Proceedings of the Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia*, 2024, pp. 8–11.

[46] K. Bhattarai, I. Y. Oh, J. M. Sierra, J. Tang, P. R. Payne, Z. Abrams, and A. M. Lai, "Leveraging gpt-4 for identifying cancer phenotypes in electronic health records: a performance comparison between gpt-4, gpt-3.5-turbo, flan-t5, llama-3-8b, and spacy's rule-based and machine learning-based methods," *JAMIA open*, vol. 7, no. 3, p. ooae060, 2024.

[47] P. Ersoy and M. Erşahin, "Benchmarking llama 3 70b for code generation: A comprehensive evaluation," *Orclever Proceedings of Research and Development*, vol. 4, no. 1, pp. 52–58, 2024.

[48] M. Merkel and J. Dörpinghaus, "A case study on the transformative potential of ai in software engineering on leetcode and chatgpt," *arXiv preprint arXiv:2501.03639*, 2025.

[49] P. Törnberg, "How to use llms for text analysis," *arXiv preprint arXiv:2307.13106*, 2023.

[50] T. Fu, R. Ferrando, J. Conde, C. Arriaga, and P. Reviriego, "Why do large language models (llms) struggle to count letters?" *arXiv preprint arXiv:2412.18626*, 2024.

[51] L. Nguyen and Y. Yan, "Evaluating the structural awareness of large language models on graphs: Can they count substructures?" *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, 2024.

[52] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. G. Monteith, "An examination of the use of large language models to aid analysis of textual data," *International Journal of Qualitative Methods*, vol. 23, p. 16094069241231168, 2024.

[53] W. S. Saba, "Llms' understanding of natural language revealed," *arXiv preprint arXiv:2407.19630*, 2024.

[54] A. Videla, "Echoes of intelligence: Textual interpretation and large language models," *Queue*, vol. 21, no. 3, pp. 36–53, 2023.

[55] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, "Position: Will we run out of data? limits of llm scaling based on human-generated data," in *Forty-first International Conference on Machine Learning*, 2024.

[56] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, "A survey on llm-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, pp. 1–66, 2025.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 693–697
DOI: 10.15439/2025F4025
ISSN 2300-5963 ACSIS, Vol. 43

Knitwear Production Scheduling

Stefka Fidanova
Ruse University "Angel Kanchev",
Ruse, Bulgaria
Centre of Excellence in Informatics and Information and Communication Technologies
Sofia, Bulgaria
E-mail: stefka.fidanova@iict.bas.bg

Jelenko Stanchov
Institute of Information and Communication Technology
Bulgarian Academy of Sciences
Sofia, Bulgaria
E-mail: elenko.stanchov@iict.bas.bg

Maria Ganzha
System Research Institute
Polish Academy of Sciences
Warsaw, Poland
E-mail: maria.ganzha@ibspan.waw.pl

***Abstract*—Clothing production is an important sector within manufacturing. It includes the sewing, knitting and leather industries. It is important for a manufacturer to organize the production process well. This organization includes personnel allocation, machine loading, and material allocation. The goal is to complete the given order in the shortest time and, if possible, at the lowest cost. This paper addresses the task of producing knitted shirts. An ant colony optimization algorithm has been proposed to solve the problem. The objective is to complete the order in the shortest possible time.**

***Index Terms*—Feshion production, Optimization, Scheduling**

I. Introduction

OPTIMIZING knitted garment production is a special part of job shop scheduling problem. Job shop scheduling problem is an important problem arising in manufacturing and industry. The production consists of producing various elements/parts on different machines. Sometimes it matters in what order the individual elements are produced. Individual machines may have some specialization, as well as different productivity. Therefore, it is important to find the optimal distribution of the production across machines so that the final product can be produced in minimal time or within a certain period of time and, if possible, at a lower cost. There is a wide variety of job shop scheduling problems, depending on the specific production processes, requirements, and constraints. A survey was conducted in about the various job shop problems and algorithms solving them [1]. This is a typical machine scheduling problem. For the first time, an algorithm for solving a variant of this problem, with machine setup time included, was made in 1954 in [2]. Next years other variants of the problem called flow shop scheduling problem is solved in [3], [4]. These publications solve small-sized problems. There are 1 to 4 operations that are performed on 2 to 4 machines. Exact methods are used to solve the problem. The term job shop was first used by Sisson in 1959 [5].

In general, the job shop scheduling problem can be defined as follows:

There is a set of m machines $M = \{M_1, M_2, \ldots, M_m\}$.

There is a set of n jobs $J = \{J_1, J_2, \ldots, J_n\}$.

The job J_i has a set of operations $O_i = \{O_1^1, O_2^i, \ldots, O_{n_i}^i\}$. The operations have predefined technological sequence.

The operation O_i^j has an operation time p_{ij} and is assigned on machine M_k from the set of machines M.

The solution is a scheduling of the jobs on the machines. The aim is to minimize the completing time of all the jobs.

When the problem is large, it is difficult to solve it in a reasonable time using exact methods. For this reason, metaheuristic methods are applied, which find an approximate solution in a short time.

In this article, we solve a variant of the job shop scheduling problem, related to the production of knitted clothes. To solve the problem, we propose an algorithm based on the ant colony optimization.

The organization of the rest of the paper is as follows. Section 2 gives literature review of the variants of job shop scheduling problems and methods to solve them. Section 3 provides the definition of the considered variant of the job shop scheduling problem. Section 4 describes the algorithm for solving the problem. Computational examples are given in Section 5. Section 6 gives some concluding remarks and direction for future work.

II. Literature Review

When manufacturing a product, we have various measures and constraints. There may be a requirement that jobs be completed in a minimum time or within a given interval. Another requirement may be minimizing the cost of the production, and others. The problem may relate to optimizing one of the measures or to meeting a maximum number of criteria.

In [6] scheduling considers flexible manufacturing system. They solve the problem by applying an adaptive genetic algorithm. Machine flexibility is the focus of the problem solved in [7], [8]. They consider energy efficiency too. Job shop scheduling problem is applied on semiconductor production [9], manufacturing of machines [10], automobile industry [11], metallurgical industry [12].

Wang [13] solves robust job shop problem with a specific structure. Dynamic flexible job shop problem is solved in [14]. Biogeography-based optimization algorithm for solving flexible multi-objective job shop problem has been proposed in [15].

Anghinolfi [16] apply a split greedy heuristic for parallel machine scheduling. In [17] has been proposed a hybrid genetic algorithm for energy adaptive production. Scalia [18]

proposed mixed integer programming model for solving job shop scheduling problem. In [19] is applied implementation in Python using *pyomo* library for solving the problem. A review of flexible job shop scheduling problem is done in [20].

III. Knitwear Production Problem

In this article, we consider the problem of optimizing the production of knitted garments. This is a variant of job shop scheduling problem. Suppose there is an order to produce knitted shirts. The shirts come in $s = 3$ sizes: M, L and XL. $N_{i,j}$ units of type i and size j must be produced, where $i = 1, \ldots, d$ and $j = 1, \ldots, s$. Each shirt consists of $d = 4$ types of parts: front, back, collar and two sleeves. t_{ij} is the time for manufacturing part i of size j, $i = 1, \ldots, d$ and $j = 1, \ldots, s$. There is a set M of m machines available. $M_{i,j,k}$ is the number of parts of type i and size j that are manufactured on machine k, $i = 1, \ldots, d$, $j = 1, \ldots, s$ and $k = 1, \ldots, m$. All machines are identical. The material used to make the shirts is the same and therefore does not require recharging.

The time during which machine k operates is:

$$T_k = \sum_{i=1}^{d} \sum_{j=1}^{s} M_{i,j,k} * t_{i,j} \tag{1}$$

We assume that all machines start working on the order at the same time. Thus the order completion time is:

$$T = \max_{k=1}^{m} T_k \tag{2}$$

or this is the time by which all the machines will have finished working.

The objective is to minimize T, or to find the minimal order completion time, $\min T$. The solution is described by the matrix M with elements $M_{i,j,k}$, indicating which parts will be manufactured on which machine.

The mathematical description of the problem is as follows: We look for:

$$min \left(max_{k=1}^{m} \left(\sum_{i=1}^{d} \sum_{j=1}^{s} M_{i,j,k} * t_{ij} \right) \right) \tag{3}$$

fulfilling the conditions:

$$\textstyle\sum_{k=1}^{m} M_{i,j,k} = N_{i,j} \quad i = 1, \ldots d \quad j = 1 \ldots s \tag{4}$$

The aim is to finish the order for the minimal time.

IV. Production Scheduling

The job shop scheduling problem is NP-hard combinatorial optimization problem. This type of problem requires a large amount of computer memory and an exponential number of calculations. For this reason, when the size of the problem is large, it could not be solved by an exact method or traditional numerical method. In such cases, metaheuristic algorithms are applied. Metaheuristic methods are iterative stochastic methods, based on random searches or searches with probability and statistical feedback from the results of previous iterations. Prior knowledge of the task and estimates, if any, are also used to improve and guide the search.

One of the most successful method for solving combinatorial optimization problems is Ant Colony Optimization (ACO) [21], [22]. In the early 1990s, Marco Dorigo proposed a stochastic method for solving combinatorial optimization problems, based on observations of the behavior of ants in nature [23]. The main thing is how ants manage to find the shortest path to a food source, marking the paths with pheromone and following the one with the highest concentration of pheromone. In order for ants to be imitated, the problem to be solved must be reduced to finding the shortest path. This is done by representing the problem in terms of a graph. A shortest path is searched in the graph under certain conditions, using numerical information (weight) on the edges or nodes of the graph, imitating the pheromone in nature. Ants use a combination of pheromone and random search to improve their path.

Artificial ants move to a new node in the task graph using a transition probability. This consists of heuristic information and the pheromone corresponding to that transition.

$$P_{i,j} = \frac{\tau_{i,j}^{a} \cdot \eta_{i,j}^{b}}{\sum\limits_{k \in Unused} \tau_{i,k}^{a} \cdot \eta_{i,k}^{b}}, \tag{5}$$

where P_{ij} is the transition probability to go from node i to node j, η_{ij} is the heuristic information related to the problem and τ_{ij} is the quantity of the pheromone, $Unused$ is the set of nodes which are not used yet in the solution.

It starts with a small positive pheromone value, the same for all elements of the graph τ_0, $0 < \tau_0 < 1$. After each iteration, the pheromone is updated depending on whether the corresponding vertex belongs to a solution and how good this solution is compared to the other solutions found. The main pheromone update rule is:

$$\tau_{i,j} \leftarrow \rho \cdot \tau_{i,j} + \Delta\tau_{i,j}, \tag{6}$$

ρ mimics evaporation in nature, which reduces the amount of pheromone over time. Then, a new pheromone is added $\Delta\tau_{i,j}$, which depends on the quality of the solution and is inversely proportional to the value of the objective function.

The iteration ends when all ants have built their solutions, i.e. the transition probability is 0. The algorithm ends its work when: it has performed its predetermined number of iterations; when there is no improvement in the result for a fixed number of iterations; when the value of the objective function is sufficiently close to a predetermined lower bound.

One of the important elements in applying the ant method is the description of the problem with a graph. The graph we use to describe the problem we are solving is three-dimensional, and node (i, j, k) corresponds to element i of size j manufactured on machine k, $i = 1, \ldots, d$, $j = 1, \ldots, s$, $k = 1, \ldots, m$. This node is assigned a value $M_{i,j,k}$, which indicates how many units of element i of size j to be manufactured on machine k. So if on machine k produces parts of type i, size j, then $M_{i,j,k} > 0$. If on machine k

does not produced parts of type i, size j, then $M_{i,j,k} = 0$. This means that a single part and size can be manufactured on several machines, and different parts with different sizes can be manufactured on one and the same machine.

Graph of the problem is completely connected. An ant begins building the solution from a random node in the graph. At the begining the value of $M_{i,j,k} = 0$, $i = 1, \ldots, d$, $j = 1, \ldots, s$, $k = 1, \ldots, m$, because the distribution of the parts along the machine is not fixed. When the ant passes through node (i, j, k), the value of $M_{i,j,k}$ increases by 1 or $M_{i,j,k} \leftarrow M_{i,j,k} + 1$. The ant can go through the same vertex (i, j, k) multiple times adding 1 to $M_{i,j,k}$. The important thing is not to violate the constraints of the problem.

The pheromone will be associated with the nodes of the graph. It indicates the importance of the node to the solution. Thus, pheromone $\tau_{i,j,k}$ corresponds to node (i, j, k) and its value indicates how important it was for the solutions of the previous iteration. So, the quantity of the pheromone $\tau_{i,j,k}$ indicates how significant the node was for the decisions of previous iterations.

At the beginning, we randomly select one of the elements to be manufactured and assign it to a randomly selected machine. Every time an element i of size j is assigned to machine k, we increase the value of $M_{i,.j,k}$ by 1.

To assign the next elements for fabrication, we select element i of size j at random and we calculate the transition probability to assign it to one of the machines. The transition probability is a product of the amount of pheromone and the heuristic function uses prior knowledge of the problem to guide the search for a good solution. We propose the following heuristic function:

$$\eta_{i,j,k} = \frac{1}{T_k} \tag{7}$$

where the value of T_k is calculated for the current values of the elements of the matrices M by increasing the value of $M_{i,j,k}$ by 1. We calculate the transition probability for the selected element i of size j for all possible machines k. We choose the node (machine) with maximum probability. If there are two or more nodes $(i, j, k_1), \ldots, (i, j, k_s)$, $1 \leq s \leq m$, with the same transition probability value, then we choose one of them randomly.

After an element i of size j is selected to be manufactured on machine k, we increase the value of $M_{i,j,k}$ by 1, $M_{i,j,k} \leftarrow M_{i,j,k} + 1$. At the same time we decrease the value of $N_{i,j}$ by 1 $N_{i,j} \leftarrow N_{i,j} - 1$. The current value of $N_{i,j}$ shows how many items of type i and size j have not yet been assigned to the machines for the production.

The algorithm repeats until $N_{i,j} = 0$, for all $i = 1, \ldots, d$ and $j = 1, \ldots, s$. This means that all the parts have already been assigned for the production on the machines, i.e. the ant has built the solution. The matrix M shows for each type and size of parts on which machine and how many are produced.

The order of the production of the parts is not important, so for each machine k it is possible to produce first all the parts of type 1 and size 1, then all the parts of type 1 size 2 and so on, and at the end all the parts of type d and size s.

When all the ants have built their solutions, we calculate the value of the objective function, i.e. the time to complete the order. Then we update the pheromone. After reducing/evaporating the pheromone with the evaporation parameter ρ, we add a new pheromone.

$$\begin{array}{lll} \Delta\tau_{i,j,k} = & (1-\rho)\frac{1}{T} & \text{if} \quad M_{i,j,k} \neq 0 \\ \Delta\tau_{i,j,k} = & 0 & \text{if} \quad M_{i,j,k} = 0 \end{array} \tag{8}$$

This concludes the current iteration. The algorithm performs a predetermined number of iterations or there is a fixed number of iterations without updating the results. In our problem, updating the results occurs when the time for filling the document is reduced.

V. Computational examples

The organization of the production is very important for its success. Given a variety of elements and a fixed number of the production machines, the elements must be distributed so that the order can be completed in the shortest possible time. In the task of producing knitted garments, there are knitting machines that can produce each of the elements. The material used to make the garments is the same for the entire order. Therefore, there is no delay in moving from one item to another because there is no need to load new material. The garments produced come in three sizes: *M*, *L*, *XL*. The machines available for the production are identical and the same item is produced on each of the machines in the same amount of time. Each garment consists of four parts/elements, a front, a back, a collar and two sleeves. The individual elements can be manufactured on different machines. It is important that for each size the number of fronts is equal to the number of backs, is equal to the number of collar and the number of sleeves is twice as large as the number of other elements. This is a necessary requirement so that the garments can then be assembled. Let the time for making the individual elements be as indicated in Table I.

TABLE I: Knitted shirt production time

element	size M	size L	size XL
front	17 min	18 min	19 min
back	13 min	14 min	15 min
sleeve	10.5 min	11 min	12.5 min
collar	3 min	3 min	3 min

The data on the production time for each element was taken from a Bulgarian company producing knitted garments.

Let the company have 3 machines. The machines are identical and each of them can knit each of the elements. Let an order be received for the production of knitted shirts as noted in Table II.

TABLE II: Knitted shirt order

shirts	size M	size L	size XL
	10 pieces	20 pieces	10 pieces

Each size could be produced on a separate machine. This is one possible solution, but not optimal. If the production of

the parts is distributed in this way, each of the machines will complete its assigned work in the following time Table III.

TABLE III: Knitted shirt production time solution 1

shirts	machine 1	machine 2	machine 3
shirt quantity	10M	20L	10 XL
front	10M	20L	10XL
back	10M	20L	10XL
sleeve	20M	40L	20XL
collar	10M	20L	10XL
production time	540 min	1140 min	620 min

With this distribution of work, the order will be completed in 1140 minutes, which is 19 hours. This is the time by which all machines will have finished working. The machine 2 finished last and it determines the order execution time.

Another valid solution is to have machine 1 produces all 10 risques of size M and only 4 of size L. Machine 2 produces 13 risques of size L, and machine 3 produces all 10 risques of size XL and 3 risques of size L. This way, the distribution of work across machines will be almost even. The machines will complete its assigned work in the following time Table IV.

TABLE IV: Knitted shirt production time solution 2

shirts	machine 1	machine 2	machine 3
shirt quantity	10 M + 4 L	13 L	3 L + 10 XL
front	10M, 4L	13L	10XL
back	10M, 4L	13L	10XL
sleeve	20M, 8L	20L	10XL
collar	10M, 4L	13L	10XL
production time	768 min	741 min	791 min

With this distribution of work, the order will be completed in 791 minutes, which is 13 hours and 11 minutes, which is shorter than previous solution. The machine 3 finished last and it determines the order execution time.

Let's apply the ant colony optimization to solving the problem of making knitted shirts. We use 2 ants. Every ant constructs his own solution and after we compare them and choose the better and it will be the iteration best solution. We compare it with the current global best solution and if it is better it is the new global best solution. The ant chose randomly an element from the order and assigns it to randomly chosen machine. After the ant chose randomly other element from the order and calculates the transition probability to be assigned to every one of the existing machines. The element is assigned for the production to the machine with a higher probability. This procedure is repeated till all elements from the order are assigned. After several iterations our ant algorithm achieves following solutions, see Table V and Table VI.

TABLE V: Ant's solution 1

element	machine 1	machine 2	machine 3
front	10M, 4L	14L, 1XL	2L, 9XL
back	10M, 4L	14L, 1XL	2L, 9XL
sleeve	20M, 8L	28L, 2XL	4L, 18XL
collar	10M, 4L	14L, 1XL	2L, 9XL
time	768 min	760 min	770 min

With this distribution of work, the order will be completed in 770 minutes, which is 12 hours and 50 minutes, which is shorter than previous solution. The machine 3 finished last and it determines the order execution time.

TABLE VI: Ant's solution 2

element	machine 1	machine 2	machine 3
front	10M, 4L	14L, 1XL	2L, 9XL
back	10M, 4L	14L, 1XL	2L, 9XL
sleeve	20M, 8L	28L, 2XL	4L, 18XL
collar	10M, 3L	16L, 1XL	1L, 9XL
time	765 min	766 min	767 min

With this distribution of work, the order will be completed in 767 minutes, which is 12 hours and 47 minutes, which is shorter than previous solution. The machine 3 finished last and it determines the order execution time. We observe that the time during which the machines work to fulfill the order is almost the same. The difference in time is within 1 to 2 minutes, which is less than the time for producing any of the elements. Therefore, if we move a part from one machine to another, the total order execution time will not be reduced any further. This gives us reason to assume that the proposed algorithm, based on the ant colony optimization, has found the optimal solution for this example.

VI. Conclusion

This paper proposes an ant colony optimization algorithm for solving the knitwear production scheduling problem. This is a problem coming from the industry. The simplest variant of the task is solved, where we have a single type of machine and the manufactured parts are of the same type and differ only in size. In our future work, we will test the algorithm with large examples with a range of several hundred numbers of elements. We will include the creation of clothes in different colors, which will require a recharge of the machine, which takes time , but it will not require a reconfiguration. Another variant of this problem involves producing multiple types of garments, which would require machine reconfiguration time.

Acknowledgment

The work is supported by the project "Russe ressearch university" financed by European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project number BG-RRP-2.013-0001 and by the Polish-Bulgarian collaborative grant "Practical aspects for scientific computing". The work was partially supported by the Centre of Excellence in Informatics and ICT under the Grant No BG16RFPR002-1.014-0018, financed by the Research, Innovation and Digitalization for Smart Transformation Programme 2021-2027 and co-financed by the European Union.

References

[1] Xiong H., Shi S., Ren D., Hu J.,*A survey of job shop scheduling problem: The types and models*, Computers & Operations Research, Vol. 142,2022, 105731.

[2] Johnson S.M., *Optimal Two- and Three-stage Production Schedules With Setup Times Included*, Naval Res. Logist. Quart., Vol. 1(1), 1954, 61-68.

[3] Akers S.B., Friedman J., *A Non-Numerical Approach to Production Scheduling Problems*, J. Oper. Res. Soc. Am., Vol. 3(4), 1955, 429-442.
[4] Akers S.B., *A Graphical Approach to Production Scheduling Problems*, Oper. Res., Vol.4(2), 1956, 244-245.
[5] Sisson R.L., *Methods of Sequencing in Job Shops- A Review*, Oper. Res., Vol.7(1), 1959, 10-29.
[6] Jerald J., Asokan P., Saravanan R., Rani A.D.C., *Simultaneous scheduling of parts and automated guided vehicles in an FMS environment using adaptive genetic algorithm*, Int. J. Adv. Manuf. Technol., Vol.29(5–6), 2006, 584-589.
[7] Yin L., Li X., Gao L., Lu C., Zhang Z., *A novel mathematical model and multi-objective method for the low-carbon flexible job shop scheduling problem*, Sustainable Comput. Inf. Syst., Vol.13, 2017, 15-30.
[8] Meng L., Zhang C., Shao X., Ren Y., *MILP models for energy-aware flexible job shop scheduling problem*, J. Cleaner Prod., Vol. 210, 2019, 710-723.
[9] Ham A., *Flexible job shop scheduling problem with parallel batch processing machine*, Proceedings of the 2016 Winter Simulation Conference, 2016, 2740-2749.
[10] Gao K.Z., Suganthan P.N., Pan Q.K., Tasgetiren M.P., A. Sadollah A., *Artificial bee colony algorithm for scheduling and rescheduling fuzzy flexible job shop problem with new job insertion*, Knowl.-Based Syst., Vol. 109, 2016, 1-16.
[11] Liu Q., Zhan M., Chekem F.O., Shao X., Ying B., Sutherland J.M., *A hybrid fruit fly algorithm for solving flexible job-shop scheduling to reduce manufacturing carbon footprint*, J. Cleaner Prod., Vol. 168, 2017, 668-678.
[12] Yang Y.Z., Gu X.S., *Pareto-based complete local search and combined timetabling for multi-objective job shop scheduling problem with no-wait constraint*, J. Donghua Univ. (English Ed.), Vol. 33(4), 2016, 601-624.
[13] Wang X., Wang B., Zhang X., Xia X., Q. Pan Q., *Two-objective robust job-shop scheduling with two problem-specific neighborhood structures*, Swarm Evol. Comput., Vol. 61, 2021, 100805.
[14] Li K., Deng Q., Zhang L., Fan Q., Gong G., Ding S.,*An effective MCTS-based algorithm for minimizing makespan in dynamic flexible job shop scheduling problem*, Comput. Ind. Eng., Vol. 155(2), 2021, 107211.
[15] An Y., Chen X., Li Y., Han Y., Zhang J., Shi H., *An improved non-dominated sorting biogeography-based optimization algorithm for the (hybrid) multi-objective flexible job-shop scheduling problem*, Appl. Soft Comput.,Vol. 99, 2021, 106869.
[16] Anghinolfi D., Paolucci M., Ronco R., *A bi-objective heuristic approach for green identical parallel machine scheduling*, Eur. J. Oper. Res., Vol. 289(2), 2021, 416-434.
[17] Grosch B., Kohne T., Weigold M,, *Multi-objective hybrid genetic algorithm for energy adaptive production scheduling in job shops*, Procedia CIRP, Vol. 98, 2021, 294-299.
[18] Scalia B., Leite C.F., The Job-Shop Scheduling Problem: Mixed-Integer Programming Models, TCS Archive, 2023.
[19] Bynum, M. L. et al., *Pyomo-optimization modeling in python*, Springer, 2021.
[20] Dauzère-Pérès S., Ding J., Shen L., Tamssaouet K., *The flexible job shop scheduling problem: A review*, European Journal of Operational Research, Vol. 314(2), 2024, 409-432.
[21] Dorigo M, Stutzle T., *Ant Colony Optimization*, MIT Press, 2004.
[22] Fidanova S. Hybrid Ant Colony Optimization Algorithm for Multiple Knapsack Problem. 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), IEEE, 2021, DOI:10.1109/ICRAIE51050.2020.9358351, 1-5.
[23] Bonabeau E., Dorigo M. and Theraulaz G., *Swarm Intelligence: From Natural to Artificial Systems*, New York,Oxford University Press, 1999.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 699–704
DOI: 10.15439/2025F9981
ISSN 2300-5963 ACSIS, Vol. 43

AI-MTD: Zero-Trust Artificial Intelligence Model Security Based on Moving Target Defense

Daniel Gilkarov
0009-0008-9274-802X
Department of Computer and Software Engineering
Ariel Cyber Innovation Center
Ariel University, Israel
daniel.gilkarov1@msmail.ariel.ac.il

Ran Dubin
0000-0002-2055-2211
Department of Computer and Software Engineering
Ariel Cyber Innovation Center
Ariel University, Israel
rand@ariel.ac.il

***Abstract*—This paper examines the challenges in distributing AI models through file transfer mechanisms. Despite advancements in security measures, vulnerabilities persist, necessitating a multi-layered approach to mitigate risks effectively. The physical security of model files is critical, requiring stringent access controls and attack prevention solutions. This paper proposes a novel solution architecture that protects the model architecture and weights from attacks by using Moving Target Defense (MTD), which obfuscates the model, preventing unauthorized access, and enabling detection of changes to the model. Our method is shown to be effective at detecting alterations to the model, such as steganography; it is faster than encryption (0.1 seconds to obfuscate vs. 18 seconds to encrypt for a 2500 MB model), and it preserves the accessibility of the original model file format, unlike encryption. Finally, our code is available at https://github.com/ArielCyber/AI-model-MTD.git.**

I. Introduction

THE swift evolution of Artificial Intelligence (AI) technology has made it a top priority for cybercriminals looking to obtain confidential information and intellectual property or cause damage in different ways. These malicious individuals may try to exploit AI systems for their own gain, using specialized tactics alongside conventional IT methods. In particular, attackers may wish to gain access to a trained AI model for any number of reasons: to exfiltrate valuable data from victims, including stealing trained AI model data, to tamper with the model, for example, implanting backdoors [1], etc. Given the broad spectrum of potential attack strategies, safeguards must be extensive.

While adversarial AI model security [2], [3], privacy [4] and operational security aspects of AI receive much attention [5], [6], it's equally important to address the physical file security aspects of AI models. In this day and age, an AI model is a valuable asset; training AI models is a long and extremely expensive process, hence, the resulting trained models are the sole artifacts that carry the effort put into the training process.

Figure 1 shows the simplified AI model architecture. Inside the model, we have the architecture metadata (in some cases), model weights (numeric parameters such as weights/biases), and optional metadata such as labels and notes. The model weights essentially carry the frozen state of the model after training, they represent the model's "knowledge" of the task it was trained on. Therefore, deploying trained AI models necessitates protecting the model file, and in particular, the trained weights. The model might be exposed to danger in different scenarios: in training, transit, deserialization, and inference.

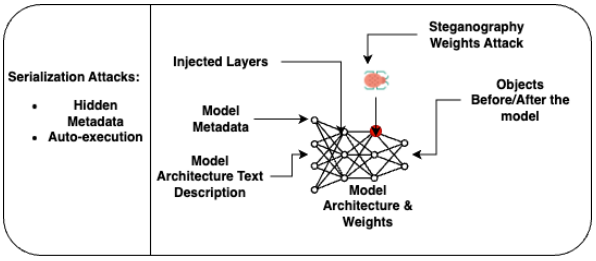


Fig. 1. Simplified AI model file architecture, and serialization attacks and steganography attacks that utilize the AI model file.

The National Security Agency (NSA) recently released a Cybersecurity Information Sheet (CSI) [5]. The CSI is intended to support National Security System owners and Defense Industrial Base companies deploying and operating AI systems designed and developed by an external entity. The CSI illustrates the steps to protect the deployment and two main points directly related to this work. The first is to validate and test the AI model. The second is protecting AI model weights. The NSA CSI suggests the following protection steps:

- Harden interfaces for accessing model weights to increase the effort it would take for an adversary to exfiltrate the weights.
  - Implement hardware protections for model weight storage, disable unnecessary hardware communication capabilities, and protect against emanation or side-channel techniques.
  - Aggressively isolate weight storage. For example, store model weights in a protected storage vault or enclave

The Open Web Application Security Project (OWASP) has been at the forefront of web application security for years. Recognizing the burgeoning significance of machine learning, OWASP introduced the "Top 10 Machine Learning Risks"

[7], a compilation that categorizes and outlines the most prevalent threats targeting AI systems, specifically about the content of the model and the threats around it. Examples of the discussed attacks are inference attacks and poisoning attacks. Of all the discussed threats, the AI Supply Chain Attacks are the most relevant to this work, while the other attacks are related to the content and integrity of the model. In the AI supply chain attack use case, an attacker modifies or replaces a machine learning library or model used by a system. This can also include the data associated with the machine learning models. MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™) [8] is a comprehensive knowledge base. It's designed to document adversary tactics, techniques, and case studies pertinent to machine learning (ML) systems. These details are amassed from real-world observations, demonstrations by ML red teams and security groups, and academic research. In terms of this research, ATLAS helps to understand the attack surface, but it is not a cyber solution. This work is positioned in MITRE ATLAS as a solution against user execution of unsafe ML Artifacts (AML.T0011.000) [9].

While growing, the literature on physical AI model security is still in its early stages [10], [11], [12], [13]. As the adoption of AI and Machine Learning (ML) technologies continues to rise, the need for robust security measures becomes increasingly critical. This work focuses on providing a secure solution to train, transfer, and ingest AI models.

In this work, we propose model weight separation based on Moving Target Defense (MTD) of model weights. Our MTD also provides verification and authentication mechanisms to guarantee the model's validity. By adopting an agreed-upon framework of obfuscation of serialized models and verifying models upon deserialization, model receivers can ensure they ingest models safely, whether as part of a supply chain or directly using the model.

The paper is organized as follows: Section II summarizes the paper's contribution. Section III reviews the related work. Section IV presents the MTD solution architecture. Section V describes our threat model. Section VI offers a detailed description of our dataset. In Section VII, we present our research evaluation and validation methodology. Subsequently, in Section VIII, we discuss our limitations. Finally, Section IX provides our conclusions.

II. Contribution

This work reviews the state-of-the-art physical Artificial Intelligence (AI) model security solutions. For the first time, as far as we know, we suggest a novel zero-trust solution based on MTD for AI model weights.

- We suggest a novel zero-trust MTD solution designed to bolster the robustness and security of AI models against potential threats. This approach ensures protection throughout the entire model life cycle, from its initial training phase to its final distribution. Consequently, it provides a strong defense against physical file-based attacks on the model weights and prevents unauthorized use of the trained model by obfuscating the model weights.
- We fully publish our code [14].

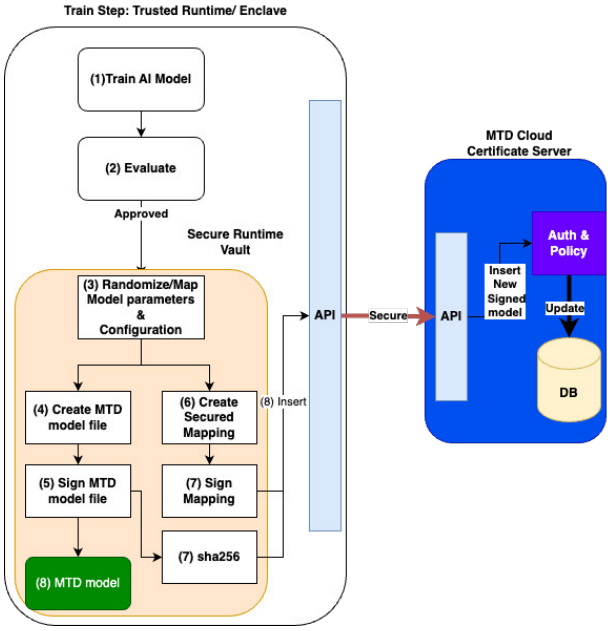


Fig. 2. MTD model creation inside a secure enclave

III. Related Work

Moving Target Defense (MTD) [15] is a proactive cybersecurity strategy that increases the complexity and cost for attackers by continuously changing the attack surface. Unlike traditional defense mechanisms, MTD introduces variability and unpredictability into the system, making it more challenging for attackers to exploit vulnerabilities. MTD encompasses techniques like IP address hopping [16], port rotation [16], virtual machine migration [17], code diversification [18], software obfuscation [19], memory randomization [20], and data randomization [21]. It is helpful in network and software security and represents a paradigm shift in cybersecurity by emphasizing dynamic and adaptive defense mechanisms. Our AI model MTD uses model weight randomization and reconstruction to prevent all model weight attacks and report an attempt of an attack.

IV. Architecture & Solutions

In this section, we describe our MTD methodology, which provides comprehensive protection against physical attacks on the model weights and prevents unauthorized use of the trained model. It is used to save and load models.

The architecture is divided into two parts: MTD creation (Fig. 2) and MTD loading (Fig. 3). If the model is not MTD protected, a fallback step optionally uses Content Disarm and Reconstruction (CDR) methods that were discussed in previous works [22] (Fig. 3).

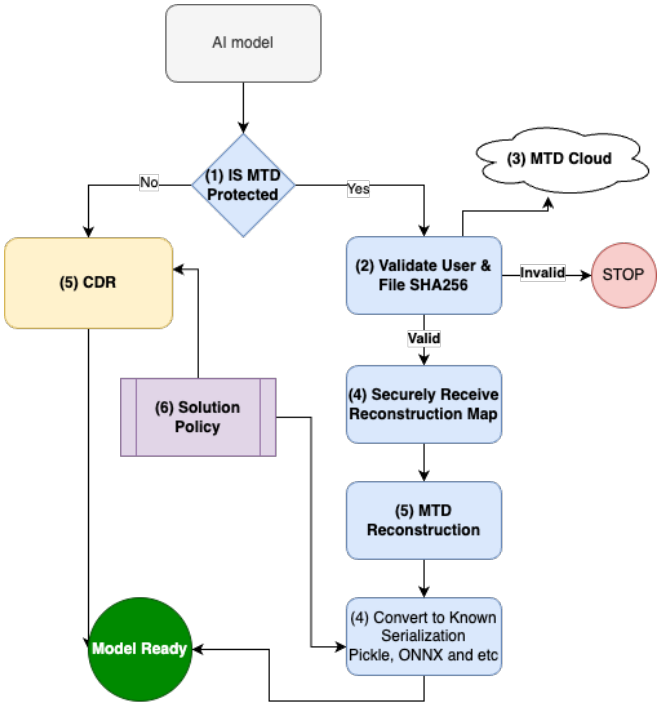


Fig. 3. MTD model loading and fallback to CDR

The model training process powered by MTD is demonstrated in Fig. 2 and occurs within a secure runtime environment or enclave [23]. Once the training is complete (1) and model evaluation passes (2), the MTD process (3) begins. The NSA [5] recommends separating the model and its weights. Therefore, the MTD algorithm randomizes the model weights, separates the model mapping from the randomized weights, and prevents unauthorized use of the trained model. To provide additional security, the MTD model's weights can be encrypted. This produces a file (4) that is protected by MTD. We sign and send this file to the cloud (5, 8) along with a mapping (6). We also sign and store the mapping in the cloud (6, 7). To further protect the model, upon saving the MTD model, we save a sha256 hash of the model weights and verify the hash on load.

Figure 3 shows how the MTD-protected model is loaded. First, it checks (1) if the model is MTD protected. If it's not, CDR (5) is used instead. If the model is protected, it's validated (2) with the MTD cloud API (3). If any modifications are detected, an alert is sent, and the process is stopped. If the model is valid, the mapping is received, and the model is reconstructed. Once the model is reconstructed, it can be saved to a file or kept in memory. It's important to note that this process assumes that if the user is compromised, it's already too late, and the focus is on securing the model and not the device. AI inference can also run inside a secure enclave. Our future work will focus on securing AI PC, which is expected to be the future of running large language models by utilizing a Neural Processing Unit (NPU) (see [24]).

V. Threat Model

To ground our security analysis, we first formalise the adversary we seek to resist, the assumptions we place on system components, and the boundaries of our evaluation. The threat model presented here guides the design choices in Section IV and underpins the evaluation in Section VII.

A. System Participants and Assets

We consider three roles:

- **Model Owner** - trains the network inside a trusted hardware enclave, applies the proposed Moving-Target Defense (MTD) to obfuscate the weights, and signs the resulting artefacts.
- **Distribution Service** - a cloud repository that stores the MTD-protected weight file and the corresponding reconstruction mapping, exposes a mutually-authenticated API, and returns the mapping only to authorised consumers.
- **Model Consumer** - fetches the protected file, verifies the signature, retrieves the mapping via the API, reconstructs the model, and performs inference.

The primary assets are: (i) the trained weights, (ii) the architecture/mapping that re-orders the weights, and (iii) the Model Owner's signing key.

B. Adversary Capabilities

The adversary ("*A*") is *active* and may:

1) Observe, intercept, replay, or tamper with any file transferred between participants (supply-chain attack, man-in-the-middle, malicious mirror).
2) Obtain *direct* read–write access to any storage location that holds the protected file at rest (e.g., compromise of a consumer endpoint or removable media).
3) Craft and deliver malicious model artefacts that embed steganography, hidden payloads, or altered weights in an attempt to bypass verification.
4) Deny or delay network access to the Distribution Service.

A cannot (by assumption):

- Break standard cryptographic primitives (digital signatures, SHA-256 hashes, authenticated encryption).
- Extract secret keys from the hardware enclave or the Model Owner.
- Compromise the integrity of the Distribution Service's authentication and access-control logic.

C. Trusted Computing Base (TCB)

Our TCB consists of:

- The hardware enclave used for training and MTD obfuscation, including its firmware and attestation chain.
- The Model Owner's private signing key and the correctness of the signature-generation process.
- The Distribution Service's storage and access-control mechanisms (but *not* its surrounding OS or network, which may be hostile).
- Standard cryptographic libraries and their implementations.

TABLE I
SUBSET OF PYTORCH MODEL ARCHITECTURES USED TO VALIDATE THE SUGGESTED MTD METHOD.

Model Type	Model Architecture	#Models	Size [MB]
Vision (Classification)	AlexNet [25]	1	230
	ConvNext [26]	4	110-750
	DenseNet [27]	4	30-110
	EfficientNet [28], [29]	12	20-450
	RegNet [30]	34	20-2460
Vision (Segmentation)	DeepLabV3 [31]	3	40-230
Vision (Detection)	Faster R-CNN [32]	4	70-160
Vision (Video)	SwinTransformer [33]	4	110-360
Text	BERT [34]	8	440-1440
Audio	Wav2Vec [35]	2	380-1270

D. Out-of-Scope Scenarios

The following are expressly excluded from this work:

- **Data poisoning, backdoor insertion during training, or adversarial-example attacks at inference time.** These alter model *behaviour* rather than the on-disk artefact and are orthogonal to file-level protection.
- **Full host compromise *before* MTD creation** (if the training environment is already under attacker control, any file- level defence is moot).

VI. DATASET

For evaluation of our suggested method, we create 2 datasets: benign AI models, and AI models attacked with steganography.

Dataset 1 - Benign PyTorch models: A dataset of benign pre-trained PyTorch models. We compile this dataset using the torchvision and HuggingFace APIs. Table I contains a subset of the pre-trained model architectures we used. In total, this dataset contains 129 vision, text, and audio models. We use this to validate our method on benign data, and it is used as a base for creating a dataset of models attacked with steganography (dataset 2) for further evaluation.

Dataset 2 - Attacked models for MTD method evaluation: We use LSB model weight steganography attacks [36] from our previous work on the benign models from dataset 1 to validate the MTD verification, ensuring that the MTD process works as intended with untempered models and also that the method successfully detects changes (i.e., physical attacks) in the model as intended.

VII. EVALUATION

This section evaluates our proposed AI model MTD method. The experiments are designed to prove that the method has practical value for protecting AI model ecosystems from physical attacks such as data exfiltration or unauthorized access to the trained AI model.

Experimental Setup: All experiments are run on a Ubuntu Linux 24.04 server, equipped with an Intel(R) Xeon(R) w5-2445 CPU (20 cores @ 3.10GHz), and 128 GB RAM.

Baseline Method: In terms of obfuscating the model to prevent unauthorized access, the immediate solution is to encrypt the serialized payload before transmission. This is also suggested by the NSA CSI [5] to ensure security of the model files. In our experiments, we compare our suggested MTD method to the Fernet encryption algorithm in terms of runtime. The MTD approach has the property of keeping the model file accessible, i.e., recognizable as the original model file, but the trained state is only available to the rightful owners of the model. This can be a benefit, for example, in scenarios such as model hubs that index the model and give an overview of its metadata, like HuggingFace does; with encryption, this won't be possible.

A. Experimental Results

We evaluate the MTD method for protecting training models with 2 main concerns: validity and runtime. Validity means we assert that the method prevents the attacks it's supposed to prevent and that the models stay intact after the whole process. We measure runtime to analyze the time overhead that the MTD methods cause. On the model sender side, there are the obfuscation and saving (serialization) methods, and on the model receiver side, there are the loading (deserialization) and the deobfuscation methods. We created automated Quality Assurance (QA) tests to check the method's validity. The tests check the following:

1) MTD model is the same after obfuscation and deobfuscation.
2) MTD model is the same after saving and loading.
3) Attack simulation: An MTD Model is constructed, obfuscated, and saved. Then the model weights are attacked, and we assert that the change is detected upon load.

For steps 1 and 2, we use the benign PyTorch models dataset (dataset 1, see Section VI); for step 3, we use dataset 2. All tests passed. Figure 4 plots the mean runtime of saving and loading MTD functions on all models in dataset 1, repeated 10 times. Additionally, we compare the runtimes to regular saving and loading to quantify the additional overhead our suggested method adds. Saving and loading were done with a virtual file since we are only concerned with the runtime overhead our method adds. We can see our MTD method adds approximately 4 seconds of runtime overhead to the serialization of a 2500 MB model; we feel this is within reason. Additionally, looking at Table II, we can see MTD obfuscation/deobfuscation has almost no runtime overhead, while the alternative method of encrypting the payload has a significant runtime overhead (about 17 seconds for a 2500 MB model). Therefore, in total, choosing our MTD method will be faster than using encryption. Finally, the MTD deserialization method has nearly identical runtime to regular runtime. In conclusion, the suggested MTD method is faster than encryption, especially as model size scales, and it also keeps the model file accessible in terms of metadata, structure, etc., which can be useful for sites like HuggingFace that give high-level overviews based on the model metadata.

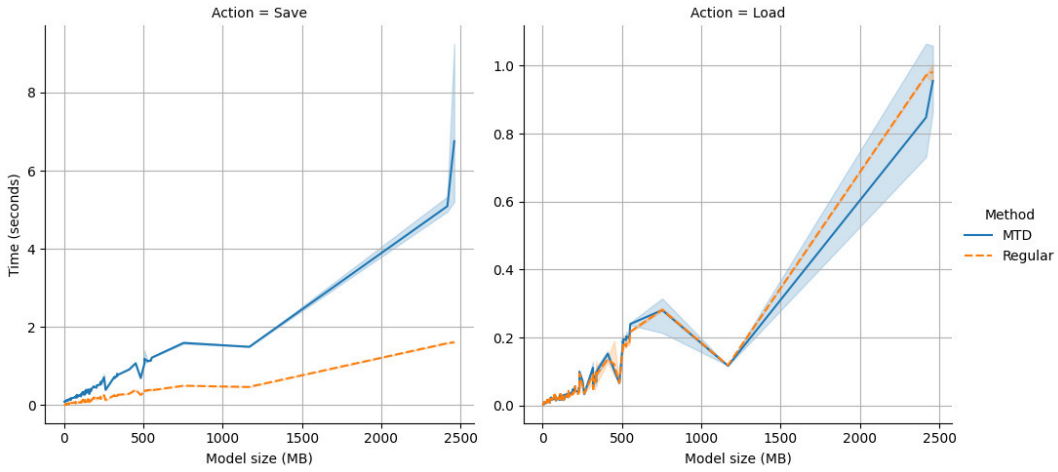


Fig. 4. Mean time (seconds) measurement vs. model size (MB) of saving/loading MTD actions. We show the MTD variant (solid blue line) and the regular variant (dashed orange line) for save and load. The action procedures were repeated 10 times with each model in the MTD datasets, and we show confidence intervals. We can see a linear increase in serialization (saving) runtime as the model size grows, and MTD has little to no overhead on loading the model after serialization.

TABLE II
RUNTIME (SECONDS) OF SERIALIZATION/DESERIALIZATION USING OUR SUGGESTED MTD METHOD, AND THE BASELINE FERNET ENCRYPTION METHOD.

	Model Serialization		Model Deserialization	
	MTD Obfuscate	Fernet Encrypt	MTD Deobfuscate	Fernet Decrypt
$0 \leq$ Model Size (MB) $<$ 100	**0.003820 ± 0.002000**	0.244292 ± 0.115333	**0.005795 ± 0.003956**	0.245123 ± 0.123005
100 $\leq$ Model Size (MB) $<$ 200	**0.008666 ± 0.002466**	0.680410 ± 0.126848	**0.013184 ± 0.005763**	0.724327 ± 0.133060
200 $\leq$ Model Size (MB) $<$ 300	**0.014490 ± 0.004417**	1.225329 ± 0.231984	**0.021955 ± 0.009834**	1.197840 ± 0.278559
300 $\leq$ Model Size (MB) $<$ 400	**0.017144 ± 0.002250**	1.648033 ± 0.388462	**0.025935 ± 0.002614**	1.746837 ± 0.373579
400 $\leq$ Model Size (MB) $<$ 500	**0.024434 ± 0.006219**	2.898500 ± 0.236313	**0.040424 ± 0.018690**	2.574974 ± 0.215552
500 $\leq$ Model Size (MB) $<$ 600	**0.019980 ± 0.001987**	3.002284 ± 0.217727	**0.021760 ± 0.004777**	2.574076 ± 0.170747
700 $\leq$ Model Size (MB) $<$ 800	**0.037661 ± 0.005977**	5.552645 ± 0.103848	**0.051598 ± 0.006328**	5.999168 ± 0.116049
1100 $\leq$ Model Size (MB) $<$ 1200	**0.051498 ± 0.000437**	7.697417 ± 0.152804	**0.069661 ± 0.002010**	8.562226 ± 0.158626
2400 $\leq$ Model Size (MB) $<$ 2500	**0.109626 ± 0.023599**	17.901734 ± 0.261984	**0.134600 ± 0.019062**	19.233927 ± 0.467096

VIII. LIMITATIONS

The proposed solution introduces a novel architecture for physically securing AI model files from malicious exposure. The architecture prevents all file alterations when the model is secured with the suggested MTD method. This builds upon previous work on CDR against steganography attacks [37] and steganalysis [36] (steganography attack detection). Since AI model file security is still in its early stages, we expect to see more sophisticated attacks and vulnerabilities in this domain. The proposed MTD architecture requires the user to fully use the various AI model functions, such as training and serialization within it, but widely-used methods like SafeTensors [38] do so too. Moreover, the proposed method adds a certain performance overhead that is added to all models since it works in a zero-trust manner. However, the performance overhead is small, as described in the results section. It is important to note that this work focuses solely on the security of the AI model's physical file. Other adversarial attacks exist, such as backdoors in machine learning models [1] and dataset poisoning attacks [39]. Dataset attacks are outside this paper's scope, but future work should view them as part of a holistic platform.

IX. CONCLUSIONS

Our work proposes a new solution for securing AI model files by preventing file modifications and unauthorized access through the use of MTD and authentication. This architecture aims to secure AI model file transfers from malicious attacks that threaten the industry. By incorporating MTD, we follow the NSA's [5] recommendation to separate the model architecture and its weights. This is accomplished by randomizing the model weights and separating the model from the deconstruction mapping. This ensures that only authenticated users can reconstruct and use the model when it is distributed. Based on our experimental results, the proposed architecture provides a 100% attack prevention rate (129/129 tampered models were correctly alerted) under the assumed conditions. In addition, we show that our method is significantly faster than using encryption (0.1 seconds to obfuscate vs. 18 seconds to encrypt

for a 2500 MB model), and contrary to encryption, it also keeps the model file in the original format, which can allow to get a high-level overview of the model, based on the metadata (e.g., number of parameters, etc.). Of course, users can also choose to use encryption on top of our suggested method if they wish to. Our method was evaluated on a wide range of data, which includes diverse benign models and steganography attacks from the literature and real-world scenarios found in HuggingFace [40]

X. Acknowledgement

This work is under US Provisional Patent Application No. 63/536,420

References

[1] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6206–6215.

[2] M. Chen, G. He, and J. Wu, "Zddr: A zero-shot defender for adversarial samples detection and restoration," *IEEE Access*, 2024.

[3] K. Nguyen, T. Fernando, C. Fookes, and S. Sridharan, "Physical adversarial attacks for surveillance: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[4] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.

[5] N. S. Agency, "Deploying ai systems securely, best practices for deploying secure and resilient ai systems," 2024, accessed: 2024-05-01. [Online]. Available: https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF

[6] M. ATLAS, "Mitre atlas," 2024, accessed: 2024-05-01. [Online]. Available: https://atlas.mitre.org/

[7] OWSAP, "Owasp machine learning security top ten," 2023, accessed: 2023-10-15. [Online]. Available: https://owasp.org/www-project-machine-learning-security-top-10/

[8] MITRE, "Mitre atlas," 2023, accessed: 2023-10-15. [Online]. Available: https://atlas.mitre.org/

[9] M. ATLAS, "Mitre atlas, user execution: Unsafe ml artifacts," 2024, accessed: 2024-05-01. [Online]. Available: https://atlas.mitre.org/techniques/AML.T0011.000

[10] R. Dubin, "Disarming attacks inside neural network models," *IEEE Access*, 2023.

[11] E. Sultanik, "Never a dill moment: Exploiting machine learning pickle files," 2022, accessed: 2022-12-19. [Online]. Available: https://blog.trailofbits.com/2021/03/15/never-a-dill-moment-exploiting-machine-learning-pickle-files/

[12] P. Zhou, "How to make hugging face to hug worms: Discovering and exploiting unsafe pickle.loads over pre-trained large model hubs," accessed: 2024-08-01. [Online]. Available: https://i.blackhat.com/Asia-24/Presentations/Asia-24-Zhou-HowtoMakeHuggingFace.pdf

[13] M. Slaviero, "Sour pickles, a serialized exploitation guide in one part," accessed: 2023-05-01. [Online]. Available: https://media.blackhat.com/bh-us-11/Slaviero/BH_US_11_Slaviero_Sour_Pickles_Slides.pdf

[14] D. Gilkarov, "AI-MTD Code Repository," 2025, accessed: 2025-06-04. [Online]. Available: https://github.com/ArielCyber/AI-model-MTD.git

[15] J.-H. Cho, D. P. Sharma, H. Alavizadeh, S. Yoon, N. Ben-Asher, T. J. Moore, D. S. Kim, H. Lim, and F. F. Nelson, "Toward proactive, adaptive defense: A survey on moving target defense," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 709–745, 2020.

[16] V. Heydari, "Moving target defense for securing scada communications," *IEEE Access*, vol. 6, pp. 33 329–33 343, 2018.

[17] M. Azab and M. Eltoweissy, "Migrate: Towards a lightweight moving-target defense against cloud side-channels," in *2016 IEEE security and privacy workshops (SPW)*. IEEE, 2016, pp. 96–103.

[18] M. Styugin, V. Zolotarev, A. Prokhorov, and R. Gorbil, "New approach to software code diversification in interpreted languages based on the moving target technology," in *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, 2016, pp. 1–5.

[19] S. Banescu and A. Pretschner, "A tutorial on software obfuscation," *Advances in Computers*, vol. 108, pp. 283–353, 2018.

[20] G. Mordehai, Y. Elovici, and G. Kedma, "Method and system for protecting computerized systems from malicious code," Jul. 11 2017, uS Patent 9,703,954.

[21] D. Evans, A. Nguyen-Tuong, and J. Knight, "Effectiveness of moving target defenses," *Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats*, pp. 29–48, 2011.

[22] R. Dubin, "Disarming attacks inside neural network models code repository," 2022, accessed: 2023-07-15. [Online]. Available: https://github.com/ArielCyber/AI-MODEL-CDR

[23] Intel, "Reference architecture for privacy preserving machine learning with intel® sgx and tensorflow* serving," accessed: 2023-05-01. [Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/technical/privacy-preserving-ml-with-sgx-and-tensorflow.html

[24] "The ai pc powered by intel is here. now, ai is for everyone." accessed: 2023-05-01. [Online]. Available: https://www.intel.com/content/www/us/en/products/docs/processors/core-ultra/ai-pc.html

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545

[27] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[28] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[29] ——, "Efficientnetv2: Smaller models and faster training," 2021. [Online]. Available: https://arxiv.org/abs/2104.00298

[30] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," 2020. [Online]. Available: https://arxiv.org/abs/2003.13678

[31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1706.05587

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://arxiv.org/abs/1506.01497

[33] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," 2021. [Online]. Available: https://arxiv.org/abs/2106.13230

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[35] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[36] D. Gilkarov and R. Dubin, "Steganalysis of ai models lsb attacks," *IEEE Transactions on Information Forensics and Security*, 2024.

[37] R. Dubin, "Content disarm and reconstruction of steganography malware in neural network models," 2023, accessed: 2022-01-15. [Online]. Available: https://github.com/randubin/CDR-NN

[38] H. Face, "Safetensors," accessed: 2024-08-01. [Online]. Available: https://huggingface.co/docs/safetensors/en/index

[39] J. Zheng, P. P. Chan, H. Chi, and Z. He, "A concealed poisoning attack to reduce deep neural networks' robustness against adversarial samples," *Information Sciences*, vol. 615, pp. 758–773, 2022.

[40] H. Face, "Hugging face model zoo," accessed: 2023-05-01. [Online]. Available: https://huggingface.co/models

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 705–713
DOI: 10.15439/2025F1129
ISSN 2300-5963 ACSIS, Vol. 43

Entangled by Design: A structured Overview of Management Challenges concerning AI Adoption in Organizations

Tim Klos
0009-0003-3041-8432
University of Applied Sciences Augsburg
An d. Hochschule 1
86161 Augsburg, Germany
Email: tim.klos@hs-augsburg.de
University of Applied Sciences Konstanz
Alfred-Wachtel-Str. 8
78462 Konstanz, Germany
Email: tim.klos@htwg-konstanz.de

Alessandra Zarcone
0000-0001-9380-4638
University of Applied Sciences Augsburg
An d. Hochschule 1
86161 Augsburg, Germany
Email: alessandra.zarcone@tha.de

Christopher Rentrop, Constanze Riedinger, Niculin Prinz
0000-0003-3629-3821, 0009-0003-0226-4114, 0000-0002-3656-2668
University of Applied Sciences Konstanz
Alfred-Wachtel-Str. 8
78462 Konstanz, Germany
Email: {christopher.rentrop, constanze.riedinger, niculin.prinz}@htwg-konstanz.de.

Melanie Huber
0000-0001-8020-9055
BITCO³ GmbH
78467 Konstanz, Germany
Email:
melanie.huber@bitco3.com

***Abstract*—Artificial Intelligence (AI) adoption poses complex, multidimensional challenges for organizations that extend beyond technical implementation. While strategic interest in AI is rising, many initiatives struggle to scale sustainably. This paper addresses the fragmented state of AI adoption research by offering a theory-driven synthesis of the key management challenges organizations face. Based on a systematic literature review and grounded in the socio-technical systems (STS) perspective, the study identifies challenges across technological, organizational, and social domains—such as data governance, organizational inertia, skill shortages, and ethical ambiguity—as mutually reinforcing. The resulting framework highlights the systemic entanglement of these barriers, underscoring the limits of isolated interventions. This study contributes to Information Systems (IS) research by conceptualizing AI adoption as a socio-technical transformation and providing an integrative typology of adoption challenges. The findings offer a foundation for future empirical research and guide strategic decision-makers in navigating the organizational complexity of AI integration.**

***Index Terms*—AI Adoption, STS, Organizational Change, AI Challenges, Information Systems**

I. Introduction

ARTIFICIAL Intelligence (AI) has recently evolved rapidly into a strategic priority for organizations across industries. In a 2019 survey, already over 84% of 1,500 surveyed C-suite executives from high-profit international companies across various industries emphasized that AI is essential for achieving their growth targets [1]. Despite this ambition, the majority of organizations continue to face substantial barriers in scaling AI effectively to this day [1], [2], [3]. Many initiatives remain confined to isolated pilot projects, lacking cross-functional integration and strategic alignment [1], [4]. Even as AI tools become more accessible, only a minority of organizations have redesigned their core workflows accordingly, and governance responsibilities often remain unclear or underdeveloped [2]. These discrepancies reveal a deeper issue: AI adoption is not merely a matter of technology implementation but requires ongoing adaptation to complex structural, procedural, and human dynamics. This paper, therefore, aims to identify and systematize the management challenges organizations must overcome to successfully adopt AI, using a socio-technical systems (STS) lens to account for the intertwined nature of these challenges.

Although existing research has addressed various barriers to AI adoption, ranging from technical and infrastructural issues to strategic misalignments or cultural resistance, it has typically examined these challenges in isolation. Several studies investigate specific barriers through expert interviews, systematic literature reviews, or conceptual models (e.g., [5], [6], [7], [8], [9]). However, few contributions adopt an integrative, theory-driven perspective that accounts for the socio-technical entanglement of AI adoption management challenges in organizations. On the other side, scholars increasingly argue that technological, structural, and social dimensions of AI implementation are tightly coupled and mutually reinforcing [9], [10], [11]. Without a holistic understanding of these interdependencies, even advanced AI initiatives risk stagnation or failure to scale sustainably [10], [12]. This paper responds to this gap by proposing a structured, STS-based

synthesis of existing knowledge on AI management challenges in an organizational context. To do so, the paper addresses the following research questions:

RQ1: *What are the management challenges of AI adoption in organizations?*

RQ2: *How does existing literature consider the socio-technical entanglement of the challenges?*

By providing a structured overview of AI adoption challenges from a socio-technical perspective, this paper contributes to Information Systems (IS) research in two ways: first, by enhancing theoretical understanding of AI adoption as an entangled socio-technical process; and second, by offering a systematic categorization of challenges that can support both academic inquiry and practical decision-making. The findings aim to support researchers, strategists, and policymakers in navigating the complex landscape of organizational AI management.

The paper is structured as follows: Section 2 provides the theoretical background on AI adoption and socio-technical systems. Section 3 outlines the methodological approach used to identify and synthesize the relevant challenges. Section 4 presents the results of the systematic analysis. Section 5 discusses the socio-technical interdependencies and implications for research and practice. Section 6 concludes with a summary and outlines directions for future research.

II. Background

AI is broadly understood as the science and engineering of developing intelligent systems capable of replicating or augmenting human cognitive functions such as reasoning, problem-solving, and decision-making [13], [14], [15]. When it comes to technology, AI comprises a diverse set of approaches including machine learning, deep learning, natural language processing, and computer vision [6], [7], [15], [16]. Functionally, AI can be defined as the capacity of systems to autonomously interpret data, learn from experience, and apply that learning to achieve goal-oriented behavior [8], [12]. Due to its general-purpose nature [17], AI is increasingly integrated into organizational routines, enabling automation, augmentation, and innovation across industries. In contrast to traditional IT systems, AI is characterized by three interrelated properties: autonomy, learning, and inscrutability [10]. Autonomy refers to AI's ability to make decisions without human intervention; learning captures the system's capacity to evolve through experience; and inscrutability addresses the opacity of AI decision-making, particularly in complex models such as deep neural networks [12], [18]. These characteristics complicate organizational control and validation practices and contribute to uncertainty in implementation outcomes[19]. As a result, AI systems often act as semi-autonomous agents, embedded in decision-making contexts with significant organizational and societal implications [20].

While the concept of technology adoption has been widely studied in the IS field (e.g., TAM, TOE, DOI), AI adoption presents a new level of complexity [9], [21]. The unique characteristics of AI challenge traditional assumptions regarding transparency, accountability, and control, rendering many existing models insufficient [21]. AI adoption refers not only to the technical integration of AI tools but also to the organizational transformation required to generate sustained business value [9], [12]. This includes the alignment of strategic priorities, development of technical and organizational capabilities, stakeholder coordination, and iterative adaptation of business processes [10], [14]. Several studies have already contributed to identifying and categorizing challenges related to AI implementation, drawing on methods such as qualitative interviews (e.g., [22]), conceptual frameworks (e.g., [14]), and systematic literature reviews (e.g., [8], [19]). These works have yielded valuable insights into key barriers, including data governance, infrastructure limitations, skill gaps, and organizational inertia. Collectively, they have helped establish a rich foundation for understanding critical factors that affect AI adoption. However, despite these advances, most studies tend to examine these challenges in isolation or emphasize predominantly technological or organizational dimensions. While a socio-technical entanglement of these challenges is often implied or acknowledged in principle [12], [18], a systematic and theory-based integration that explicitly maps their interdependencies across functional and structural boundaries remains limited. This fragmented view is problematic. Given the socio-technical nature of AI, challenges cannot be addressed in isolation. A detailed understanding of these challenges supports more effective implementation by revealing organizational gaps and enabling proactive risk mitigation. Yet, there is limited research that systematically examines how challenges from different organizational levels and functional domains interact and reinforce each other [5], [7], [8], [11], [12], including no systematic STS-classification of AI management challenges of AI adoption in organizations.

III. Method

In order to identify and synthesize current academic discourse on challenges in managing AI adoption in organizations, we conducted a concept-centric literature review following the guidelines of [23] and [24]. The objective was to provide a structured overview of the key challenges associated with organizational AI adoption and to analyze their socio-technical interdependencies using a theory-informed framework. We followed the five-stage model of [23] for literature reviews: (1) problem formulation, (2) data collection, (3) data evaluation, (4) analysis and interpretation, and (5) presentation. Stages (1) and (5) are addressed in the introduction and results sections, respectively; this section covers stages (2) through (4).

A. Data collection and evaluation

For our data collection, we use Litbaskets.io, a curated and reproducible search tool optimized for IS research, covering over 1,000 IS-relevant journals indexed in Scopus. This tool allows dynamic filtering of the "medium bucket" of 51 core IS journals and conferences, mitigating risks of over- or under-inclusiveness often associated with traditional search

strategies [25], [26]. To define the search space, we applied the string ("artificial intelligence" OR "AI") AND "manag*" AND "challenge*" across abstracts, titles, and keywords, covering the period from 2000 to January 2025.

The initial result of 196 hits was filtered by applying multiple exclusion criteria. First, duplicates and non-English entries were removed. Second, only papers addressing AI management challenges within an organizational context were considered, excluding papers focused solely on technical or specific domain implementations (e.g., in healthcare, logistics, or mobility domains) without addressing overall organizational implications. Third, the search was extended via forward and backward snowballing, as suggested by [24].

By adopting the described method, a targeted yet inclusive identification of AI-related literature was achieved, ensuring that both high-impact contributions and emerging research trends were systematically captured. At the end of the entire process, 15 publications specifically included challenges of AI management on an organizational level.

B. Data analysis and interpretation

The final corpus of 15 contributions was analyzed using the six-step qualitative thematic analysis procedure from [27]. This approach is well-established in IS research and supports systematic and transparent theme identification. The goal was to extract, classify, and interpret explicit challenges regarding AI management in organizations. We employed a deductive coding strategy guided by the Socio-Technical Systems (STS) model [28], [29], which conceptualizes work systems as interactions between four components: technology, task, structure, and people (cf. Fig. 1).

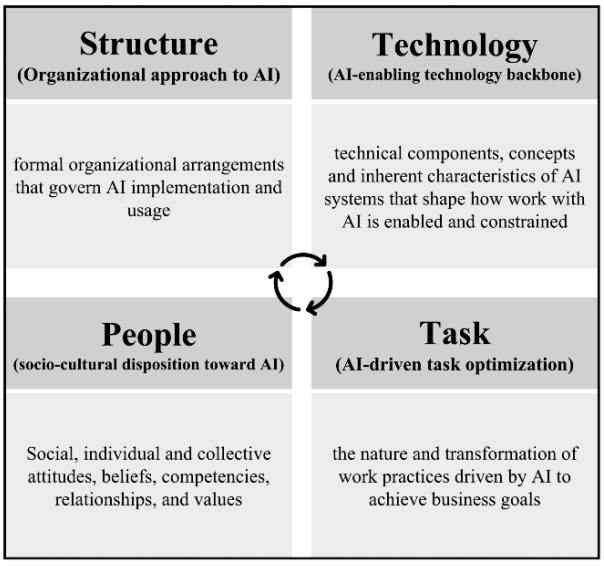


Fig 1 - STS perspectives applied on AI based on [28]

To explain the STS lens for AI adoption, we briefly outline the relevance of each of the four core perspectives. The *technology perspective* addresses the "AI-enabling technology backbone" and therefore refers to the technical components, concepts, and inherent characteristics of AI systems that shape how work is enabled and constrained. This includes, for example, physical infrastructure such as computing power and storage, as well as virtual resources such as pre-trained models, algorithms, training data, or IT-security efforts to keep the new technology safe. These elements determine not only the feasibility of AI deployment but also the system's autonomy, learning capacity, and opacity—factors which are central to how AI affects organizational routines (e.g., [8], [10]).

The *task perspective* considers the "AI-driven task optimization" and captures the nature and transformation of work practices driven by AI. It concerns the activities and processes through which organizational goals are pursued and how these are augmented, redefined, or automated by AI technologies. AI integration can shift task boundaries, introduce new dependencies, or alter performance criteria, thereby affecting both productivity and work design (e.g., [7], [19]). The *structure perspective* consolidates aspects of the "organizational approach to AI", which addresses the formal organizational arrangements that govern AI implementation and use. This includes decision rights, coordination mechanisms, and governance structures that define how responsibilities are distributed, risks are managed, strategies are defined, and AI capabilities are scaled across organizational units. As AI blurs traditional boundaries between human and machine decision-making, structural adaptation becomes a prerequisite for responsible and effective use (e.g., [22], [30]). Finally, the *people perspective* focusing on the "socio-cultural disposition toward AI" focuses on the human actors—such as employees, managers, and other stakeholders—and their perceptions, beliefs, competencies, and values. AI implementation is shaped by individual and collective attitudes, digital skills, trust in AI systems, and willingness to engage with algorithmic processes (e.g., [8], [12]).

This framework allows for theory-informed mapping of challenges and their systemic interrelations. The operationalization of the STS components was informed by prior IS research that applied similar socio-technical logics to technology and platform adoption (e.g., [26], [31]). Coding was conducted manually in Excel using a structured matrix and visualized on a collaboration board. One researcher initially coded all 15 papers, and afterwards the coding was refined across three iterative rounds in expert workshops with another 3 researchers. Disagreements in coding were resolved through discussion within the expert workshops until consensus was reached. During coding, no new categories emerged outside the STS framework, indicating thematic sufficiency. The final coding structure was thus stable and consistent with the four original dimensions. To ensure validity, definitions and representative indicators for each category were documented in a shared collaborative board and reviewed across the rounds.

IV. RESULTS

This section presents the identified management challenges in the literature that organizations face when adopting AI, answering RQ1 and visualizing the interdependencies regarding

RQ2 (cf. Table 1). It is oriented on the four STS model components and their derived categories.

A. Technology

The literature analysis resulted in 10 challenges related to the technology dimension of the STS model clustered by the categories AI security, AI infrastructure, AI characteristics, and data management.

In the context of the category **AI security**, addressing technical vulnerabilities emerged as a critical concern due to the growing deployment of AI systems in safety- and mission-critical domains. A key challenge lies in *mitigating security vulnerabilities* inherent to AI technologies, particularly those resulting from adversarial examples, data poisoning, or insufficient model robustness. These vulnerabilities present new attack vectors and expose AI systems to manipulation or malfunction, potentially leading to catastrophic outcomes in real-world applications [11], [32]. AI models, particularly those based on machine learning, are susceptible to inputs that are deliberately designed to cause misclassification while appearing benign to human observers —for example, when strategically placed stickers on a stop sign lead a vision model to interpret it as a speed limit sign [11]. This complicates the detection and mitigation of adversarial attacks. Furthermore, training data can be subtly manipulated to introduce harmful biases or reduce model performance, underscoring the urgency of robust data validation and attack-resilient learning algorithms [11].

In the category of **AI infrastructure**, *legacy integration* challenges persist as organizations attempt to embed AI within outdated or fragmented IT landscapes. Technical debt, rigid architecture, and incompatible data structures hinder seamless deployment and scalability. Consequently, AI implementation efforts are often delayed or limited in scope due to the inability of legacy systems to support dynamic and data-intensive AI functionalities [7], [9], [12], [16]. Beyond integration, infrastructure resilience has been identified as a prerequisite for the sustainable operation of AI systems [5], [7]. Ensuring resilience entails mitigating system downtimes, enabling rapid recovery, and maintaining consistent performance under stress [11], [16]. AI-specific demands, such as continuous learning and high-volume data processing, impose additional strain on traditional infrastructures [8], necessitating architectural adaptability and robust failover mechanisms [12]. Simultaneously, *ensuring infrastructure resilience* proves to be challenging, as there is a constant need to adapt to an increasingly volatile technological environment. Frequent shifts in tools, platforms, and vendor ecosystems place additional pressure on organizations to maintain flexibility without compromising reliability. A lack of strategic foresight or modularity in AI infrastructure design can result in systemic inefficiencies or incompatibilities when transitioning between rapidly evolving AI ecosystems [8], [9], [14].

The concerns within the category **characteristics of AI** stress the distinct challenges complicating its effective deployment and governance. A central issue arises from *the limited transferability of AI models*. Context-specific training and performance constraints hinder model generalization, thereby requiring resource-intensive retraining and calibration efforts when AI is applied across diverse domains or tasks [8], [10], [11], [16]. Compounding this is the difficulty of *understanding computable human cognition* within organizational settings. AI maturity is often overestimated, and the disparity between perceived and actual capabilities leads to misaligned expectations and strategic missteps in adoption efforts [7], [9], [14], [18], [19], [22]. Another key challenge lies in *handling probabilistic behavior*. AI systems frequently produce non-deterministic outcomes due to inherent stochastic modeling. This unpredictability can undermine user confidence, especially in critical applications requiring consistent and repeatable decisions [8], [11], [12]. Equally critical is the need to *address information processing bias*. Biases stemming from data selection, model assumptions, or systemic societal inequities can propagate through AI outputs, leading to skewed or unfair results that compromise ethical and operational standards [5], [8], [11], [14]. Finally, *handling the inscrutability* of advanced AI systems presents substantial governance risks. Complex architectures such as deep neural networks often lack interpretability, making it difficult to justify decisions or intervene in case of failure. This opacity impedes transparency and undermines accountability structures [8], [12], [14], [32].

Another category of **data management** addresses the management challenge of *ensuring a robust data foundation* as a critical prerequisite for the effective deployment of AI systems. Challenges arise when data is fragmented, incomplete, poorly structured, or inaccessible across organizational silos, thereby limiting model training quality and reliability. These limitations constrain the scalability and performance of AI applications and increase the risk of biased or unstable outcomes [5], [7], [8], [10], [16], [22]. Insufficient governance and weak integration between data sources further aggravate these problems. Inadequate metadata, inconsistent taxonomies, and a lack of data lifecycle management practices impede effective AI use and model interpretability. Moreover, the absence of real-time data pipelines delays responsiveness and undermines decision-making precision in dynamic environments [5], [7], [16]. As a result, building a solid data foundation not only requires a robust technical infrastructure but also strategic alignment between data stewardship, organizational objectives, and AI initiatives. Without such alignment, data management inefficiencies will propagate into model behavior, reducing trust, scalability, and business value across AI-driven systems [8], [10], [22].

B. Task

The literature analysis resulted in 4 challenges related to the task dimension of the STS model, all clustered within the category AI lifecycle management.

Within the category of **AI lifecycle management**, the *identification of valuable AI use cases* represents a persistent challenge. Strategic misalignment, lack of domain understanding, and unclear business objectives often hinder the selection of

use cases with real organizational impact. Without a systematic process to prioritize AI initiatives, resources risk being diverted to projects with limited value creation [5], [7], [8], [10], [12], [14], [30]. Once use cases are identified, *the optimization of AI solution design* becomes critical. Challenges emerge when requirements are poorly captured, stakeholder needs remain unstructured, or technical feasibility is not properly assessed. This leads to mismatches between designed systems and operational realities, reducing solution effectiveness and user acceptance [5], [8], [10], [12], [19], [22]. Equally complex is the *development of AI solutions*, where iterative prototyping, model training, and validation demand high levels of coordination and technical expertise. Challenges arise from data constraints, evolving requirements, and difficulties in transferring conceptual models into deployable systems, often resulting in extended timelines and reduced reliability [10], [11], [12], [14]. The *operationalization of AI solutions* further introduces barriers related to integration, scalability, and governance. Insufficient monitoring, unclear ownership, and a lack of standardized practices for maintenance and update cycles impede sustainable AI operations, placing long-term reliability and compliance at risk [8], [10], [12], [14], [16].

C. Structure

The literature analysis resulted in 10 challenges related to the structure dimension of the STS model clustered within the categories AI strategy, AI readiness, AI risk & compliance.

Within the category of **AI strategy**, the *alignment of strategic AI initiatives across organizational levels* presents a persistent challenge. Discrepancies between top-level vision and operational execution often result in fragmented initiatives, conflicting priorities, and inefficient resource use, undermining enterprise-wide AI coherence and effectiveness [6], [7], [10], [19], [30]. Closely related is the need to *align cross-domain collaboration for AI adoption*. Functional silos and disciplinary boundaries impede the integration of technical, business, and operational expertise, thereby restricting the co-creation of viable AI solutions and slowing organizational learning [5], [7], [9], [10], [18], [30]. Compounding these challenges is the difficulty of *identifying business value from AI integration*. Without clear mechanisms to assess and translate AI capabilities into measurable outcomes, initiatives risk remaining technology-driven rather than impact-oriented, limiting stakeholder support and long-term investment [5], [7], [8], [9], [10], [14]. A further obstacle lies in *clarifying the role of AI within the broader organizational vision*. Ambiguities regarding AI's strategic purpose or future role hinder commitment and direction, leading to ad hoc experimentation rather than coordinated transformation efforts [5], [6], [8], [18], [30]. Finally, the *navigation of resource constraints* remains a cross-cutting issue. Budget limitations, talent shortages, and technical debt constrain the scope, pace, and sustainability of AI initiatives, often forcing organizations to deprioritize foundational efforts in favor of short-term experimentation [5], [7], [8], [9], [14], [16], [19].

In the category of **AI readiness**, *ensuring transparency in information processing for organizational decision-making* has emerged as a critical concern. The opacity of AI-driven analytics impedes trust and accountability, particularly when decision outcomes cannot be traced back to interpretable data logic or model behavior, thereby limiting managerial acceptance and regulatory compliance [8], [11], [12], [14]. A further barrier is the *insufficient understanding of the organizational AI status quo*. Many firms lack a clear assessment of their current AI maturity, adoption level, and capabilities, which hinders strategic planning and investment prioritization, while also impairing the identification of internal gaps and alignment needs [7], [9], [14], [18], [19], [22]. Moreover, *managing the disruption caused by AI* constitutes a central readiness challenge. The transformation of workflows, redefinition of roles, and shift in decision authority induced by AI adoption frequently trigger organizational resistance, role ambiguity, and cultural friction, thereby complicating effective change management and sustainable implementation [5], [8], [12], [14], [18], [19].

Within the category of **risk and compliance**, the *management of AI-related risks* has emerged as a core organizational concern. Especially, in high-stakes environments, the inability to anticipate or explain AI behavior under attack undermines trust and usability. Therefore, increasing emphasis is placed on developing risk-aware AI systems capable of assessing, quantifying, and managing the multifaceted security threats posed by intelligent algorithms [11], [32]. Additionally, the probabilistic nature of AI systems, combined with limited transparency and unpredictable interactions, generates uncertainties that may result in unintended consequences, undermining both operational reliability and stakeholder trust [5], [8], [11], [16]. Closely connected to risk management is the challenge of *ensuring AI compliance*. Regulatory frameworks also increasingly demand transparency, fairness, and accountability in AI systems, yet technical opacity and unclear governance structures hinder compliance with standards such as the GDPR or emerging AI-specific regulations like the European AI Act [5], [8], [11], [12].

D. People

The literature analysis resulted in 3 challenges related to the people dimension of the STS model clustered within the categories AI competence development and AI workforce transformation.

In the category of **AI competence development**, *overcoming a lack of organizational AI competence* has been identified as a fundamental barrier to successful adoption. Deficits in technical skills, strategic understanding, and interdisciplinary collaboration inhibit the effective design, implementation, and governance of AI systems, leading to misaligned initiatives and underutilized potential [8], [9], [14], [19].

Looking at the category of **workforce transformation**, *facilitating human-AI collaboration* presents a significant challenge. Unclear role definitions, insufficient interface design, and a lack of trust impede effective interaction between employees and AI systems. Without structured integration and mutual augmentation, the potential of hybrid intelligence remains largely underexploited [6], [8], [12], [19], [22]. Moreover, *overcoming cultural distance toward AI* constitutes a critical barrier to transformation. Deep-seated skepticism, fear of replacement, and resistance to change create a climate of mistrust that slows down adoption and undermines engagement. Addressing these tensions requires not only communication and training, but also a shift in organizational mindset [5], [7], [8], [9], [18], [22].

TABLE II.
CHALLENGES OF MANAGING AI ADOPTION IN ORGANIZATIONS FROM A TECHNOLOGY SYSTEM PERSPECTIVE

STS Perspective	Category	Challenges	15 contributions resulting from the literature review														
			Alsheibani et al. (2020) – [Ref. 9]	Asatiani et al. (2021) – [Ref. 12]	Benbya et al. (2020) – [Ref. 14]	Borges et al. (2021) – [Ref. 6]	Dwivedi et al. (2021) – [Ref. 8]	Enholm et al. (2022) – [Ref. 5]	Hansen et al. (2024) – [Ref. 7]	Holmström (2022) – [Ref. 18]	Jan et al. (2023) – [Ref. 16]	Jöhnk et al. (2021) – [Ref. 22]	Kučević and Brandes (2025) – [Ref. 30]	Lee et al. (2023) – [Ref. 19]	Wang and Siau (2019) – [Ref. 32]	Weber et al. (2023) – [Ref. 10]	Zhang et al. (2022) – [Ref. 11]
TECHNOLOGY	AI Security	*Mitigate AI Security Vulnerabilities*													X		X
	AI Infrastructure	*Overcome Legacy Integration*	X	X					X		X						
		Ensure Infrastructure Resilience		X				X	X		X						X
		Manage rapid AI Evolution	X		X		X										
	AI Characteristics	*Manage limited AI Model Transferability*					X				X					X	X
		Understand computable Human Cognition			X	X	X	X									
		Handle Information Processing Bias			X		X	X									X
		Handle Probabilistic Behavior		X			X										X
		Handle Inscrutability		X	X		X								X	X	
	Data Management	*Ensure robust Data Foundation*					X	X	X		X	X				X	
TASK	AI Lifecycle Management	*Align AI Use Case Derivation*		X	X		X	X	X				X			X	
		Handle AI Solution Design		X			X	X				X		X		X	
		Handle AI Development Phase		X	X											X	X
		Handle AI Operations Phase		X	X		X				X					X	
STRUCTURE	AI Strategy	*Align AI Strategy across organizational Levels*				X			X				X	X		X	
		Align cross-domain collaboration for AI Adoption	X					X	X	X			X			X	
		Identify Business Value from AI Integration	X		X		X	X	X							X	
		Clarify Role of AI (Vision)				X	X	X		X			X				
		Navigate Resource Constraints	X		X		X	X	X		X			X			
	AI Readiness	*Ensure transparency in AI-driven decision-making*	X		X		X										X
		Understand organizational AI Status Quo	X		X				X	X		X		X			
		Manage Disruption caused by AI		X	X		X	X		X				X			
	AI Risk & Compliance	*Manage AI Risks*					X	X			X						X
		Manage AI Compliance		X			X	X									X
PEOPLE	AI Competence Development	*Overcome lacking AI Competence*	X		X		X							X			
	AI Workforce Transformation	*Facilitate Human-AI Collaboration*		X		X	X					X		X			
		Overcome Cultural Distance toward AI	X				X	X	X	X		X					

V. Discussion

This section reflects on the findings, considering the two guiding research questions. First, we revisit RQ1 and summarize the identified barriers. Second, we address RQ2 by evaluating the extent and limitations of current research through the lens of the STS framework.

While the challenges identified—such as data bias [11], legacy infrastructure limitations [10], unclear strategic anchoring [30], or persistent skill gaps [7]— on their own are well-documented in the literature, their significance shifts when examined through a structured sociotechnical (STS) lens. The findings support and extend prior research by demonstrating that these issues do not merely coexist but are systematically interlinked across organizational levels. This multi-level mapping reveals not just the breadth but the vertical distribution of management challenges in the context of AI adoption. For instance, earlier studies have pointed to a misalignment between strategic intent and operational realization [7], whereas now results show how such misalignment frequently stems from disconnects between abstract ambitions and the concrete readiness of middle and technical layers. Moreover, it also becomes visible that, for example, technical-level data issues—such as infrastructure bottlenecks or lack of clean training data—can reverberate upward, fostering cultural skepticism toward AI and weakening governance capacity [12]. Although certain challenge areas are known previously from other contexts, our observations align with the broader literature, which increasingly acknowledges the systemic nature of AI adoption challenges. For example, [18] emphasizes that organizational readiness must be understood as a socio-technical configuration, not a purely technological state. Similarly, [19] stresses the need for integrated frameworks that bridge organizational, technical, and human dimensions. Our findings offer empirical reinforcement of these positions by illustrating how misalignments are often not incidental but structurally embedded, emerging from insufficiently coordinated transformation efforts across different layers of abstraction.

The second RQ examined how literature addresses the sociotechnical entanglement of AI management challenges. Our analysis shows that while many studies acknowledge interdependencies between technological, organizational, and human factors, few analyze them through a theory-informed lens. The application of the STS framework has demonstrated value not as a descriptive taxonomy but as a diagnostic tool. By revealing how challenges in one dimension—e.g., *inscrutability* (Technology)—can constrain *transparency* (Structure) and foster *mistrust* (People), the analysis substantiates the STS proposition that AI adoption is inherently a system-wide transformation rather than a technical integration exercise. The study confirms through the STS view that managing one subsystem effectively requires anticipating feedback from others. This insight also resonates with current literature from a perspective of organizational change management [33], which also posits that the four STS core components must be kept in dynamic balance for change efforts to succeed. Imbalances or unilateral interventions in one domain are likely to trigger unintended effects in others. Thus, managing AI requires more than technical readiness—it demands systemic alignment [34]. Nonetheless, the review also reveals blind spots. Some entanglements are better theorized than others. Vertical interconnections (e.g., from vision to implementation) are rarely articulated, pointing to a need for longitudinal research. Horizontally, identical challenges—such as AI governance—are framed differently depending on disciplinary or epistemic perspectives. While this study offers a structured and timely synthesis of AI adoption challenges, it has certain limitations. The final analysis draws on a relatively small set of 15 peer-reviewed papers, which—despite a broad search window (2000–2025)—mostly stem from recent years. This likely reflects the evolving discourse, shaped by the rise of generative AI and increasing accessibility. Still, the limited sample and primarily descriptive approach call for further empirical validation and broader literature inclusion. Future research should explore to what extent challenges are recognized across domains (e.g., medicine, agriculture) and functions (e.g., supply chain, finance), and how IS research can address such pluralism without fragmenting. Several implications emerge from this analysis:

Theoretical integration: To validate this work's diagnostic potential, future research should examine cross-dimensional dependencies in organizational contexts. This aligns with recent calls for theoretically grounded and empirically tested models of AI readiness and maturity (e.g., [7], [18])

Methodological pluralism: To address the complexity and context-specific nature of sociotechnical entanglements, we advocate mixed-method designs—e.g., comparative case studies combined with systems modeling. This reflects broader demands in AI research to integrate qualitative depth with quantitative generalizability (e.g., [8], [19]).

Constructive interdisciplinarity: Management challenges in AI adoption often reflect the disciplinary lens through which they are viewed. Future work should distinguish between "perspective-conditioned" and universally critical challenges. This supports calls for integrative frameworks bridging technological, organizational, and societal viewpoints (e.g., [6]).

Governance-focused modeling: Our results suggest that STS-informed maturity or capability models can help assess AI readiness systemically, beyond narrow metrics of technology or compliance. The need for such governance-oriented frameworks has also been underscored in recent work on sociotechnical envelopment [12], organizational AI capabilities [10], and risk-aware AI systems [11].

In summary, the recent literature increasingly acknowledges AI as a socio-technical transformation but stops short of treating it as such in analytical practice. Our study bridges that gap and opens a trajectory for more systematic analyses and empirical verifications.

VI. Conclusion

This study examined AI adoption challenges in organizations through the socio-technical systems (STS) lens. Addressing two core research questions, it synthesized fragmented literature and revealed interrelations across the STS dimensions of Technology, Task, Structure, and People.

To advance AI operationalization amid rising regulatory and strategic demands, this study is both timely and necessary. It shows that AI adoption is not merely technical, but a socio-organizational transformation. Key challenges do not occur in isolation—they reinforce each other. The study's novelty lies in its theory-driven, cross-dimensional mapping of entangled AI challenges. Unlike prior work treating obstacles as standalone, this paper highlights how misalignments across layers can hinder adoption. It thus extends the IS discourse beyond techno-centric models. The findings address multiple audiences: researchers gain a conceptual framework for empirical validation, practitioners a diagnostic lens for identifying coordination gaps, and policymakers criteria to expand governance beyond technical performance toward structural and cultural readiness.

Future research should empirically validate the proposed framework across sectors through case studies, develop STS-based maturity models for AI readiness, and explore governance mechanisms that account for cross-dimensional dependencies. These steps are critical for ensuring that AI is not just implemented—but implemented responsibly, strategically, and sustainably.

Acknowledgment

This research was partially funded by the Bavarian State Ministry for Science and the Arts (StMWK: Bayerische Staatsministerium für Wissenschaft und Kunst- StMWK) as part of the Project "CHIASM" (Chancenreiche industrielle Anwendungen für vortrainierte Sprachmodelle).

References

[1] Accenture, "AI: Built to Scale – Achieving High Performance with AI at Scale," Accenture, 2019.

[2] A. Singla, A. Sukharevsky, L. Yee, M. Chui, and B. Hall, "The State of AI: How Organizations are Rewiring to Capture Value," McKinsey & Company, 2025.

[3] N. Maslej, "The AI Index 2025 Annual Report," Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford, CA, Apr. 2025.

[4] World Economic Forum and Deloitte, "AI in Action: Beyond Experimentation to Transform Industry," World Economic Forum, Genf, Jan. 2025.

[5] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial Intelligence and Business Value: a Literature Review," *Inf Syst Front*, vol. 24, no. 5, pp. 1709–1734, Oct. 2022, doi: 10.1007/s10796-021-10186-w.

[6] A. F. S. Borges, F. J. B. Laurindo, M. M. Spínola, R. F. Gonçalves, and C. A. Mattos, "The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions," *Int J Inf Manage*, vol. 57, 2021, doi: 10.1016/j.ijinfomgt.2020.102225.

[7] H. F. Hansen, E. Lillesund, P. Mikalef, and N. Altwaijry, "Understanding Artificial Intelligence Diffusion through an AI Capability Maturity Model," *Inf. Syst. Front.*, 2024, doi: 10.1007/s10796-024-10528-4.

[8] Y. K. Dwivedi *et al.*, "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int J Inf Manage*, vol. 57, 2021, doi: 10.1016/j.ijinfomgt.2019.08.002.

[9] S. Alsheibani, C. Messom, Y. Cheung, and M. Alhosni, "Artificial Intelligence Beyond the Hype: Exploring the Organisation Adoption Factors," *Australasian Conference on Information Systems*, 2020.

[10] M. Weber, M. Engert, N. Schaffer, J. Weking, and H. Krcmar, "Organizational Capabilities for AI Implementation—Coping with Inscrutability and Data Dependency in AI," *Inf. Syst. Front.*, vol. 25, no. 4, pp. 1549–1569, 2023, doi: 10.1007/s10796-022-10297-y.

[11] X. Zhang, F. T. S. Chan, C. Yan, and I. Bose, "Towards risk-aware artificial intelligence and machine learning systems: An overview," *Decis Support Syst*, vol. 159, 2022, doi: 10.1016/j.dss.2022.113800.

[12] A. Asatiani *et al.*, "Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems," *JAIS*, vol. 22, no. 2, pp. 325–352, 2021, doi: 10.17705/1jais.00664.

[13] S. AlSheibani, Y. Cheung, Y. Cheung, and C. Messom, "Re-thinking the Competitive Landscape of Artificial Intelligence," *HICSS*, 2020.

[14] H. Benbya, T. H. Davenport, and S. Pachidi, "Artificial Intelligence in Organizations: Current State and Future Opportunities," *SSRN Journal*, 2020, doi: 10.2139/ssrn.3741983.

[15] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, pp. 63–71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[16] Z. Jan *et al.*, "Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities," *Expert Sys Appl*, vol. 216, 2023, doi: 10.1016/j.eswa.2022.119456.

[17] S. Raisch and S. Krakowski, "Artificial Intelligence and Management: The Automation–Augmentation Paradox," *AMR*, vol. 46, no. 1, pp. 192–210, Jan. 2021, doi: 10.5465/amr.2018.0072.

[18] J. Holmström, "From AI to digital transformation: The AI readiness framework," *Business Horizons*, vol. 65, no. 3, pp. 329–339, May 2022, doi: 10.1016/j.bushor.2021.03.006.

[19] M. C. M. Lee, H. Scheepers, A. K. H. Lui, and E. W. T. Ngai, "The implementation of artificial intelligence in organizations: A systematic literature review," *Inf Manage*, vol. 60, no. 5, 2023, doi: 10.1016/j.im.2023.103816.

[20] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons*, vol. 61, no. 4, pp. 577–586, Jul. 2018, doi: 10.1016/j.bushor.2018.03.007.

[21] P. Mikalef and M. Gupta, "Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance," *Information & Management*, vol. 58, no. 3, p. 103434, Apr. 2021, doi: 10.1016/j.im.2021.103434.

[22] J. Jöhnk, M. Weißert, and K. Wyrtki, "Ready or Not, AI Comes—An Interview Study of Organizational AI Readiness Factors," *Bus Inf Syst Eng*, vol. 63, no. 1, pp. 5–20, Feb. 2021, doi: 10.1007/s12599-020-00676-7.

[23] H. M. Cooper, "Scientific Guidelines for Conducting Integrative Research Reviews," *Review of Educational Research*, vol. vol. 52, pp. 291–302, 1982, doi: 10.3102/00346543052002291.

[24] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly*, vol. 26, no. 2, pp. xiii–xxiii, 2002, doi: 10.2307/4132319.

[25] S. Boell and B. Wang, "www.litbaskets.io, an IT Artifact Supporting Exploratory Literature Searches for Information Systems Research," in *Proceedings of the 30th Australasian Conference on Information Systems (ACIS 2019)*, Perth, Australien: Australasian Association for Information Systems, 2019, pp. 1–12. Accessed: Feb. 20, 2025. [Online]. Available: https://aisel.aisnet.org/acis2019/71/

[26] S. Kass, S. Strahringer, and M. Westner, "Drivers and Inhibitors of Low Code Development Platform Adoption," in *2022 IEEE 24th Conference on Business Informatics (CBI)*, Amsterdam, Netherlands: IEEE, Jun. 2022, pp. 196–205. doi: 10.1109/cbi54897.2022.00028.

[27] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

[28] R. P. Bostrom and J. S. Heinen, “MIS Problems and Failures: A Socio-Technical Perspective, Part II: The Application of Socio-Technical Theory,” *MIS Quarterly*, vol. 1, no. 4, p. 11, Dec. 1977, doi: 10.2307/249019.

[29] R. P. Bostrom and J. S. Heinen, “MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes,” *MIS Quarterly*, vol. 1, no. 3, p. 17, Sep. 1977, doi: 10.2307/248710.

[30] E. Kučević and J. J. Brandes, “Guidance for Formulating an AI Strategy: A Taxonomy for Organizational Strategic Perspectives Towards AI,” presented at the Hawaii International Conference on System Sciences, 2025. doi: 10.24251/HICSS.2025.724.

[31] N. Prinz, M. Huber, C. Riedinger, and C. Rentrop, “Two Perspectives of Low-Code Development Platform Challenges – An Exploratory Study,” 2022.

[32] W. Wang and K. Siau, “Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda,” *J. Database Manage.*, vol. 30, no. 1, pp. 61–79, 2019, doi: 10.4018/JDM.2019010104.

[33] J. G. March, Ed., *Handbook of organizations*, 1 vols. Abingdon, Oxon: Routledge, 2013.

[34] C. Riedinger, M. Netscher, and S. Zimmermann, “Organizational Capabilities for Business-IT Integration in Digital Enterprises,” presented at the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, Oct. 2024, pp. 493–500. doi: 10.15439/2024F9800.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 715–720
DOI: 10.15439/2025F2449
ISSN 2300-5963 ACSIS, Vol. 43

An application of BURA solver to fractional super-diffusion problems

Nikola Kosturski, Ivan Lirkov
0000-0002-2215-9835, 0000-0002-5870-2588
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Acad. G. Bonchev, bl. 25A
1113 Sofia, Bulgaria
Email: kosturski@parallel.bas.bg, ivan.lirkov@iict.bas.bg
http://parallel.bas.bg/~kosturski/, http://parallel.bas.bg/~ivan/

Marcin Paprzycki
0000-0002-8069-2152
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw, Poland,
Email: paprzyck@ibspan.waw.pl
http://www.ibspan.waw.pl/~paprzyck/

***Abstract*—In this contribution, the numerical solution of the spectral fractional elliptic equation with power $\alpha \in (1,2)$ is studied. After discretization, the problem is reduced to solving a system of linear algebraic equations. We apply the Best Uniform Rational Approximation (BURA) method and focus on the numerical aspects, related to its implementation. An extensive experimental study is conducted to evaluate the accuracy of the proposed approach. The numerical results confirm the theoretical analysis and demonstrate the effectiveness of using the BURA method for the fractional super-diffusion problems.**

I. Introduction

LET US CONSIDER the spectral fractional elliptic equation with power $\alpha \in (1,2)$, given by

$$\mathcal{A}^\alpha u(x) = f(x), \tag{1}$$

where $\mathcal{A}$ is a self-adjoint elliptic operator in Ω, satisfying homogeneous Dirichlet boundary conditions. The non-local operator $\mathcal{A}^\alpha$ is defined via the spectral decomposition of $\mathcal{A}$, namely

$$\mathcal{A}^\alpha u = \sum_{j=1}^{\infty} \lambda_j^\alpha (u, \psi_j)\psi_j,$$

where $\lambda_j > 0$ are the eigenvalues, respectively, ψ_j are the corresponding normalized eigenfunctions of $\mathcal{A}$, and $(.,.)$ is the L^2 inner product.

Let ω_h be a uniform rectangular mesh, while the Finite Difference Method (FDM) is applied to approximate the operator $\mathcal{A}$. In this case, the FDM discretization of (1) leads to a system of linear algebraic equations with respect to the mesh functions $\mathbf{u}$ and $\mathbf{f}$, in the form

$$A^\alpha \mathbf{u} = \mathbf{f}. \tag{2}$$

Here, $A \in \mathbb{R}^{N\times N}$ is a sparse, symmetric and positive definite (SPD) matrix, and

$$A^\alpha \mathbf{u} = \sum_{j=1}^{N} \lambda_{j,h}^\alpha (\mathbf{u}, \Psi_{j,h})\Psi_{j,h}. \tag{3}$$

As in the continuous case, $\{\lambda_{j,h}\}_{j=1}^N$ is the spectrum of A, and the eigenvectors $\Psi_{j,h}$ are normalized with respect to the Euclidean dot product $(.,.)$.

When $\alpha \in (0,1)$, the problem (1), corresponding to the fractional sub-diffusion, has been extensively studied in the past decade (see, for instance, [1]). When $\alpha \in (1,2)$, the problem (1) is related to the fractional super-diffusion [2], which is the focus of this study. Notably, many numerical methods developed for the sub-diffusion case rely directly on the condition $\alpha \in (0,1)$. Therefore they do not extend naturally to the super-diffusion case. For $\alpha \in (1,2)$, the BURA has one positive pole and one positive zero, making one of the resulting linear systems non-positive definite. To address this, we employ the BURA method, based on the Best Uniform Rational Approximation $r_{\alpha,k}(t)$ of degree k of t^α in the interval $[0,1]$. Specifically, we investigate the numerical behavior of the BURA method [3], [4] when applied to the fractional super-diffusion problems.

II. The BURA Method

The abbreviation BURA stands for Best Uniform Rational Approximation. Let us consider the min-max problem: find $r_{\alpha,k} \in \mathcal{R}(k,k)$ such that

$$\max_{t\in[0,1]} |t^\alpha - r_{\alpha,k}(t)| = \min_{r_k(t)\in\mathcal{R}(k,k)} \max_{t\in[0,1]} |t^\alpha - r_k(t)|, \quad \alpha \in (0,1), \tag{4}$$

where $r_k(t) = P_k(t)/Q_k(t)$, P_k and Q_k are polynomials of degree k. Then the error $E_{\alpha,k}$ of the k-BURA element $r_{\alpha,k}$ is defined as

$$E_{\alpha,k} := \max_{t\in[0,1]} |t^\alpha - r_{\alpha,k}(t)|. \tag{5}$$

A sharp estimate of $E_{\alpha,k}$ has been derived in [5] and has the form:

$$E_{\alpha,k} = 4^{\alpha+1} \sin(\alpha\pi) e^{-2\pi\sqrt{\alpha k}}. \tag{6}$$

Following [6] we introduce the approximation of $A^{-\alpha}$ in the form

$$A^{-\alpha} \approx \lambda_{1,h}^{-\alpha} r_{\alpha,k}\left(\lambda_{1,h} A^{-1}\right),$$

and then

$$\mathbf{u}_k = \lambda_{1,h}^{-\alpha} r_{\alpha,k}\left(\lambda_{1,h} A^{-1}\right)\mathbf{f}, \tag{7}$$

where $\mathbf{u}_k$ is the BURA numerical solution of the linear algebraic system (2). In other words, $\mathbf{u}_k$ is the BURA approximation of the solution $u(x)$ of the fractional elliptic equation (1). Scaling A^{-1} by $\lambda_{1,h}$, above, ensures that all eigenvalues of the matrix are in the interval $(0, 1]$ where the BURA error is uniform.

III. An application of the BURA method

In this work we consider the original problem (2) in the form:

$$A^{\alpha_1} A^{\alpha_2}\mathbf{u} = \mathbf{f},$$

where $\alpha = \alpha_1 + \alpha_2$, $\alpha_{1,2} \in (0, 1]$. Now, the BURA method can be used for solving the following sub-diffusion problems:

$$A^{\alpha_1}\mathbf{w} = \mathbf{f}, \tag{8}$$

$$A^{\alpha_2}\mathbf{u} = \mathbf{w}. \tag{9}$$

The application of the BURA method for solving such problems has been relatively well studied, during the last years.

In [2] the authors used similar approach where $\alpha = \sum_{i=1}^{l} \alpha_i$. In our study, we limit our analysis to the case $l = 2$, for the following reasons. First, using $l > 2$ requires solving larger number of linear systems, which increases the computational complexity, and reduces the efficiency of the solver. Second, for small values of α, a larger k must be used, to achieve good accuracy, which further increases the computational cost.

IV. Numerical experiments

The presented numerical experiments are designed to determine the optimal combination of parameters, balancing the numerical error and the computational time. Specifically, we consider the one dimensional problem (1) with the right hand side

$$f(x) = sin(\pi x) - sin(2\pi x) + sin(3\pi x),$$

where the computational domain Ω is the unit interval $[0, 1]$ and $\alpha = 1.25, 1.5, 1.75$. For this problem, it is possible to compute the exact solution of the problem (1). Therefore, the error in the numerical solution can be also precisely calculated.

The problem is discretized on a uniform mesh ω_h, $h^{-1} = n = 8000, 16000, 32000, 64000$, by applying the standard three point stencil. Next, the BURA method is applied with:

- $\alpha_1 = 1$;
- $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$;
- $\alpha_2 = 1$.

Specifically, the additive BURA method (see [4]) is used, where

$$r_{\alpha,k}(t) = c_0 + \sum_{i=1}^{k} \frac{c_i}{t - d_i}. \tag{10}$$

In this case, the equation (7) can be rewritten as

$$\mathbf{u}_k = \left[c_0 \lambda_{1,h}^{-1} I + \sum_{i=1}^{k} c_i \left(A - \lambda_{1,h} d_i I\right)^{-1}\right]\mathbf{f}, \tag{11}$$

The BRASIL [7], [8] software, version 1.2.1, was used for the computation of the coefficients of the rational approximation $r_{\alpha,k}(t)$ in the interval $[0, 1]$.

The l_2 norm of the error was computed as:

$$\|\mathbf{u} - \mathbf{u}_{exact}\|_{l_2} = \sqrt{\frac{\sum_{i=1}^{n+1}(u_i - u_{exact}(x_i))^2}{n+1}}.$$

The code has been implemented in Python. (See the Appendix for the main part of the code.) The equation (11) requires a solver for the tridiagonal systems of equations. In the implemented solver, the LAPACK [9] subroutine DGTSV was used for solving these tridiagonal systems. Across the solver, double precision real numbers were used.

The code has been tested on a server which has Intel(R) Xeon(R) Silver 4309Y CPU with 16 cores at 2.8 GHz and 32 GB of RAM. However, it should be noted that no attempt was made to parallelize the solution process. This direction of research may be pursued in the future, when substantially larger problems are going to be solved.

V. Numerical results

The l_2 norm of the error for $n = 8000, 32000$ is presented in Tables I and II, the l_2 norm of the error for $n = 64000$ is presented in Fig. 1. Let us note that the error of the numerical solution, in the experiments, consists of a composition of several types of errors. First, the discretization leads to an error. Here, when the Laplace equation is formed, it can be estimated as $O(h^2)$. Second, the rational approximation has an error $E_{\alpha,k}$ and the sharp estimate of $E_{\alpha,k}$ is presented in (6). Thus, the error of the rational approximation depends on αk. For $\alpha \in (0, 1)$, a more accurate rational approximation for α close to 1, and for large k, is obtained. Third, there is also an error in the numerical solution of the systems of linear algebraic equations, in (11). It should be noted that increasing the size of these systems leads to a larger error in the solution. Finally, at every step of the numerical solution of the problem there are rounding errors. Unfortunately, there is no easy way to decouple these errors, and capture them independently.

From Tables I and II it follows that for the same α and the same k a better accuracy is obtained when $\alpha_1 = \alpha_2$. Next, as expected, for α close to 1, the accuracy is higher. However, higher accuracy comes at an additional cost. When $\alpha_1 = \alpha_2$, the BURA method is applied twice to solve (8–9). When $\alpha_1 = 1$ (or $\alpha_2 = 1$) the BURA method is applied only for solving equations (9) (or (8)). Thus, in total, for $\alpha_1 = \alpha_2$, $2k$ linear systems are solved, while for $\alpha_1 = 1$ (or $\alpha_2 = 1$) only $k + 1$ linear systems have to be solved. The practical realization of this comparison can be seen in Table III, where the average execution time (from multiple runs) for $n = 8000$ is presented in milliseconds.

The next observation that can be derived from Tables I and II is that the difference in the accuracy for $\alpha_1 = 1$ and $\alpha_2 = 1$ is less than one percent and thus can be neglected.

From the results represented in Fig. 1 it is possible to compare the error obtained for $\alpha_1 = \alpha_2$, $k = 10$ and for $\alpha_1 = 1$, $k = 19$. In both cases, in total, 20 linear systems

TABLE I
l_2 NORM OF THE ERROR FOR $n = 8000$.

k	(α_1, α_2)								
	(1, 0.25)	(0.625, 0.625)	(0.25, 1)	(1, 0.5)	(0.75, 0.75)	(0.5, 1)	(1, 0.75)	(0.875, 0.875)	(0.75, 1)
12	2.64e-6	1.86e-08	2.64e-6	3.02e-08	1.67e-09	3.02e-08	3.34e-10	2.08e-10	3.34e-10
16	4.91e-7	1.33e-09	4.91e-7	2.47e-09	4.24e-10	2.47e-09	2.86e-10	3.07e-10	2.86e-10
20	1.14e-7	7.50e-10	1.14e-7	3.50e-10	5.01e-10	3.50e-10	3.09e-10	3.05e-10	3.08e-10
24	2.99e-8	8.82e-10	2.99e-8	4.63e-10	5.10e-10	4.63e-10	3.11e-10	3.11e-10	3.11e-10

TABLE II
l_2 NORM OF THE ERROR FOR $n = 32000$.

k	(α_1, α_2)								
	(1, 0.25)	(0.625, 0.625)	(0.25, 1)	(1, 0.5)	(0.75, 0.75)	(0.5, 1)	(1, 0.75)	(0.875, 0.875)	(0.75, 1)
12	2.64e-6	1.97e-08	2.64e-6	3.07e-08	1.62e-09	3.07e-08	4.30e-10	3.10e-10	4.30e-10
16	4.91e-7	1.59e-09	4.91e-7	2.91e-09	1.11e-10	2.91e-09	4.08e-11	1.67e-10	4.08e-11
20	1.14e-7	1.42e-10	1.14e-7	2.85e-10	2.32e-10	2.85e-10	7.51e-11	3.21e-11	7.51e-11
24	3.06e-8	1.87e-10	3.06e-8	4.90e-11	1.94e-11	4.92e-11	8.93e-12	1.91e-10	8.91e-12

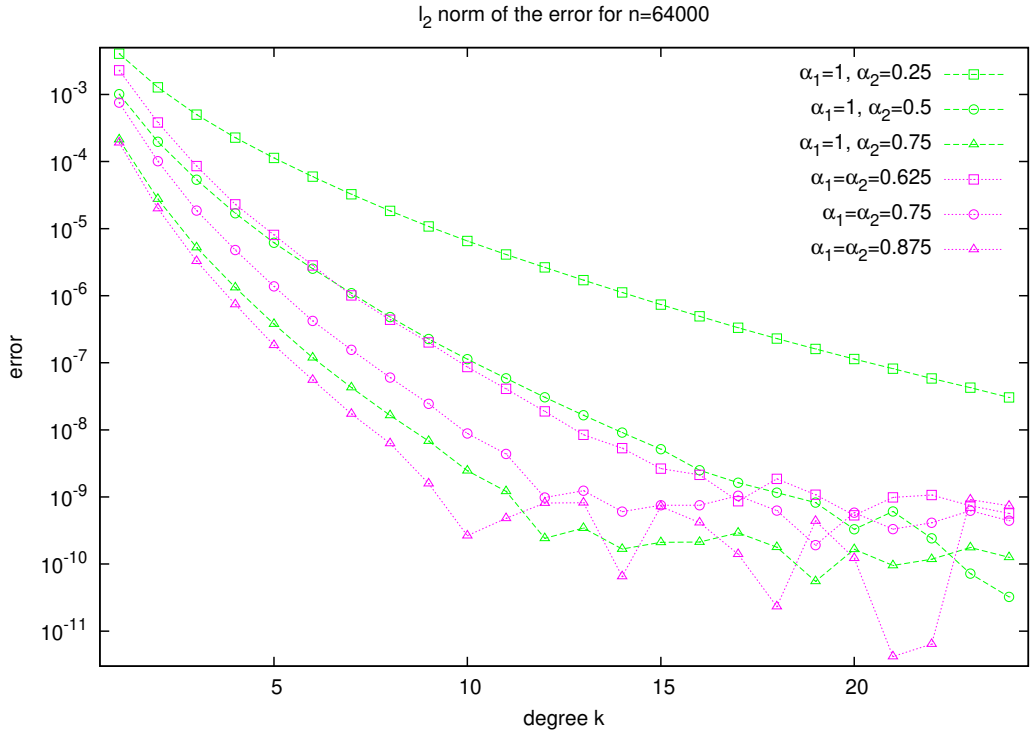


Fig. 1. l_2 norm of the error for $n = 64000$.

have been solved, and the execution time is almost the same. For $\alpha = 1.25$ the obtained errors are $\approx 9.10^{-8}$ and 2.10^{-7}, for $\alpha = 1.5$, the errors are $\approx 9.10^{-9}$ and 8.10^{-9}, for $\alpha = 1.75$ the errors are $\approx 3.10^{-10}$ and 6.10^{-11}. Therefore, it can be concluded that, depending on the value of α, it is possible to choose the optimal order k and the parameter α_1.

The next observation concerns the results depicted in Fig. 1. There, one can see the strange behavior of the error for $\alpha_1 = \alpha_2 = 0.875$ for large k. As noted above, the error consists (among others) of the discretization error and the rational approximation error. For $n = 64000$, the discretization error is $\approx 2.5 \times 10^{-10}$ and this error could not be improved for $k > 10$.

Finally, Fig. 2 shows the average execution time (from multiple runs) for $n = 8000, 16000, 32000, 64000$, in seconds. As noted above, for $\alpha_1 = 1$ the algorithm solves $k + 1$ linear systems of size $n + 1$, while for $\alpha_1 = \alpha_2$ — $2k$ systems of the same size. As expected, the execution time is proportional to $k + 1$ or to k and the linearity is clearly visible in the plot.

VI. CONCLUDING REMARKS

In this contribution, the numerical solution of the spectral fractional elliptic equation with power $\alpha \in (1, 2)$ was studied.

TABLE III
THE EXECUTION TIME FOR $n = 8000$ IN MILLISECONDS.

k	(α_1, α_2)								
	(1, 0.25)	(0.625, 0.625)	(0.25, 1)	(1, 0.5)	(0.75, 0.75)	(0.5, 1)	(1, 0.75)	(0.875, 0.875)	(0.75, 1)
12	2.82	5.09	2.79	2.82	5.09	2.79	2.83	5.09	2.79
16	3.63	6.72	3.62	3.64	6.72	3.62	3.65	6.71	3.61
20	4.44	8.35	4.42	4.44	8.37	4.43	4.46	8.35	4.36
24	5.27	9.96	5.23	5.26	9.98	5.23	5.33	9.96	5.24

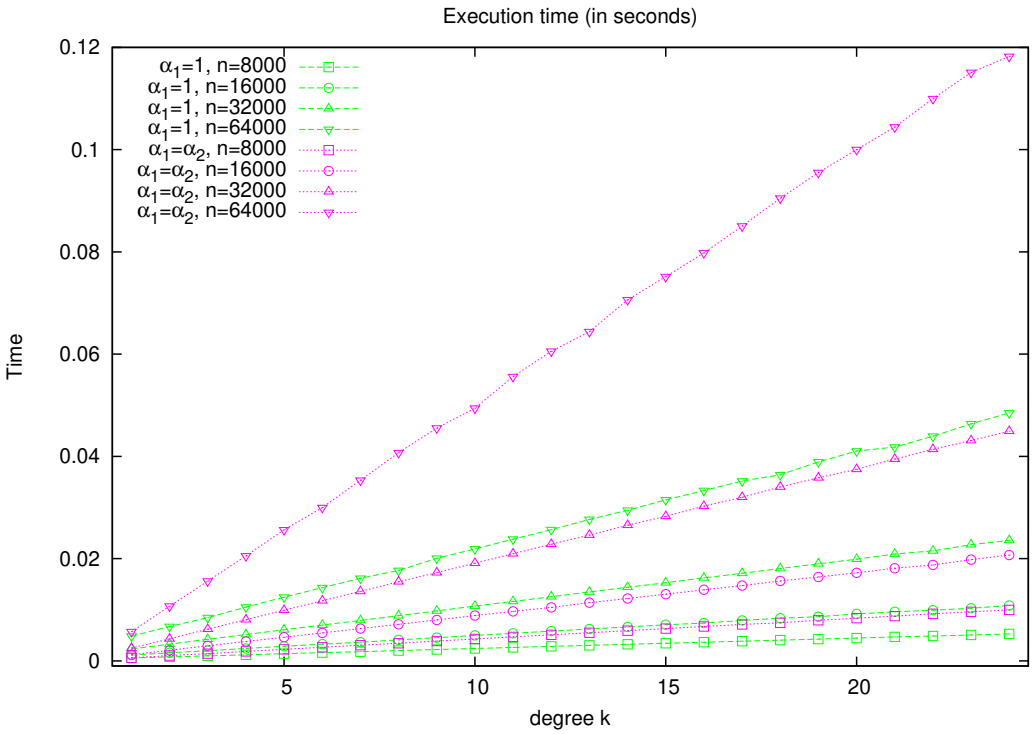


Fig. 2. Execution time in seconds for $n = 8000, 16000, 32000, 64000$.

Specifically, for the problem in question, the Best Uniform Rational Approximation method has been applied. Next a comprehensive set of experiments has been completed. The obtained results match the theoretical expectations. In particular, the fact that the solution error is composed of multiple components has been captured for large values of k and n. This may constitute the actual limit of applicability of the proposed approach (in its standard form).

ACKNOWLEDGMENT

Work presented here is part of the collaborative grant between the Polish Academy of Sciences and the Bulgarian Academy of Sciences, IC-PL/01/2022-2023, "Practical aspects of scientific computing". The work was partially supported by the Bulgarian National Science Fund under grant No. KP-06-N72/2.

REFERENCES

[1] C. Hofreither, "A unified view of some numerical methods for fractional diffusion," *Comp Math Appl.*, vol. 80, no. 2, pp. 332–350, 2020. doi: 10.1016/j.camwa.2019.07.025

[2] D. Slavchev, S. Harizanov, and N. Kosturski, "Analysis on computational issues when approximating fractional powers of sparse SPD matrices," *International Journal of Applied Mathematics*, vol. 37, no. 6, pp. 699–711, 2024. doi: 10.12732/ijam.v37i6.7

[3] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and Y. Vutov, "Optimal solvers for linear systems with fractional powers of sparse SPD matrices," *Numer Linear Algebra Appl.*, vol. 25, no. 5, p. e2167, 2018. doi: 10.1002/nla.2167

[4] S. Harizanov, N. Kosturski, I. Lirkov, S. Margenov, and Y. Vutov, "Reduced multiplicative (BURA-MR) and additive (BURA-AR) best uniform rational approximation methods and algorithms for fractional elliptic equations," *Fractal and Fractional*, vol. 5, no. 3, p. 61, 2021. doi: 10.3390/fractalfract5030061

[5] H. Stahl, "Best uniform rational approximation of x^{α} on [0, 1]," *Acta Math*, vol. 190, no. 2, pp. 241–306, 2003.

[6] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and J. Pasciak, "Analysis of numerical methods for spectral fractional elliptic equations based on the best uniform rational approximation," *Journal of Computational Physics*, vol. 408, p. 109285, 2020. doi: 10.1016/j.jcp.2020.109285

[7] "Software BRASIL," accessed on September 1, 2022. [Online]. Available: https://baryrat.readthedocs.io/en/latest/#baryrat.brasil

[8] C. Hofreither, "An algorithm for best rational approximation based on barycentric rational interpolation," *Numerical Algorithms*, vol. 88, pp. 365–388, 2021. doi: 10.1007/s11075-020-01042-0

[9] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed. Philadelphia: SIAM, 1999.

APPENDIX

```
from baryrat import brasil
from mpmath import mp
from numpy import linalg
from scipy.linalg import lapack
import argparse
import numpy as np
# ...
lambda1 = (2*(args.n)*np.sin(np.pi/2/(args.n)))**2
# solver
def thomas(dl, d, du, b):
    _, _, _, x, _ = lapack.dgtsv(dl, d, du, b)
    return x

# BURA
class bura:
    def __init__(self, alpha, k):
        if alpha == 0 or alpha == 1:
            self._alpha = alpha
            return
        rat = brasil(
            lambda x: x**alpha,
            (0, 1),
            k,
            tol=np.finfo(np.double).eps,
            npi=-30,
            maxiter=10000,
        )

        poles = np.real_if_close(np.array(rat.poles(True), dtype=np.complex128))
        zeros = np.real_if_close(np.array(rat.zeros(True), dtype=np.complex128))

        d = 1 / np.flip(np.sort(poles))
        z = 1 / zeros
        b = np.zeros(k + 1)
        b[-1] = rat(1)
        m = 1 / np.subtract.outer(np.append(z, 1), d)
        m = np.hstack((m, np.ones((m.shape[0], 1))))
        c = np.real(np.array(mp.lu_solve(np.real(m), b), dtype=np.complex128))
        self._alpha, self._k, self._c, self._d = alpha, k, c, d

    def solve(self, dl, d, du, b, lambda1=lambda1):
        if self._alpha == 0:
            return b
        if self._alpha == 1:
            return thomas(dl, d, du, b)
        n = len(b)
        r = self._c[-1] * b
        for i in range(self._k):
            r += (
                lambda1
                * self._c[i]
                * thomas(dl, d - lambda1 * self._d[i] * np.ones(n), du, b)
            )
        return lambda1**-self._alpha * r
```

```
bura1, bura2 = bura(args.alpha1, args.k1), bura(args.alpha2, args.k2)
num = bura1.solve(lower_diag, main_diag, upper_diag, b)
num = bura2.solve(lower_diag, main_diag, upper_diag, num)

exact = exact_solution(dofs, alpha)
err = exact - num
max_err = np.max(np.abs(err))
print("Max error:", max_err)
print("L2 error:", linalg.norm(err)/((args.n+1)**0.5))
```

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 721–726
DOI: 10.15439/2025F5317
ISSN 2300-5963 ACSIS, Vol. 43

An algorithm for Direct Construction of all Pareto Optimal Biobjective Minimum Spanning Trees

Lasko M. Laskov
0000-0003-1833-818
Informatics Department
New Bulgarian University
21 Montevideo Str., 1618 Sofia, Bulgaria
Email: llaskov@nbu.bg

Marin L. Marinov
0009-0003-9544-819X
Informatics Department
New Bulgarian University
21 Montevideo Str., 1618 Sofia, Bulgaria
Email: mlmarinov@nbu.bg

***Abstract*—In this paper we describe an exact algorithm that constructs all Pareto optimal solutions of the biobjective minimum risk minimum length spanning trees problem. The method constructs directly the Pareto optimal spanning trees without constructing the entire classes of optimal spanning trees with respect to the length and with respect to the risk criterion.**

We formulate the theorems that prove the correctness and computational complexity of the proposed algorithms. Also, we illustrate the method using a comprehensive numerical example. The computational complexity of the algorithm that constructs the complete Pareto front of the problem is $O(s(m + n \lg n))$, where s is a constant that depends on the number of all Pareto optimal solutions and a predefined constant that can be used to limit the number of constructed solutions.

I. Introduction

THE minimum spanning tree (MST) problem is the problem to construct a single spanning tree in an undirected weighted graph that has minimum total weight. A spanning tree is an acyclic subset of the edges of the graph that connects all its vertices. In the standard single-objective case the weight of the spanning tree is defined as the sum of the weights of all edges that are contained in it (see for example [1]).

While in practice even the basic binary heap implementation gives a good performance of the Prim's algorithm [2], the advanced data structure Fibonacci heap [3] can result in even more significant speedup. In the general case, the Prim's algorithm implemented with Fibonacci heap runs in $O(m+n \lg n)$ computational complexity, where n is the number of vertices and m is the number of edges in the graph. In the case of sparse graphs for which $m = \Omega(n \lg n)$, the computational complexity is improved to $O(m)$.

On the other hand, by introducing of additional objective function to the MST problem, in specialized literature is reported that it results into a NP-hard problem [4], [5], [6], [7]. Also, in the case of multiple objective functions, we are no longer looking for a single optimal solution, but a whole set of non-dominated optimal spanning trees that are called Pareto optimal (see for example [8]).

Even though the NP-hardness of the biobjective MSTs problem suggests that exact methods are inefficient to apply, and heuristic or metaheuristic methods, like for example ant colony optimization [9], are more appropriate, our research shows that for particular objective functions the problem turns to be weak NP-hard. The latter means that an exact pseudo-polynomial or even polynomial algorithms can be formulated.

In this paper we describe an exact algorithm that constructs the complete Pareto front of the biobjective minimum risk minimum length spanning trees problem. The algorithm does not construct the entire classes of optimal spanning trees with respect to the length and with respect to the risk criterion and directly finds the Pareto optimal solutions of the problem with computational complexity $O(s(m + n \lg n))$, $s = \min\{M, |P|\}$, where M is a predefined constant that can be used to limit the number of constructed solutions, and P is the complete Pareto optimal set.

The paper is organized as follows. In the next section we provide definitions and problem formulation. Section III-A describes our algorithm that constructs the complete set of MSTs for the standard single-objective version of the problem. Section III-B describes our polynomial complexity algorithm that finds the minimum complete Pareto front. The algorithm that solves the main problem in this paper is given in section IV, and section V contains our conclusions.

II. Problem definition

We denote with $G = (V, E)$ a connected undirected graph with $n = |V|$ number of vertices and $m = |E|$ number of edges. The vertices of G are labeled with the first n natural numbers and $V = \{1, 2, \ldots, n\}$. The edges of the G are symmetric and we will denote each edge $(u, v) \equiv (v, u)$ as a pair of its endpoint vertices $u, v \in V$. We define two weight functions on the set of the graph edges: $f : E \to \mathbb{R}_+$ and $g : E \to \mathbb{R}_+$. The function f maps to each edge $(u, v) \in E$ the positive number $f(u, v)$ which we will call the *length* of (u, v). The function g maps to each edge $(u, v) \in E$ the positive number $g(u, v)$ which we will call the *risk* of (u, v).

The graph G together with the two weight function defines the network $N = (V, E, f, g)$ which is represented by the adjacency lists $N.Adj$ which is a vector of n elements, where each element $N.Adj[u]$ is a list of the adjacent vertices to the vertex u and corresponding edges weights:

$$\langle (v_{u_1}, f(u, v_{u_1}), g(u, v_{u_1})), (v_{u_2}, f(u, v_{u_2}), g(u, v_{u_2})), \ldots \rangle$$

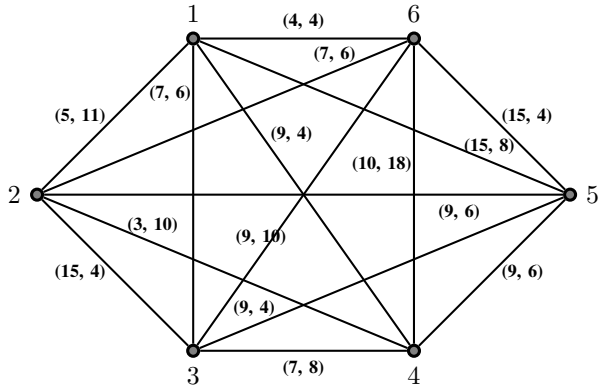


Fig. 1. Example complete network N_1

TABLE I
ADJACENCY LISTS OF EXAMPLE COMPLETE NETWORK N_1 GIVEN FOR EACH VERTEX $u \in V$

u	Adjacency list
1	$\langle(2,5,11),(3,7,6),(4,9,4),(5,15,8),(6,4,4)\rangle$
2	$\langle(1,5,11),(3,15,4),(4,3,10),(5,9,6),(6,7,6)\rangle$
3	$\langle(1,7,6),(2,15,4),(4,7,8),(5,9,4),(6,9,10)\rangle$
4	$\langle(1,9,4),(2,3,10),(3,7,8),(5,9,6),(6,10,18)\rangle$
5	$\langle(1,15,8),(2,9,6),(3,9,4),(4,9,6),(6,15,4)\rangle$
6	$\langle(1,4,4),(2,7,6),(3,9,10),(4,10,18),(5,15,4)\rangle$

This representation is illustrated with the adjacency lists given in Table I of the example complete network N_1 (see Fig. 1).

For each subset of edges $A \subseteq E$ we define the *length* of A:

$$x(A) = \sum_{i=1}^{|A|} f(u_i, v_i), \tag{1}$$

and the *risk* of A with:

$$y(A) = \max\{g(u,v) : (u,v) \in A\}. \tag{2}$$

We denote with W the set of all spanning trees in the network N. The set of all spanning trees with minimum length is W_f, where:

$$W_f = \{T' \in W : x(T') \leq x(T), \forall T \in W\}, \tag{3}$$

and the set of all spanning trees with minimum risk is W_g, where:

$$W_g = \{T' \in W : y(T') \leq y(T), \forall T \in W\}. \tag{4}$$

Definition 1. *We call the tree $T' \in W$ **Pareto optimal spanning tree** (POST) when there does not exist another tree $T \in W$, for which any of the following two conditions is fulfilled:*

- $x(T) < x(T')$ *and* $y(T) \leq y(T')$*;*
- $x(T) \leq x(T')$ *and* $y(T) < y(T')$*.*

We denote with P the set of all POSTs in the network N. P is the Pareto optimal set and is represented in the form:

$$P = \bigcup_{i=1}^{K} P_i, \tag{5}$$

where the subsets P_i satisfy the following two conditions.

1) All trees in P_i, $i = 1, 2, \ldots, K$, are equivalent.
2) For each $i = 1, 2, \ldots, K-1$ the following inequalities hold:
$$l_i < l_{i+1} \text{ and } r_i > r_{i+1}, \tag{6}$$
where l_i is the length, r_i is the risk of each tree in P_i.

The set P has an unique representation (5). We will call each subset P_i *class of equivalent POSTs*, and we will call the set P *complete Pareto front* (CPF).

Definition 2. *We will call each set $P_{min} = \{t_1, t_2, \ldots, t_K\}$, where $t_i \in P_i$, $i = 1, 2, \ldots, K$ a **minimum complete Pareto front** (MCPF).*

Let $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ are arbitrary edges in the network N. We will use the following edge relations.

- The edges e_1 and e_2 are *equivalent*, denoted $e_1 \sim e_2$, when $f(e_1) = f(e_2)$ and $g(e_1) = g(e_2)$.
- The edge e_2 is *dominated by* e_1 with respect the *length*, denoted $e_2 \prec_f e_1$ when: (i) $f(e_1) < f(e_2)$ or (ii) $f(e_1) = f(e_2)$ and $g(e_1) < g(e_2)$.
- The edge e_2 is *dominated by* e_1 with respect the *risk*, denoted $e_2 \prec_g e_1$ when: (i) $g(e_1) < g(e_2)$ or (ii) $g(e_1) = g(e_2)$ and $f(e_1) < f(e_2)$.

When $e_2 \sim e_1$ or $e_2 \prec_f e_1$ is fulfilled, we will denote it by $e_2 \precsim_f e_1$, and when $e_2 \sim e_1$ or $e_2 \prec_g e_1$ is fulfilled, we will denote it by $e_2 \precsim_g e_1$.

The aim of this paper is the solution of the problem given in Problem 1.

Problem 1. *(Main Problem) Let N be a connected network and let M be a natural number. Find a set of POSTs such that:*

1) *if $M \leq K$, then $|P'| = M$ and $P'[i] \in P_i$, for $i = 1, 2, \ldots, M-1$, and $P'[M] \in P_K$;*
2) *if $K < M \leq |P|$, then $|P'| = M$ and P' contains at least one MCPF;*
3) *if $|P| < M$, then $P' = P$.*

The solution of Problem 1 that we propose in this paper is given in Sec. IV. The proposed algorithm finds the list P' and determines which of the three cases of the problem is fulfilled. It also determines whether P' contains at least one MCPF. If P' contains a MCPF, the algorithm reports whether the CPF is found, and if it is not found, which classes of equivalent POSTs, given in equation (5), may contain elements that are not included in P'.

III. HELPER ALGORITHMS

A. Complete list of all minimum spanning trees

In this section we will solve the helper Problem 2.

Problem 2. *Let N be a connected network and let M be a natural number. Construct the list W'_f of MSTs in N that has the following properties:*

1) *if $M \geq |W_f|$, then $W'_f = W_f$;*
2) *if $M < |W_f|$, then $|W'_f| = M$.*

Algorithm 1 Function MSTLIST(N, r, M)

Input: network N, root vertex r, natural number M
Output: list of MSTs L, Boolean flag all

```
1: let S be an empty stack, L be an empty list
2: push into S the initial subproblem
3: while S ≠ ∅ and |L| ≠ M do
4:    X ← POP(S), (t, S) ← SUBPROBLEMS(S, N, X)
5:    if t ∉ L then
6:       PUSHBACK(L, t)
7:    end if
8: end while
9: if S ≠ ∅ then
10:   all ← false
11: else
12:   all ← true
13: end if
14: return L, all
```

The solution of Problem 2 is given by the function MSTLIST in Algorithm 1. The input of the algorithm are the network N, a root vertex of the constructed trees r and a natural number M. The output are the list of lists of parents of the constructed trees L that contains the set W'_f, and a Boolean value all. The set W'_f forms the solution of Problem 2. When $all = true$, it shows that $W'_f = W_f$, and when $all = false$, then $|W'_f| = M < |W_f|$.

The helper function SUBPROBLEMS is described in details in our work [10]. The computational complexity of MSTLIST is evaluated to $O(s(m + n \lg n))$, where $s = \min\{|W_f|, M\}$.

B. Minimum complete Pareto front

In this section we describe Algorithm 3 that solves the helper Problem 3.

Problem 3. *Let M be an arbitrary natural number. Find a subset P' of the set of POSTs P that has the following properties:*

1) $|P'| = \min\{M, K\}$, *where K is the number of classes of equivalent POSTs;*
2) $|P' \cap P_i| = 1$, *for $i = 1, 2, \ldots, |P'|$.*

Algorithm 3 uses the following two modifications of the Prim's algorithm.

1) Function MINTREERISK that finds a POST that has minimum risk. The pseudocode of the function is given in Algorithm 2.
2) Function MINTREELENGTH that finds a POST that has minimum length. In this case the algorithm is analogous to Algorithm 2, but the relation $\prec_f$ is applied.

Both functions MINTREERISK and MINTREELENGTH extend the greedy principle of the Prim's algorithm using the relations $\prec_g$ and $\prec_f$ respectively.

Theorem 1. *Let A is a tree that corresponds to the predecessor list t_{bar} calculated with* MINTREERISK. *Then A is a POST that has minimum risk among all POSTs.*

Algorithm 2 Function MINTREERISK(N, r)

Input: network N, root vertex r
Output: POST that has minimum risk t_{bar}

```
1: a ← [false, false, ..., false]
2: d ← [∞, ∞, ..., ∞], d[r] ← (0, 0), t ← [0, 0, ..., 0]
3: insert each u ∈ V into Q with key d[u]
4: while Q ≠ ∅ do
5:    u ← EXTRACTMIN(Q, ≺_g)
6:    a[u] ← true
7:    for all (v, f(u,v), g(u,v)) ∈ N.Adj[u] do
8:       if a[v] = false and d[v] ≺_g (u, v) then
9:          t[v] ← u, d[v] ← (f(u,v), g(u,v))
10:         DECREASEKEY(Q, v, (f(u,v), g(u,v)))
11:      end if
12:   end for
13: end while
14: E' ← ∅
15: for all e ∈ E do
16:   if g(e) ≤ y(t) then
17:      E' ← E' ∪ e
18:   end if
19: end for
20: E ← E', t_bar ← MSTList(N, r, 1)
21: return t_bar
```

Algorithm 3 Function MINCPF(N, r, M)

Input: network N, root vertex r, natural number M
Output: POSTs list L, barrier tree t_{bar}, integers c_{dif} and M'

```
1: let L be an empty list
2: M' ← M, t_bar ← MINTREERISK(N, r)
3: c_bar ← y(t_bar), c_dif ← 1
4: while M' > 0 and c_dif > 0 do
5:    t ← MINTREELENGTH(N, r)
6:    PUSHBACK(L, t)
7:    c ← y(t)
8:    E' ← ∅
9:    for all e ∈ E do
10:      if g(e) < c then
11:         E' ← E' ∪ e
12:      end if
13:   end for
14:   E ← E'
15:   M' ← M' − 1, c_dif ← c − c_bar
16: end while
17: return L, t_bar, c_dif, M'
```

If A is a tree with a predecessor list calculated with MINTREELENGTH, *then A is a POST that has a minimum length among all POSTs.*

Theorem 2. *The computational complexity of both functions* MINTREERISK *and* MINTREELENGTH *is* $O(m + n \lg n)$.

Algorithm 3 applies as a step also the restriction operation (lines 8-13) which separates a subnetwork N' of the input

network N. The resulting subnetwork N' has the same set of vertices, however it contains only these edges of the original network, which have a risk *strictly less* than the fixed risk value c.

Let the list L is a result of the calculations of Algorithm 3. We define the set of trees P' such that:

1) the number of elements of P' is equal to the length of the list, $|P'| = |L|$;
2) $P'[i]$ is the tree that has the predecessor list $L[i]$, for $i = 1, 2, \ldots, |L|$.

Then the following Theorem 3 holds.

Theorem 3. *The set P' is a solution of Problem 3.*

The proof of Theorem 3 directly follows from Theorem 1.

Corollary 1. *For the above defined set P', barrier tree t_{bar}, integers c_{dif} and M' that result from Algorithm 3, the following four statements hold.*

1) *The tree T_{bar} with predecessor list t_{bar} is a POST and $T_{bar} \in P_K$.*
2) *$P'[i] \in P_i$, for $i = 1, 2, \ldots, |L|$.*
3) *If $c_{dif} = 0$, then the set P' is a MCPF. In this case, if $T_{bar} \neq P'[K]$, then besides the MCPF P', Algorithm 3 have constricted a second tree in the class P_K.*
4) *If $c_{dif} > 0$ and $M' = 0$, then $M < K$. Then for each $j \in \{M+1, \ldots, K\}$ the inequalities $x(P'[M]) < l_j \leq x(T_{bar})$ and $y(T_{bar}) \leq r_j < y(P'[M])$ hold. In this case $M+1$ number of POSTs are constructed from the same MCPF. An additional verification is needed in order to determine whether the set $P' \cup \{T_{bar}\}$ is a MCPF.*

From Theorem 2 it follows that the running time of Algorithm 3 can be evaluated $O(\alpha(m + n \lg n))$, where $\alpha = \min\{M, K\}$. Because K cannot exceed the number of edges of the network m, we say that Algorithm 3 has polynomial complexity.

IV. Complete Pareto front

The helper algorithms, given in section III, are used in the definition of our algorithm for direct construction of all Pareto optimal biobjective minimum spanning trees, which is described in this section.

Let the CPF P of the network N is given by the equality (5). The solution of the main Problem 1 is given by the function DIRECTCPF, described in Algorithm 4. The algorithm stores the predecessor lists of the discovered POSTs in the list of lists C. Each list of lists in C corresponds to one class of equivalent POSTs. The algorithm itself implements the following **inductive procedure**:

1) **Base case.** Store in the list C the first MCPF. Complement C with the remaining POSTs from the class P_1.
2) **Inductive step.** Suppose that in C are stored all POSTs from the class P_j. The procedure complements C with the remaining POSTs from the class P_{j+1}.
3) The inductive procedure continues until all POSTs are constructed or until in C are stored M number of POSTs.

Algorithm 4 Function DIRECTCPF(N, r, M)

Input: network N, root vertex r, natural number M
Output: list of lists of POSTs C and seven indicators

let C be an empty list of lists of POSTs
$M' \leftarrow M$
$(L, c_{dif}, t_{bar}, M') \leftarrow$ MINCPF(N, r, M')
$k \leftarrow |L|$, $t_{last} \leftarrow L[k]$, $j \leftarrow 0$, $all \leftarrow false$
if $c_{dif} = 0$ **then**
  while $j < k$ **and** $M' > 0$ **do**
    $M' \leftarrow M' + 1$, $E' \leftarrow \varnothing$
    for all $e \in E$ **do**
      if $g(e) \leq y(L[j+1])$ **then**
        $E' \leftarrow E' \cup e$
      end if
    end for
    $E \leftarrow E'$
    $(L', all) \leftarrow$ MSTLIST(N, r, M')
    $M' \leftarrow M' - |L'|$
    PUSHBACK(C, L')
    $j \leftarrow j + 1$
  end while
  while $j < k$ **do**
    PUSHBACK$(C, \langle L[j+1] \rangle)$
    $j \leftarrow j + 1$
  end while
else
  for all $t \in L$ **do**
    PUSHBACK$(C, \langle t \rangle)$
  end for
end if
return C, c_{dif}, t_{last}, t_{bar}, M', j, k, all

The function DIRECTCPF takes as parameters the input network N, a root vertex r and a natural number M that defines the maximum number of POSTs. As a result of its calculations, the function returns the list of lists of POSTs C and the following indicators:

- c_{dif} shows whether a MCPF is constructed by the function MINCPF;
- t_{last} is the last POST contained in the list of the MCPF L;
- t_{bar} is the barrier tree that belongs to the class P_K constructed by MINCPF;
- M' shows if the number of discovered POSTs is M and if not, how many POSTs are required;
- j shows which is the last class of equivalent POSTs that is stored in C
- k shows whether in C is stored at least one MCPF;
- all is a Boolean flag that shows whether the last call of MSTLIST found all MSTs in the restricted network.

Theorem 4, given below, follows from Theorem 1, Theorem

3 and the correctness of algorithms in section III.

Theorem 4. *Algorithm 4 solves correctly Problem 1.*

Besides that, the following six properties hold.

Property 1. *Each element of C is a list of the predecessors lists of the same class of equivalent POSTs. For example, $C[1]$ is is a list of the predecessors lists of the class P_1.*

Property 2. *If $M' = 0$, then in C are stored exactly M number of POSTs. If $M' > 0$, then exactly M' number of POSTs are still required.*

Property 3. *The indicator k can take exactly two values. It is equal to 0, when C does not store any MCPF. It is equal to the number of classes of equivalent POSTs K, when in C is stored at least one MCPF.*

Property 4. *The indicator $j \in \{0, 1, \ldots, K\}$. If $j = 0$ then C stores at most one element of each class of equivalent POSTs. If $j \in \{1, 2, \ldots, K\}$, then the calculations of the algorithm terminated at the stage in which in C were included the predecessors lists of the POSTs of the class P_j. When $j \in \{2, 3, \ldots, K\}$, then $C[i]$ stores the complete classes P_i, for $i = 1, 2, \ldots, j-1$.*

Property 5. *If $M' = 0$ and $k = 0$, then M number of POSTs has been discovered before a MCPF was completed. The trees t_{bar} and t_{last} define the rectangular area in which are located the POSTs of the MCPF that are not stored in C.*

Property 6. *When $j \neq 0$ and $all = false$, then P_j may contain elements that are not stored in C. If $j \neq 0$ and $all = true$, then all elements of P_j are stored in C.*

From Theorem 2 it follows that the computational complexity of Algorithm 4 is $O(s(m + n \lg n))$, where $s = \min\{M, |P|\}$.

The calculations of Algorithm 4 are illustrated in Example 1. Because of the small size of the network used in the example, the above six properties can be directly verified.

Example 1. *Consider the example network N_1 given in Fig. 1 with adjacency lists in Table I. We select the vertex $r = 1$ for the root vertex of the POSTs that will be constructed. We will track the calculations of Algorithm 4 in the following four cases: (a) $M = 2$; (b) $M = 5$; (c) $M = 18$ and (d) $M = 25$.*

Solution. Before we solve the example, in order to illustrate graphically the results of the calculations, first we find all the spanning trees in the network N_1 using function MSTLIST given in Algorithm 1. The number of all spanning trees of N_1 which has number of vertices $n = 6$ is $n^{n-2} = 1256$. For each spanning tree t we store the pairs $(x(t), y(t))$, and we store only those pairs that does not coincide. For N_1 there are 116 such pairs, which means that the set of all spanning trees in N_1 is divided in 116 classes of equivalent trees. In Fig. 2 and 3 we illustrate all classes of equivalent trees with a black dot with the corresponding Cartesian coordinates.

(a) We solve Problem 1 with Algorithm 4 for $M = 2$. The

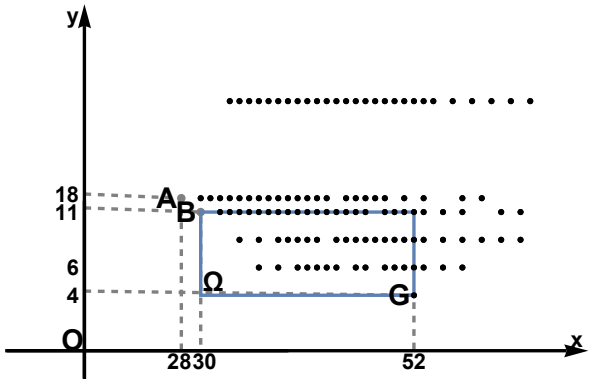


Fig. 2. Partial MCPF and the rectangle area Ω for the example network N_1. The POST that represent the class P_1 is denoted with A, and the POST that represent the class P_2 is denoted with B. The barrier tree is denoted with G

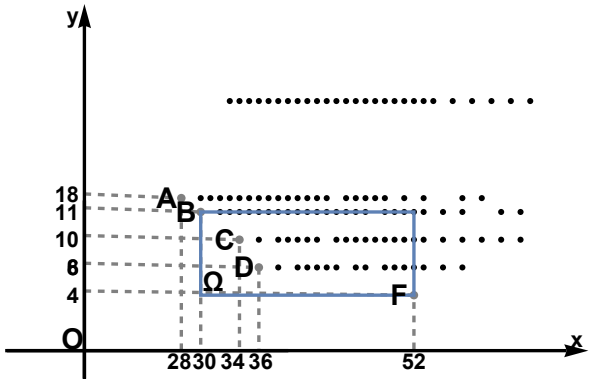


Fig. 3. MCPF for the example network N_1. The POSTs that represent the classes P_1 to P_5 are denoted with A, B, C, D and F respectively

resulting list is:

$$C = \langle\langle[0, 1, 1, 2, 3, 1]\rangle, \langle[0, 6, 1, 2, 3, 1]\rangle\rangle.$$

From the correctness of Algorithm 4 it follows that the two trees stored in C are POSTs and both of them are from a different class of equivalent POSTs. Since $k = 0$, from Property 5 it follows that the entire MCPF is not yet constructed. If we denote with T_1 the POST with predecessor list $[0, 1, 1, 2, 3, 1]$ and with T_2 the POST with with predecessor list $[0, 6, 1, 2, 3, 1]$, then $T_1 \in P_1$ and $T_2 \in P_2$. In this case $l_{last} = 30$ and $c_{last} = 10$ are respectively the length and the risk of the last tree T_2.

In Fig. 2 we denote the trees T_1 with the point A and T_2 with the point B. Also, the algorithm calculates the barrier tree T_{bar} with predecessor list t_{bar}. We denote with G the point with coordinates $(x(T_{bar}), y(T_{bar})) = (52, 4)$. From the correctness of Algorithm 3 it follows that POSTs that does not belong to P_1 or P_2 are contained into the rectangular area Ω with diagonal BG and sides parallel to the coordinate axes. From Corollary 1 we know that the barrier tree T_{bar} (illustrated by the point G in Fig. 2) is a POST from the last class of equivalent POSTs.

(b) We solve Problem 1 with Algorithm 4 for $M = 5$. After

the termination of the algorithm $M' = 0$ and therefore exactly $M = 5$ POSTs are found. In this case the resulting list is:

$$C = \langle\langle[0,1,1,2,3,1]\rangle, \langle[0,6,1,2,3,1]\rangle, \langle[0,1,1,2,3,1]\rangle, \langle[0,6,1,2,3,1]\rangle, \langle[0,1,1,2,3,1]\rangle\rangle.$$

Since now $k = 5 \neq 0$, from Property 3 it follows that C contains a MCPF and it is composed from exactly 5 POSTs. Besides that, $j = 0$ and from Property 4 it follows that for each class of equivalent POSTs a single element is included in the list C. Besides the trees T_1 and T_2 that were also calculated in the case (a), the list contains the trees T_3, T_4 and T_5 with predecessors lists respectively $[0,6,1,3,3,1]$, $[0,6,1,1,3,1]$ and $[0,3,5,1,6,1]$. The corresponding weights of these POSTs are $(34,8)$, $(36,6)$ and $(52,4)$. In Fig. 3 the POSTs T_3, T_4 and T_5 are denoted with the corresponding points in the Cartesian plane $C(34,8)$, $D(36,6)$ $F(52,4)$.

Also, in this case $all = false$ that gives us information that it is not clear whether in P_1 there is no more than one element. Besides that, it is directly verified that $T_5 = T_{bar}$ and in the class P_5 there is no second element that is discovered.

(c) We solve Problem 1 with Algorithm 4 for $M = 18$. After the termination of the algorithm, we get that $M' = 0$, and from Property 2 it follows that in the list C there are exactly 18 POSTs. It can be directly verified that in $C[1]$ there are 6 elements, in $C[2]$ there are 9 elements, and $C[3]$, $C[4]$ and $C[5]$ contain single element each.

Since $k = 5$, from Property 3 it follows that C contains at least one MCPF. Besides that, $j = 2$ and from Property 4 we get that the classes P_1 and P_2 are visited more than once by the algorithm and all elements of P_1 are stored in $C[1]$. Since $all = true$, from Property 6 it follows that all elements of P_2 are stored in $C[2]$. The POSTs stored in $C[3]$, $C[4]$ and $C[5]$ are the same as in the case (b).

(d) We solve Problem 1 with Algorithm 4 for $M = 25$. The algorithm returns $M' > 0$ and $k = 5 \neq 0$, and therefore the CPF is stored in C. The last class of equivalent POST that has been visited by the algorithm is defined by $j = 5$ and it is the class P_5. The classes P_1, P_2, P_3 and P_4 are stored respectively in $C[1]$, $C[2]$, $C[3]$ and $C[4]$. Since $all = true$, the entire class P_5 is stored in $C[5]$.

It can be directly verified that $|C[1]| = 6$, $|C[2]| = 9$, $|C[3]| = 3$, $|C[4]| = 5$, and $|C[5]| = 1$, and the CPF has totally 24 POSTs.

V. Conclusion

In the general case, the biobjective MSTs problem is widely considered an NP-hard problem, which means that an exact method will lead to exponential computational complexity. In our research we have shown that for particular objective functions, the problem is weakly NP-hard, and efficient exact algorithms can be found.

In this paper we have examined in details a version of biobjective MSTs problem in which the first objective function is linear (length), and the second objective function is non-linear bottleneck (risk). Our algorithm that finds the minimum complete Pareto front of the problem has polynomial complexity. The proposed algorithm in this paper that constructs the complete Pareto front has pseudo-polynomial complexity.

We have implemented and tested the described algorithms in this paper on complete random networks. The conducted experiments illustrate the effectiveness of Algorithm 4 that constructs the complete Pareto front. For example, for a complete random network with $n = 100$ number of vertices, the running time of the program written in Wolfram language is not more than 11 minutes on a standard computer configuration. In the particular case, the number of all spanning trees in the network are 100^{98}, the number of Pareto optimal spanning trees in the complete Pareto front are 4739, and it is distributed in 3986 number of classes of equivalent Pareto optimal spanning trees.

The results of the prototype program written in Wolfram language show that with a more optimized implementation (for example in C++ or Julia programming languages), we will be able to run the described algorithms with even bigger examples for much less running time. For example our Julia implementation of Algorithm 3 competes in 66.28 seconds on a complete random network with $n = 1000$ number of vertices.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, 4th ed. Cambridge, Massachusetts: MIT Press, 2022, ch. 21, pp. 585–603. ISBN 0-262-04630-X

[2] R. C. Prim, "Shortest connection networks and some generalizations," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957. doi: 10.1002/j.1538-7305.1957.tb01515.x

[3] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *Journal of the ACM*, vol. 34, no. 3, pp. 596–615, July 1987. doi: 10.1145/28869.28874

[4] R. M. Ramos, S. Alonso, J. Sicilia, and C. González, "The problem of the optimal biobjective spanning tree," *European Journal of Operational Research*, vol. 111, no. 3, pp. 617–628, 1998. doi: 10.1016/S0022-0000(05)80064-9

[5] D. Rocha, E. Goldbarg, and M. Goldbarg, "A memetic algorithm for the biobjective minimum spanning tree problem," in *Evolutionary Computation in Combinatorial Optimization*, J. Gottlieb and G. R. Raidl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. doi: 10.1007/11730095_19. ISBN 978-3-540-33179-7 pp. 222–233.

[6] S. Steiner and T. Radzik, "Computing all efficient solutions of the biobjective minimum spanning tree problem," *Computers & Operations Research*, vol. 35, no. 1, pp. 198–211, 2008. doi: 10.1016/j.cor.2006.02.023

[7] A. C. Santos, D. R. Lima, and D. J. Aloise, "Modeling and solving the bi-objective minimum diameter-cost spanning tree problem," *Journal of Global Optimization*, vol. 60, pp. 195–216, 2014. doi: 10.1007/s10898-013-0124-4

[8] H. W. Corley, "Efficient spanning trees," *Journal of Optimization Theory and Applications*, vol. 45, pp. 481–485, 1985. doi: 10.1007/BF00938448

[9] S. Fidanova and M. Ganzha, "Ant colony optimization for workforce planning with hybridization," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. IEEE, 2023. doi: 10.15439/2023F9586 p. 955–959. [Online]. Available: http://dx.doi.org/10.15439/2023F9586

[10] L. M. Laskov and M. L. Marinov, "Pareto optimal solutions of the biobjective minimum length minimum risk spanning trees problem," in *Proceedings of the of the 19th Conference on Computer Science and Intelligence Systems*, ser. ACSIS, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39. IEEE, 2024. doi: https://doi.org/10.15439/2024F2913. ISSN 300-5963 pp. 405–416.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 727–732
DOI: 10.15439/2025F9112
ISSN 2300-5963 ACSIS, Vol. 43

Towards Optimal Train Routing Using Microscopic Simulation on Moving Block Controlled Networks

Severin Lochschmidt, Stefan Engels* and Robert Wille
—, 0000-0002-0844-586X and 0000-0002-4993-7860
Chair for Design Automation, Technical University of Munich (TUM), 80333 Munich, Germany
Email: {severin.lochschmidt, stefan.engels, robert.wille}@tum.de

***Abstract*—The demand for sustainable railway transportation is increasing over time. At the same time, the capacity of railway networks is limited. Hence, efficient algorithms for generating optimal timetables are of great interest. Previous research focuses on trains being separated by classical fixed block signaling systems. With modern control systems based on moving block, e.g., within the *European Train Control System* (ETCS), the principles of safely separating trains change significantly. Only limited research on optimal routing on such modern railway networks exists. With this work, we propose a simulation approach tailored to be used with heuristic optimization algorithms to tackle this problem. Moreover, we show how such a framework can allow for more general inputs to jointly optimize what is usually planned sequentially as of today. The simulation framework is included within the open-source *Munich Train Control Toolkit* (MTCT) available on GitHub at https://github.com/cda-tum/mtct.**

I. Introduction

PASSENGER rail traffic in the European Union (EU) has increased by 35% in the last 30 years [1]. The main driver of this increase is high-speed rail traffic, which has risen on average by 6% per year. At the same time, freight rail traffic volume has stagnated, but, triggered by the European Green Deal, freight traffic should be prioritized to be transported by rail instead of road [2]. However, many railway lines are already operating at their capacity limit. A good planning process and traffic management systems are crucial to cope with the increasing demand. Some of the arising design tasks are introduced in [3].

The general railway planning process is commonly split into multiple sub-problems. Solving these sub-problems sequentially makes it possible to reduce complexity drastically. While mathematical modeling of these singular steps is feasible and widely implemented, an integrated approach remains out of reach [4], [5].

In this work, we focus on timetabling on a predefined railway network, or more precisely, how to optimally route trains through a railway network to fulfill certain conditions. The feasibility of a solution highly depends on the implemented control system. Since trains cannot operate on sight due to long braking distances, such signaling systems are crucial to prevent collisions and ensure safe operation. Classically, block signaling systems have been implemented. However, new train control principles have also been defined considering the increasing demand. One of these extensions is the introduction of *Moving Block* control systems, replacing classical block signaling with dynamic supervision. Relevant details on these systems are reviewed in Section II.

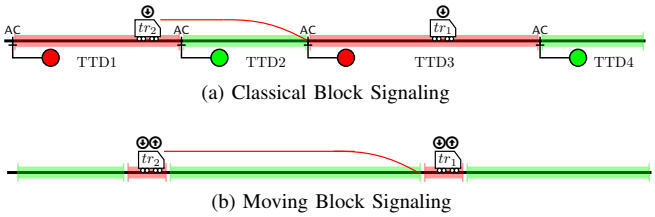
Much of the existing literature for train routing relies on the properties of fixed-block signaling. There is only limited research on optimally routing trains through networks equipped with modern moving block control systems [6], [7], [8]. They are based on Mixed Integer Linear Programming and do not (yet) scale well. We aim to close this gap by introducing an alternative heuristic approach based on simulation on a microscopic level, i.e., considering individual tracks. Existing simulation tools, such as OpenTrack [9] or OSRD [10], mainly focus on verifying and evaluating given solutions. Instead, our work focuses on applying a simulation framework already within the optimization process. Its architecture is built from the ground up with a moving-block approach, ditching old constraints. This allows us to apply novel techniques to encode solutions more efficiently to improve the performance of respective search algorithms applied to our simulation framework. Our approach can even include more complicated objectives. By doing so, it could be extended to combine multiple planning steps jointly in one optimization step. All work is integrated into the *Munich Train Control Toolkit* (MTCM) available open-source on GitHub at https://github.com/cda-tum/mtct.

For this Section II reviews the basic principles of train control systems, Section III describes the routing problem, Section IV introduces the underlying model and encoding with the aim of reducing the search space, Section V provides information on how this model can be used for optimization purposes, and, Section VI evaluates the approach by conducting a case study on a small benchmark set. Finally, Section VII concludes this work.

II. Train Control Principles

Train signaling has been based on splitting a line into discrete segments or blocks (fixed block), going back to the 19th century, Once a train enters a block, it is marked occupied, and access to other trains is forbidden. The status of a line segment is indicated to the driver via a mechanical display, called a semaphore, or later, an optical signal. This ensures mutual exclusion for track segments and thus prevents train collisions. *Trackside Train Detection* (TTD) uses electrical

*Corresponding author.



(a) Classical Block Signaling

(b) Moving Block Signaling

Fig. 1: Schematic drawings of signaling principles [8]

circuits embedded in the tracks or discrete sensors triggered by passing train axles to define the signal status.

Example II.1 (from [8]). *Consider two trains following each other on a single track as depicted in Figure 1a. Train tr_2 can only move until the end of $TTD2$. It cannot enter $TTD3$ because it is still occupied and, hence, might have to slow down in order to be able to come to a full stop before entering the occupied block section.*

With advances in communications technology, it has become possible to soften the traditional block structure into virtual and moving blocks. *Virtual Subsections* (VSS) obey the same principles as fixed blocks but can be arbitrarily defined. Position reporting relies no longer on trackside equipment but continuous bidirectional communication with the train. This alleviates the previously mentioned problems, especially with small or large block sizes, and allows for changing the block layout at any time.

Moving Block is a further evolution where blocks are abolished altogether. All trains in the network self-report and receive each other's position and speed. This allows them to calculate safe headway times dynamically, considering their individual braking curves. Moving Block has been taken up first on suburban railways, where risks are easier to manage. In metro lines, technical efficiency was improved by 11.5% by introducing moving block signaling and connected automation [11].

Example II.2 (from [8]). *In contrast to Eample. II.1, consider a moving block control implemented in Figure 1b. Because trains operate at the ideal absolute braking distance, tr_2 can move up to the actual end of tr_1 (minus a little buffer). In particular, it can already enter what has been $TTD3$ previously. Hence, trains can follow each other more closely.*

III. Train Routing under Moving Block Control

This work considers optimal train routing on a network equipped with a moving block control system. The *Train Timetabling Problem* (TTP) is part of the previously described planning process in Section I. It aims to find a feasible schedule and train routing under various constraints imposed by requirements on the desired timetable and safety conditions imposed by the control system as discussed in Section II. Train routing on a microscopic level consists of assigning trajectories and timings at the track level without imposing conflicts between different train movements.

In this work, we consider a part of a network that is equipped with a moving block control system, i.e., the problem is given as:

Problem III.1 (Train Routing under Moving Block Control). ***Given:*** *A railway network, a list of trains including relevant properties, and a set of demands for every train consisting of*

- *information on when and where the train enters the network,*
- *information on when and where the train exits the network, and*
- *a list of stations*[1] *the train should stop at.*

Problem: *Find a feasible train routing that minimizes overall travel time.*

IV. Simulation Model

This section describes the underlying model of our microscopic time-continuous simulator. First, we briefly discuss how the railway network and train in Sections IV-A and IV-B. Finally, an efficient solution encoding is discussed with the aim of preventing invalid solutions from being generated by the optimizers presented in Section V.

A. Network and Train Model

Our network model is based on [3]. At the same time, we allow trains to move on railway tracks in arbitrary directions. Thus, a railway network is an undirected graph $G = (V, E)$ with vertices $V = \{v_1, v_2, v_3, \ldots, v_m\}$ representing railway switches, signals or other points of interest and edges $E = \{e_1, e_2, e_3, \ldots, e_m\}$, $e_i = \{v_j, v_k\} \subset V$ representing railway tracks. Each edge e_i also has an associated positive weight l_{e_i} representing track length and a maximum permitted speed $v_{e_i}^{max,edge}$. Moreover, railroad switches can only be traversed in specific directions, which is also modeled within the railway network.

B. Train Model

Trains are objects of a certain length l^{train}, maximal velocity $v^{max,train}$, and maximal acceleration/deceleration[2] $a^{max,train}$. We allow negative velocity values to include directional information. If $v > 0$, the train moves forward; if $v < 0$, it moves backward. In contrast to previous work [6], [7], [8], this allows trains to also turn around if necessary, e.g., in stations where this commonly happens in practice. Without loss of generality, we define a train's position as its center point and denote by $e[t] \in E$ the edge at time t

C. Solution Encoding

Choosing a purely heuristic approach affords us freedom in the design of the entire optimization loop. This makes it possible to limit the search space by intelligently preprocessing our candidate solutions. We choose a representation

[1]This only includes the station. The platform is not yet fixed, but it is to be decided by the solving algorithm.

[2]For simplicity, the train's maximal acceleration and deceleration are equal. However, the model can also be extended to model that these are usually different since braking and accelerating behavior differ.

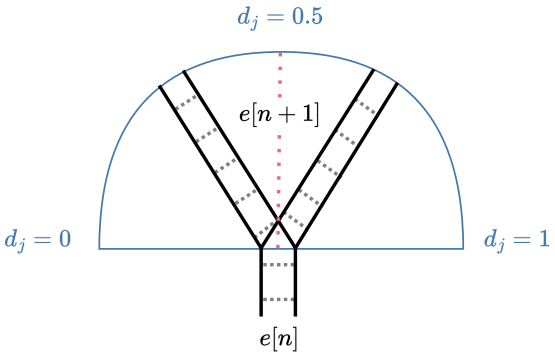


Fig. 2: Directional choice variable.

such that for each train, the decision variables map onto a physically possible train trajectory. This way, we eliminate solutions that contain overspeed, overacceleration, and invalid paths.

1) Railway Switches: Since trains in our model can reverse directions at any time, their trajectory describes an open, undirected walk. Instead of generating a valid graph walk for each solution, we compute it implicitly at simulation time with no additional cost. For this, we introduce directional choices $D = \{d_1, \ldots, d_{N^{max}_{transitions}}\}$, where each $d_j \in [0, 1]$ mimics a steering wheel for a given decision point. If $d_j = 0$, the train wants to move left; if $d_j = 1$, the train wants to move right; see Figure 2. Because $d_j \in \mathbb{R}$, this allows us to also choose middle tracks if a given switch allows for more than two options. At the first switch, the route is chosen according to d_1, at the second switch according to d_2, and so on. That means the route choice variables do not depend on time but are a list of decisions that are made in order whenever it is necessary to make a decision.

More precisely, assume that at time t, a train traverses a switch v_{trav} on edge $e[t]$ and has to make its kth decision. Assume that the list of possible successor edges allowed by the switch $\Gamma(e[t], v_{trav})$ is ordered and contains j elements $e_0, e_1, \ldots, e_{j-1}$. Then

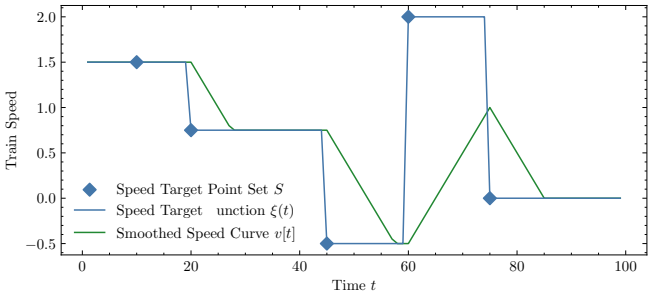
$$e[t + \Delta t] = e_{\lfloor (d_k \cdot (j-1)) \rceil}, \tag{1}$$

where $\lfloor \cdot \rceil$ denotes rounding to the closest integer.

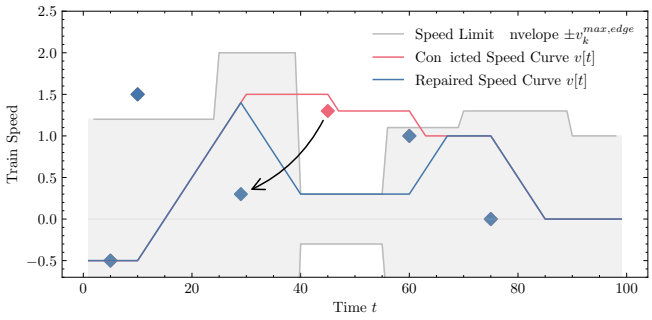
2) Speed and Acceleration: We do not directly encode each train's velocity (and acceleration) because modeling future speed limits (which might induce braking well in advance) would be challenging to model. Instead, we choose a set of speed target points S as decision variables, as time-velocity pairings. More precisely, any $s_i = (\tau_i, \nu_i) \in S$ consists of a time point $\tau_i \in [0, T_{max}]$ and a normalized velocity target $\nu_i \in [0, 1]$. From S we create a piecewise constant function $\xi(t)$ by previous-neighbor interpolation and scaling with $v^{max,train}$, i.e.,

$$\xi(t) = v^{max,train} \cdot \nu_{previous(S,t)}, \tag{2}$$

where $previous(S, t) := \max\{i \colon \tau_i \leq t\}$ is the most current speed target. From this, we induce the acceleration by attempting to reach the current speed target as quickly as possible, see Figure 3a, similar to a bang-bang controller known from optimal control theory. If $\left|\frac{\xi(t) - v[t]}{\Delta t}\right| \leq a^{max,train}$, then $v[t + \Delta t] = \xi(t)$, otherwise the train maximally accelerates toward its target.



(a) Transformation of speed decision variables into continuous curve.



(b) Single iteration of the track speed limit repair algorithm.

Fig. 3: Smoothing ξ to a feasible trajectory.

The track speed limit is much more difficult to encode since train speed $v[t]$ and edge position $e[t]$ are interdependent. Adjusting the speed can shift the time of an edge change or, when reversing, even the edges themselves. We leverage this by repairing $v[t]$ during the simulation using backtracking, a standard method for algorithmic constraint satisfaction as shown in Figure 3b. When encountering a constraint violation, the offending speed targets in S are removed and replaced with the maximum allowed speed for the track section. The simulation travels backward in time and applies just enough braking to avoid overspeed. Similarly, we can force the train to stop at a given point, e.g., at a station.

Encoding the target velocity curve as a set of discrete points in this way reduces the dimensions of the solution space. Any arbitrary curve in the solution space is still representable by increasing the number of points until every timestep is contained in S, amounting to a lookup table.

V. Optimization using Simulation

The combined encoding from Section IV-C gives us a space of decision variables consisting of speed target points S and direction choices D that map onto individually feasible train trajectories as previously discussed. Next, we evaluate the

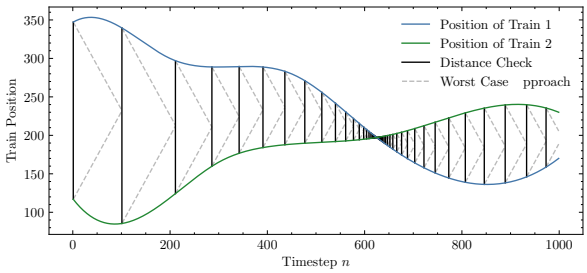


Fig. 4: Detection algorithm for headway constraint violations.

quality of each solution to define optimization criteria and apply various heuristic algorithms to find reasonable solutions.

A. Objective Function

The objective function is a composite of soft constraints $f_{headway}, f_{dest}$, and a global objective f_{stops}, which are explained next. They form the multi-objective minimization problem. Each objective can easily be normalized and weighted as desired.

1) Headway: Arguably, the most critical constraint is to avoid collisions, i.e., to apply long enough headways between two consecutive trains. It is introduced into the objective as a soft constraint $f_{headway}$ penalizing closeness beyond a minimum safety distance $h_{safe} \geq 0$.

Let $d_{i,j}(t)$ be the distance between tr_i and tr_jat time t. Additionally, a distance penalty function ω is chosen, which is zero for distances greater than h_{safe}, for example, a triangle function. The headway penalty can then be defined as

$$f_{headway} = \sum_{i=1}^{N_{trains}} \sum_{j=i+1}^{N_{trains}} \sum_{t} \omega(d_{i,j}(t)). \tag{3}$$

Using pre-computed all-pairs shortest paths for the entire network, the effort to find the distance between two trains is constant. The total number of necessary checks can be reduced by skipping time steps where a collision is physically impossible. For this, not that $d_{i,j}$ fulfills the following constraints:

$$\begin{aligned} & d_{i,j}(t) \geq 0, \\ & d_{i,j}(t) \leq longest\ distance\ in\ network, \\ & |d'_{i,j}(t)| \leq \left|v_i^{max,train}\right| + \left|v_j^{max,train}\right|, \text{ and} \\ & |d''_{i,j}(t)| \leq \left|a_i^{max,train}\right| + \left|a_j^{max,train}\right|. \end{aligned} \tag{4}$$

Using Equation (4), we can calculate a grace period for each distance check in which the trains cannot meet even when taking the shortest path towards each other at full speed. As seen in Figure 4, more checks are performed when trains are closer. Consequently, the computational effort for this objective varies based on the density and maximum speed of trains on the network. This method can be further improved by considering acceleration constraints in the grace period calculation and using a more efficient sampling of the distance function.

2) Destination: The problem specifies that each train enters and exits the network under consideration at some point in time with respective target velocities. The former constraint can be fulfilled by fixing the initial state of each train. Guidance toward a final destination is added through another soft constraint, penalizing the distance from the desired position and the difference in speed.

3) Intermediate Stops: Stops en route can be introduced by rewarding time spent with zero speed on scheduled edges. However, we can increase the prevalence of solutions with valid stops by encoding them using the repair algorithm described in Section IV-C. Each train passing an edge with a stop will be forced to halt for a minimum amount of time. In the objective evaluation, the fulfilled stops can be counted at minimal cost.

B. Optimizer

As described in Section IV, we are confronted with a general, non-linear, constrained optimization problem. In addition, the objective is non-convex and non-smooth, and its analytic form is unknown. We already defined an unconstrained version via relaxation using penalty functions; see Section V-A. This no longer guarantees solution feasibility for the combined objective. However, feasibility can be verified by evaluating the penalty functions.

The formulation gives rise to applying various heuristic approaches to find a good solution, which are described in the following paragraphs.

1) Random Search: A trivial baseline for search can be achieved by randomly assigning decision variables and keeping the best result. Such a method can be improved by introducing local search steps in each iteration. , e.g., using the Luus-Jakola method, a simple gradient-free local search algorithm [12]. An initial solution is randomly perturbed in a hypersphere with radius r. If the new candidate solution is an improvement, it replaces the old one as a starting point. The radius r contracts with each iteration by a contraction factor $c_{contract}$.

2) Greedy Search: It can be beneficial to mimic the behavior of human timetable planners when generating candidate solutions, as proposed in [13]. We, therefore, place trains sequentially using a limited random search for each instance in sequence. This approach breaks down the routing into subproblems. The order of train placement in our implementation is randomized. Classic asynchronous routing usually schedules trains by priority. Again, such an algorithm could be improved using local search, leading to the *Greedy Randomized Adaptive Search Procedure* (GRASP).

3) Genetic Algorithm: The use of genetic algorithms for timetabling is widespread [14], [15]. We implemented a primitive version, where crossover amounts to randomly choosing a trajectory for each train from the two parent solutions. Again, the iterations can be combined with local search. The population is sorted by objective score. Individuals are transferred from an elite set of top performers to the next

TABLE I: Search method comparison parameters.

Local Search and random+local and GRASP	Start Radius	$sr_{initial}$	0.4
	Stop Radius	sr_{final}	0.001
	Contraction Coefficient	$c_{contract}$	0.99
Greedy Search and GRASP	Per Train Stall Time		10ms
Genetic Search	Population	$\|P\|$	1000
	Elite Fraction	$n_{elite}/\|P\|$	0.01
	Mutation Rate		0.1
	Crossover Fraction	$n_{crossover}/\|P\|$	0.7

generation without modification. The crossover fraction determines the subset of individuals selected as parents for the next generation. Each offspring has a uniform chance to mutate, called the mutation rate; in this case, a random perturbation of all variables proportional to their maximum range is applied. The size of this perturbation is scaled down non-linearly as the generations progress. Our implementation uses the *openGA* library [16].

VI. Case Study

Due to their different model assumptions and significant limitations on publicly available data, the performance of timetabling tools is difficult to compare quantitatively. Ultimately, the most decisive measure is usefulness when applied in day-to-day planning. With this case study, we demonstrate the applicability of the proposed framework to the early stages of timetable planning.

A. Benchmark

For this, we solve three sample problems of varying size from [17]. These networks are included in the aforementioned *Munich Train Control Toolkit* (MTCT) available open-source. The first network, "Overtake," models a situation where faster trains must overtake slower ones. The second "SimpleNetwork" makes four trains cross at a central station. The last network, "Stammstrecke," is a full-sized replica of the Munich S-Bahn's core network passing through the city center.

Each train is defined with a fixed schedule for arrival and departure in the network and planned stops. In our test case, all objective functions are designed such that an objective score of zero indicates an optimal solution. This means all trains obey the distance minima, visit their scheduled stops, and reach their destination at the right speed and time. The collision objective is weighted roughly 10 to 1 against the two others.

Table I shows the parameters chosen to compare all search methods. They are a compromise across all test networks drawn from previous parameter measurements not reported in this paper due to space limitations. The numeric results for the search parameters we obtained are not universal since they depend heavily on problem parameters. However, we attempt to make generalized statements about algorithm performance wherever conclusive measurements allow.

B. Setup

Our C++ implementation is integrated into the open-source Munich Train Control Toolkit available on GitHub at https://github.com/cda-tum/mtct. Each optimizer instance was run on a single 4GHz core of a stock Intel i7-6700K processor using DDR4 RAM at 2133 MHz. For each algorithm instance, we record each improvement in objective score along with the expired search time. After collecting a statistically significant number of runs, we interpolate at 500 evenly spaced points over the runtime of the longest run. We then calculate the arithmetic mean score at each point. These mean score progressions can then be compared to assess algorithm quality.

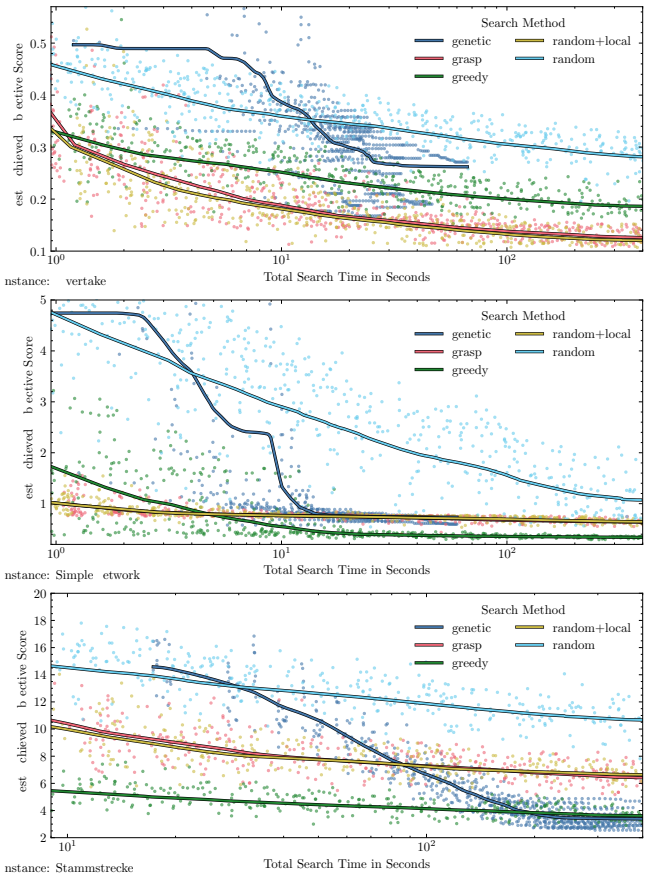


Fig. 5: Search Methods Effectiveness Measurement.

C. Evaluation

We compare the optimizers introduced in Section V-B. For the genetic algorithm, adding a local search step for each generation did not improve outcomes while increasing search time by an order of magnitude. For this reason, we excluded local improvement from the results. For our sample problems, all search methods stopped improving within the investigated time period. Thus, we focus on the best score before stalling, primarily reflecting solution quality.

Figure 5 pits all optimizers against each other. The random search baseline performed worst in all cases, as expected. The two local search methods dominated for the smallest instance. Focusing on refinement is more fitting for a smaller fraction of infeasible solutions. Greedy search was the overall best,

delivering both fast results and a low floor over time. It can efficiently generate many feasible solutions but will drop them immediately. The genetic algorithm showed progressively improved results for larger problem sizes, even overtaking greedy search in final solution quality for the Stammstrecke instance.

By tracking individual objectives and visualizing the train trajectories, we can assess the quality of the solutions for these candidates. The objective values are normalized to $[0, 1]$ or $[0\%, 100\%]$ respectively. The best solutions for the SimpleNetwork and Overtake instances exhibit no distance violations, an average destination penalty of 7% and 11%, and visit all reachable stops. The Stammstrecke solution does contain collisions, a 34% average distance penalty, and 27% average unfulfilled stops.

VII. Conclusions

We have reviewed current developments in railway planning and shown the need for modern, flexible, and accessible simulation platforms. We presented a new microscopic simulation framework for train timetabling on networks equipped with modern signaling systems based on moving block separation. This allows for more flexible train movements than previous approaches. We explained how architectural decisions and constraint encoding could be chosen to help heuristic algorithms find feasible trajectories quickly by implementing repair heuristics. Our approach allows for flexible objectives so that, e.g., line planning could be included in a joint optimization using the presented framework.

We tested multiple optimizer configurations on example timetabling problems to demonstrate functionality and gain insights into problem structure. Greedy and population-based search strategies proved superior to neighborhood-focused ones. For small instances, this trend was reversed. Overall, diversification, in combination with constraint satisfaction, dictated solution quality rather than local refinement. The smaller routing instances could be solved within a few minutes and withstand hard constraints even with limited hardware.

At the same time, this hints that the repair step introduced in the encoding to enforce feasible trajectories significantly helps the heuristics progress. Note that not all solutions firmly adhered to the headway constraints. Hence, the approach might improve from an enhanced solution encoding, which ensures safe distances by a backtracking repair step similar to the one proposed to ensure satisfying track maximum speeds. It is conjectured that this will highly benefit the quality of the solution.

Overall, this work constitutes a promising approach for using simulation frameworks to generate optimal timetables and microscopic routes on general railway networks equipped with modern moving block control systems. All code is available open-source within the *Munich Train Control Toolkit* (MTCT) on GitHub at https://github.com/cda-tum/mtct, which is still under active development.

References

[1] European Commission and Directorate-General for Mobility and Transport, *Transport in the European Union: Current trends and issues.* Publications Office of the European Union, 2024. [Online]. Available: https://doi.org/10.2832/131741

[2] European Parliament, "Position of the European Parliament and of the Council on the use of railway infrastructure capacity in the single European railway area," 2024, EP-PE_TC1-COD(2023)0271. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TC1-COD-2023-0271_EN.pdf

[3] S. Engels, T. Peham, J. Przigoda, N. Przigoda, and R. Wille, "Design tasks and their complexity for the European Train Control System with hybrid train detection," *EURO Journal on Transportation and Logistics*, vol. 14, p. 100161, 2025. [Online]. Available: https://doi.org/10.1016/j.ejtl.2025.100161

[4] R. M. Lusby, J. Larsen, M. Ehrgott, and D. Ryan, "Railway track allocation: models and methods," *OR Spectrum*, vol. 33, no. 4, pp. 843–883, dec 2009. [Online]. Available: https://doi.org/10.1007/s00291-009-0189-0

[5] N. Absi, C. Archetti, S. Dauzère-Pérès, and D. Feillet, "A two-phase iterative heuristic approach for the production routing problem," *Transportation Science*, vol. 49, no. 4, pp. 784–795, 2015. [Online]. Available: https://doi.org/10.1287/trsc.2014.0523

[6] T. Schlechte, R. Borndörfer, J. Denißen, S. Heller, T. Klug, M. Küpper, N. Lindner, M. Reuther, A. Söhlke, and W. Steadman, "Timetable optimization for a moving block system," *Journal of Rail Transport Planning & Management*, vol. 22, p. 100315, 2022. [Online]. Available: https://doi.org/10.1016/j.jrtpm.2022.100315

[7] T. Klug, M. Reuther, and T. Schlechte, "Does laziness pay off? - a lazy-constraint approach to timetabling," in *22nd Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2022)*, vol. 106. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. [Online]. Available: https://doi.org/10.4230/OASIcs.ATMOS.2022.11

[8] S. Engels and R. Wille, "Comparing lazy constraint selection strategies in train routing with moving block control," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, vol. 39. Polish Information Processing Society, 2024, pp. 585–590. [Online]. Available: https://doi.org/10.15439/2024f3041

[9] Hürlimann, Daniel, "Objektorientierte Modellierung von Infrastrukturelementen und Betriebsvorgängen im Eisenbahnwesen," Ph.D. dissertation, 2002. [Online]. Available: https://doi.org/10.3929/ETHZ-A-004223151

[10] OSRD contributors. Open source railway designer. [Online]. Available: https://osrd.fr

[11] S. Canavan, D. J. Graham, P. C. Melo, R. J. Anderson, A. S. Barron, and J. M. Cohen, "Impacts of moving-block signaling on technical efficiency: Application of propensity score matching on urban metro rail systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2534, no. 1, pp. 68–74, 2015. [Online]. Available: https://doi.org/10.3141/2534-0

[12] R. Luus, *Direct Search Luus–Jaakola Optimization Procedure*. Springer Nature Switzerland, 2024, pp. 1–5. [Online]. Available: https://doi.org/10.1007/978-3-030-54621-2_129-1

[13] K. Nachtigall and S. Voget, "A genetic algorithm approach to periodic railway synchronization," *Computers & Operations Research*, vol. 23, no. 5, pp. 453–463, 1996. [Online]. Available: https://doi.org/10.1016/0305-0548(95)00032-1

[14] N. Bešinović, R. Goverde, G. Nicholson, and H. Steele, "Review of railway timetabling developments: planning phases, goals and future directions," 2021. [Online]. Available: https://www.researchgate.net/publication/353378414_Review_of_railway_timetabling_developments_planning_phases_goals_and_future_directions

[15] F. Hauck, "An overview and categorization of approaches for train timetable generation," 2023. [Online]. Available: https://doi.org/10.17169/REFUBIUM-37849

[16] A. Mohammadi, H. Asadi, S. Mohamed, K. Nelson, and S. Nahavandi, "OpenGA, a C++ genetic algorithm library," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 2051–2056. [Online]. Available: https://doi.org/10.1109/smc.2017.8122921

[17] S. Engels, T. Peham, and R. Wille, "A symbolic design method for ETCS Hybrid Level 3 at different degrees of accuracy," in *23rd Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2023)*, vol. 115. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. [Online]. Available: https://doi.org/10.4230/OASICS.ATMOS.2023.6

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 733–738
DOI: 10.15439/2025F5355
ISSN 2300-5963 ACSIS, Vol. 43

A New Dimension of Acceptance? Introducing Perceived Public Value as Extension of UTAUT in the Smart City Context

Alisa Lorenz
0000-0002-8547-1391
Technische Hochschule Mittelhessen – THM Business School, Wiesenstr. 14, 35390 Giessen, Germany;
University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany
Email: alisa.lorenz@w.thm.de

***Abstract*— Smart city technologies aim to address urban challenges by leveraging digital innovation to contribute to livability. However, their acceptance by citizens remains a critical challenge. Technology acceptance models such as the Unified Theory of Acceptance and Use of Technology (UTAUT) assess individual benefits, overlooking community-oriented motivations. This study proposes expanding UTAUT by introducing perceived public value (PPV) as a new dimension that reflects collective benefits that individuals perceive for their community. Drawing on public value theory, we synthesize a concept of public value and evaluate UTAUT's capability to reflect resulting public value dimensions. Our findings suggest that integrating PPV as additional dimension in UTAUT could enhance understanding of technology acceptance in the smart city context and lays the groundwork for empirical validation. This research contributes to both research and practice by advocating for the inclusion of public value considerations in smart city development, aiming for more sustainable urban development.**

***Index Terms*—Smart city, technology acceptance, public value, UTAUT, sustainable development.**

I. Introduction

THE increasing development of smart city solutions is crucial for addressing urban challenges such as overpopulation, space shortages, and pollution, with the goal to provide a livable environment for citizens [1], [2], [3]. By leveraging the advantages of information and communication technology (ICT), these solutions can offer advantages across various domains, including smart economy, smart environment, smart mobility, smart governance, smart living, and smart people [4].

However, smart city applications can only fulfill their purpose if they are widely accepted and adopted by citizens [5]. The most commonly used models to evaluate technology acceptance are the technology acceptance model (TAM) and the unified theory of acceptance and use of technology (UTAUT) [6], [7], [8], [9]. However, these models are primarily used in the context of private organizations and focus on evaluating personal utility and use. In contrast, smart city applications serve a broader, more heterogeneous user base composed of diverse citizens. These applications aim to benefit the entire population and are even more dependent on citizen acceptance, as citizens are both key stakeholders and end-users [10]. Despite the growing relevance of smart city applications, their potential societal and sustainability benefits, and the need for broad acceptance, existing technology acceptance models and their adaptations have largely measured only individual advantages. To our knowledge, they have not yet assessed the societal value of such applications. In smart cities, technology is not merely a personal tool that provides individual value but a communal resource that can enhance collective well-being and support sustainable development goals (SDGs) [11], [12]. This shortcoming highlights the necessity for an integrated framework that considers both individual and community-oriented motivations in technology acceptance.

Although not originally focused on technology, the concept of public value provides a relevant framework for evaluating the societal contributions of public services [13]. Public value can help understand the societal contributions of organizations or activities which primarily result from government performance. Given that smart city initiatives are often driven and funded by government, public value could offer a valuable perspective for identifying additional acceptance dimensions centered on societal benefit. Similar to previous research on service acceptance models that included sustainability and quality of life as important factors, user acceptance of smart city applications may extend beyond individual utility [14].

Drawing on public value theory and research on technology acceptance, this article thus answers the following research questions:

RQ 1: How can public value be conceptualized in the context of smart city applications?

RQ 2: To what extent do UTAUT variables reflect public value dimensions?

RQ 3: How could UTAUT be adjusted to reflect public value dimensions?

Although this article does not provide empirical evidence yet, we bridge public value with technology acceptance and provide a first conceptualization that paves the way for further research. Our findings contribute to research on technology acceptance and public value for smart cities by introducing an extension of UTAUT that reflects public value. For practice, we highlight the importance of citizens as central stakeholders of smart city solutions that are essential for the acceptance of smart cities.

To answer our research questions, this article is structured as follows. First, we review relevant literature on technology acceptance and public value, followed by a detailed description of our research process. Next, we propose a condensed conceptualization of public value and examine the extent to which UTAUT variables reflect the related dimensions. We then present an extension of UTAUT tailored to smart city applications, introducing the variable *perceived public value* (PPV) as a new dimension designed to capture the collective value smart city applications. Finally, we discuss our findings and outline directions for future research. With this study, we aim to support the integration of public value considerations into the development of smart city solutions, ultimately contributing to more sustainable smart cities that service people, planet, and prosperity.

II. Related Literature

A. Technology Acceptance in Sustainable Smart Cities

UTAUT and its predecessor TAM have been essential in predicting technology adoption behaviors [6], [8], [9]. While TAM focuses on perceived usefulness and ease of use as predictors of the attitude towards using a technology, UTAUT extends TAM by introducing constructs such as performance expectancy, effort expectancy, social influence, and facilitating conditions to forecast behavioral intentions and use [7], [8]. UTAUT has been extended to account for factors influencing use behavior for the consumer market by adding hedonic motivation, price value, and habit [15]. While these models have provided valuable insights into the factors influencing technology acceptance in private organizations, their use has been limited in the smart city context [16], [17], [18].

However, smart city applications differ from enterprise tools in both user base and purpose. Citizens represent a more heterogeneous user group compared to employees of one company, as they vary in terms of demographics (e.g., teenagers and seniors), socio-economic backgrounds (e.g., education levels, income, social status, and living conditions), cultural backgrounds (e.g., languages, values, and norms), as well as user capabilities and tech-savviness. Second, smart city application use is typically voluntary rather than mandated. Smart city solutions aim to contribute to benefits for citizens and thus to livability in the city, making cities inclusive, safe, resilient, and sustainable [12]. Despite these distinctions, prior acceptance studies have focused primarily on individual performance benefits [19], [20]. While UTAUT incorporates social influence, it largely addresses personal acceptance drivers rather than broader expectations of technology serving the public good, benefits for the community, or societal gain. Studies examining additional variables such as trust and government support have produced mixed results and did not assess broader societal value [21], [22], [23]. Similar to previous research on service acceptance models that included sustainability and quality of life as important factors, user acceptance in UTAUT might reach beyond individual benefits [14]. In sum, to our knowledge, no study has yet examined whether perceived community or societal value also has an influence on citizen's intention to use smart city applications.

B. Public Value

The concept of public value emerged in the 1990s in government research and provides a framework for evaluating the contributions of organizations or activities to society and, as such, to common good [13]. Since then, it has evolved in several directions that conceptualize public value from strategical-managerial, psychological, and normative perspectives. Further research also detailed public values, including human dignity, citizen involvement, or integrity [24].

Moore states that "the core idea is a simple one: public managers should be focused on 'creating public value' from the assets entrusted to them by the public." (Moore, 2021, p. 1). He conceptualizes public value from a managerial and strategic perspective as a strategic triangle encompassing legitimacy and support, operational capability, and public value [13]. Bozeman expands the idea from a normative perspective and defines public values as "those providing normative consensus about (a) the rights, benefits, and prerogatives to which citizens should (and should not) be entitled; (b) the obligations of citizens to society, the state, and one another; and (c) the principles on which governments and policies should be based" (Bozeman, 2007, p. 13). In this view, public values go beyond what the public wants, being collective and socially constructed and not reducible to individual preferences. The concept is further expanded with the introduction of the Public Value Scorecard, which draws on psychological needs theory to evaluate how individuals perceive public value across five organizational dimensions: task fulfillment, morality, quality of life, social cohesion, and profitability [27]. Faulkner and Kaufman describe public value as outcome-focused and group it in four dimensions, including outcome achievement, which assesses the social and economic impacts of an initiative; trust and legitimacy, which reflect the public's confidence in an organization and its fairness; service delivery quality, denoting client satisfaction and convenience; and efficiency, which considers whether the benefits of an activity outweigh its costs [28].

In sum, while definitions vary, public value aims to ensure that public organizations and managers deliver services and outcomes that are beneficial to the public, by considering efficiency, ethical factors, and social outcomes.

C. Research Gap

Considering the call for more sustainable smart city solutions that are accepted and used by the public, we see the potential of public value as a promising concept to explain technology acceptance beyond individual benefits since smart city applications can be a form of a public service. Despite its broad application in public organizations and first applications in e-Government [29], to our knowledge public value theory has not been incorporated in technology acceptance models yet.

Despite the advantages of TAM and UTAUT in assessing influencing factors of technology acceptance, we identified a gap in their ability to account for the collective notion of technology serving a greater good, or public value. These models, and their expansions, focus predominantly on individual acceptance factors and do not consider the broader societal influence that smart city technology should contribute to. To bridge these gaps, we suggest the incorporation of public value frameworks in existing technology acceptance theory in the context of smart cities. By incorporating public value perspectives, we can better assess the potential of smart city technologies to meet the collective needs of citizens and align with broader goals such as sustainability and livability. Ultimately, balancing individual and collective value while ensuring sustainable and effective implementation may enhance public acceptance of smart city applications. Considering the concept of public value in relation to the call for "IS for the better" [30], the expansion of existing technology models could thus challenge the notion that acceptance is solely based on individual advantages and fill the proposed research gap.

III. Research Process

To answer our three research questions, we structured our research process into three phases (Fig. 1). In the first phase, we ensured a theoretical foundation by identifying and analyzing relevant literature on public value. We selected four key contributions based on two criteria: 1) the number of citations, reflecting tehri relevance to the academic field, and 2) the requirement that they provide a clear conceptualization of public value. We selected the conceptualizations of public value by Moore [13], Bozeman [26], Faulkner and Kaufman [28], and Meynhardt [27]. We then conducted a comparative analysis of their conceptualizations of public value and identified five common dimensions among them, which we mapped in a synthesized public value concept. This synthesis aimed to answer how public value can be conceptualized in the context of smart city applications (RQ 1).

In the second phase, we synthesized the concepts of public value and technology acceptance. We selected UTAUT as technology acceptance model to evaluate public value, as it builds on TAM and captures additional dimensions beyond perceived usefulness and ease of use. We did not use UTAUT2 due to its tailoring to the consumer context, which includes dimensions such as price value and hedonic motivation. As smart city applications are typically offered free of charge and serve a public rather than a consumer-oriented purpose, these extended dimensions were deemed less relevant in our context. Next, we conducted a qualitative comparative analysis of the dimensions of the synthesized public value concept from phase 1 and the UTAUT dimensions. From this, we identified gaps in UTAUT that fail to reflect public value dimensions. The resulted findings addressed the extent to which UTAUT variables reflect public value dimensions (RQ 2).

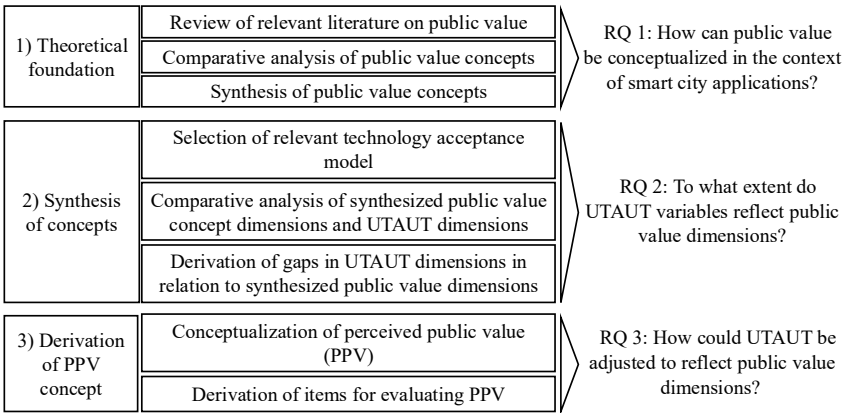


Fig 1. Research Process

In the third phase, we derived a new concept of *perceived public value* that could be added to UTAUT as an additional dimension to assess the influence of public value on the intention to use smart city applications. Furthermore, we derived initial measurement items that could be used in future evaluations of PPV. This phase aimed to answer how UTAUT could be adjusted to reflect public value dimensions (RQ 3). We present the results of this process in the following.

IV. Conceptualization of Perceived Public Value

The result from phase 1 was a synthesized conceptualization of public value. Table I summarizes the results from the comparative analysis of the selected public value concepts.

We first compared how each author defines public value to provide context for their conceptualizations. In the rows, we present the identified similarities between the concepts in five conceptual dimensions of public value:

- Legitimacy and trust,
- operational capacity,
- outcome orientation, efficiency, and profitability,
- citizen engagement and participation,
- and ethical/moral foundation.

In the columns, we assigned the related sources. The resulting table shows how each concept addresses the five conceptual dimensions. Some definitions and concepts might relate to more than one conceptual dimension, depending on the focus. For example, trust and legitimacy as defined by Faulkner & Kaufman (2018) can relate to a distinct dimension with emphasis on the trust in an organization but can also contribute in part to the dimension of ethical and moral foundation when focusing on the fairness of processes.

As a result of phase 2, Table II illustrates the conceptual overlap between the four UTAUT dimensions performance expectancy, effort expectancy, social influence, and facilitating conditions. We conducted a qualitative review of the UTAUT items and public value dimensions to assess overlap and distinction. We did not consider moderating variables of UTAUT. We used a plus sign (+) when UTAUT sufficiently reflects a public value dimension, a minus sign (-) when no overlap was found, and a combination of both (+/-) for partial overlap.

For *legitimacy and trust*, we observed partial overlap with social influence, which captures social or organizational pressure to adopt a system. This may relate to perceived legitimacy but does not reflect trust in institutions or government as a central theme. *Operational capacity* showed partial alignment with facilitating conditions, which refer to technical and organizational support for use. This partly reflects institutional readiness to deliver services. *Outcome orientation, efficiency, and profitability* aligned with performance expectancy, as both emphasize the belief that a system will improve performance. Effort expectancy also overlaps in terms of efficiency, assuming that ease of use contributes to effective service delivery. *Citizen engagement and participation* showed indirect overlap with social influence and effort expectancy. While UTAUT acknowledges social norms, it does not explicitly model participatory or co-creation processes. Effort expectancy also reflects ease of use which is somewhat reflected in service delivery quality as part of client satisfaction. Last, we did not find any overlapping dimension in UTAUT that reflects ethical or moral foundation, highlighting its value-neutral stance. Although several public value dimensions are only partially represented, the complete absence of a moral/ethical foundation stresses a significant conceptual gap between the two frameworks.

As results of phase 3, we developed a new dimension that could be added to UTAUT. While UTAUT captures drivers of individual technology acceptance, its constructs only partially address the broader dimensions of public value. This underscores the need for an expanded or integrated model when evaluating smart city application use. UTAUT assumes a rational individual actor motivated by personal benefit. However, in public administration, both managers and citizens may also be driven by collective, moral, and social values. The narrow individual focus of UTAUT thus overlooks the societal responsibilities and ethical expectations that may shape citizen attitudes towards smart city solutions.

TABLE I.
CONCEPTUALIZATION OF PUBLIC VALUE BASED ON [13], [26], [27], [28]

		Moore (1995)	**Bozeman (2007)**	**Faulkner & Kaufman (2018)**	**Meynhardt (2019)**
	Value definition	What the public values / benefits from	Rights, entitlements, and societal obligations	Achieving desirable outcomes (social, economic, etc.)	Subjective perception of value based on individual experience
Conceptual dimension	**Legitimacy and trust**	Legitimacy and support (political/legal)	Obligations of citizens to society, state, and one another; principles on which governments and policies should be based on	Trust and legitimacy (fair process, trust in organization)	Morality (is the organization decent) and social cohesion (is it acceptable)
	Operational capacity	Organizational ability to deliver	N/A	Efficiency	Implicit in task fulfillment
	Outcome orientation, efficiency, and profitability	Implicit in value creation	Implicit in value creation	Outcome achievement (strong emphasis on achieving social, economic, environmental and cultural outcomes)	Focus on individual (quality of life) and collective (morality, acceptance) well-being, but also on profitability
	Citizen engagement / participation	Implicit in focus on the mission of socially oriented public value	Obligations of citizens to society, the state, and one another	Service delivery quality (client satisfaction and suitable citizen engagement, accessibility, convenience)	Quality of life (positive experience) and social cohesion (satisfying needs for belonging)
	Ethical/moral foundation	Implicit in political and legal backing, benefits and useful outcomes for the public	Shared responsibility of citizens and government	Trust and legitimacy (transparent and fair process), service delivery quality (client satisfaction), Outcomes (social, economic, environmental, cultural)	Morality (satisfaction of moral-ethic needs)

We therefore propose a new dimension called *perceived public value* (PPV), defined as the extent to which an individual believes that a system will create value for the broader community. We chose PPV as a new term for two reasons. First, to distinguish it from prior definitions of public value. Second, to emphasize that public value may extend beyond the individual perspective, which cannot always be directly measured. While citizens likely value the personal benefits of smart city applications (performance expectancy), they may also be motivated by a sense of civic duty or community contributions. In such cases, adoption may stem as much from civic responsibility as from personal convenience.

To assess PPV, we developed six measurement items based on the public value dimensions that were not covered by UTUAT in phase 2 (Table III). We excluded the dimension *outcome orientation, efficiency, and profitability*, as it is sufficiently addressed by performance and effort expectancy. We developed the items based on the style of previous UTAUT questionnaires [6], [21], [23] and added the items according to the identified focus of each public value dimension. The proposed items are intentionally formulated broadly to allow adaptation to specific applications and use cases. We propose that incorporating PPV into UTAUT may improve understanding of citizen behavior in smart city contexts. Doing so acknowledges the civic motivations behind technology use, particularly in cases where collective adoption and use enhances overall system effectiveness.

TABLE II.
PUBLIC VALUE DIMENSIONS AND UTAUT MAPPING

Public Value Dimension (PPV)	**UTAUT Dimensions**	**Conceptual Overlap**
Legitimacy and Trust	Social Influence (some overlap)	+/-
Operational Capacity	Facilitating Conditions (some overlap)	+/-
Outcome Orientation, Efficiency, Profitability	Performance Expectancy, Effort Expectancy	+
Citizen Engagement and Participation	Social Influence (indirect), Effort Expectancy (indirect)	+/-
Ethical/Moral Foundation	Not Applicable	-

TABLE III.
MEASUREMENT ITEMS FOR PERCEIVED PUBLIC VALUE (PPV)

Measurement Item(s)	**Public Value Dimension**
PPV1: The application contributes to the city's objectives which are targeted with the application.	Legitimacy and Trust
PPV2: The application contributes to the main goal (for which it was developed).	Operational Capacity
PPV3: The application contributes to the feeling of community. PPV4: The application contributes to fulfilling citizen's needs beyond individual needs.	Citizen Engagement and Participation
PPV5: The application contributes to sustainability factors and the environment. PPV6: The application contributes to quality of life (in the city).	Ethical/Moral Foundation

V. DISCUSSION AND OUTLOOK

In this article, we introduced PPV as an additional dimension to be integrated into UTAUT to evaluate factors influencing citizens' intention to use smart city applications. Our research has both theoretical and practical implications. From a theoretical perspective, our study bridges a significant gap by proposing a novel way to account for community-centered benefits within technology acceptance models. Modifying UTAUT to incorporate public value may not only enhance its explanatory power in the context of smart cities but also align it more closely with broader sustainability goals. For practice, the inclusion of PPV emphasizes the importance of considering societal benefits in the design and evaluation of smart city applications. Our literature review demonstrated how public value represents a crucial foundation for smart city applications and how it may influence technology acceptance. We therefore encourage stakeholders to integrate assessments of public value early in the planning stages of smart cities to improve future acceptance of such solutions. This approach could promote the development of cities that are not only technologically advanced but also aligned with the SDGs.

Although we carefully conducted our research based on relevant literature and structured analyses, we acknowledge several limitations. First, our study is primarily conceptual and has not yet validated PPV as an extension of UTAUT. Future research should test this dimension and evaluate whether it improves the model's ability to explain citizens' intention to use smart city applications. This includes conceptualizing how PPV integrates into UTAUT and determining its appropriate placement within the model. In fact, we have already taken initial steps toward this goal and plan to report preliminary findings in the near future. Second, we based our comparative analysis in phase 2 on four conceptualizations of public value. While carefully selected, we recognize that our inclusion criteria may have excluded other relevant frameworks. A broader literature review may have yielded additional insights that could have extended the synthesized concept. Third, the perception of public value likely varies across demographic, cultural, and socio-economic contexts of the user base. Future research could explore how these contextual factors influence both the perception of public value and acceptance of smart city applications. Evaluating public value as a dimension of acceptance could help in developing smart city applications that include contributions to public value and thus correspond to citizens' needs, eventually improving future acceptance.

In conclusion, while empirical validation remains necessary, this research represents a foundational step in introducing public value into UTAUT. We contribute groundwork for future studies to empirically test the proposed dimension in real-world settings. We also emphasize the importance of recognizing the collective benefits of technology adoption, aligning with the broader objectives of sustainable development as key responsibility and goal of public organizations. Our work encourages further exploration of the intersection of public value and technology acceptance, urging researchers and

practitioners to consider the broader impacts of technology acceptance. The integration of public value considerations offers further opportunities for information systems research aimed at societal benefit. In this context, we expand upon our previous work on the relevance of citizen engagement for technology acceptance [31], [32].

We hope that our future empirical work on PPV will offer additional evidence for the importance of community-oriented perspectives. Through this lens, smart cities can evolve to become centers of technological advancement that while also serving citizen's needs and prioritizing sustainability, creating a more livable future for all.

REFERENCES

[1] V. Morabito, "Big Data and Analytics for Government Innovation," in *Big Data and Analytics*, Cham: Springer International Publishing, 2015, pp. 23–45, https://doi.org/10.1007/978-3-319-10665-6_2

[2] T. Chen, J. R. Gil-Garcia, and M. Gasco-Hernandez, "Understanding social sustainability for smart cities: The importance of inclusion, equity, and citizen participation as both inputs and long-term outcomes," *Journal of Smart Cities and Society*, vol. 1, no. 2, pp. 135–148, June 2022, https://doi.org/10.3233/SCS-210123

[3] H. Chourabi et al., "Understanding Smart Cities: An Integrative Framework," in *Proc. of the 2012 45th Hawaii International Conference on System Sciences*, HICSS 2012, IEEE, 2012, https://doi.org/10.1109/HICSS.2012.615

[4] R. Giffinger and G. Haindlmaier, "Smart cities ranking: an effective instrument for the positioning of the cities?," *ACE: Architecture, City and Environment*, vol. 4, no. 12, pp. 7–26, Feb. 2010, https://doi.org/10.5821/ace.v4i12.2483

[5] C. Del-Real, C. Ward, and M. Sartipi, "What do people want in a smart city? Exploring the stakeholders' opinions, priorities and perceived barriers in a medium-sized city in the United States," *International Journal of Urban Sciences*, vol. 27, no. sup1, pp. 50–74, January 2023, https://doi.org/10.1080/12265934.2021.1968939

[6] V. Venkatesh, M. G. Morris, G. B. Davis, and F. Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, Mar. 2003, https://doi.org/10.2307/30036540

[7] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Manage Sci*, vol. 46, no. 2, pp. 186–204, February 2000, https://doi.org/10.1287/mnsc.46.2.186.11926

[8] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Manage Sci*, vol. 35, no. 8, pp. 982–1003, August 1989, https://doi.org/10.1287/mnsc.35.8.982

[9] M. Blut, A. Y. L. Chong, Z. Tsigna, and V. Venkatesh, "Meta-Analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): Challenging its Validity and Charting a Research Agenda in the Red Ocean," *J Assoc Inf Syst*, vol. 23, no. 1, pp. 13–95, January 2022, https://doi.org/10.17705/1jais.00719

[10] J. Grundy, H. Khalajzadeh, and J. Mcintosh, "Towards Human-centric Model-driven Software Engineering," in *Proc. of the 15th International Conference on Evaluation of Novel Approaches to Software Engineering*, SCITEPRESS - Science and Technology Publications, 2020, http://doi.org/10.5220/0009806002290238

[11] United Nations Human Settlements Programme, "International guidelines on people-centered smart cities: Resolution HSP/HA.2/Res. 1," 2023. https://unhabitat.org/resolution-21-international-guidelines-on-people-centred-smart-cities-hspha2res1 (accessed May 21, 2025).

[12] General Assembly of the United Nations, "Global indicator framework for the Sustainable Development Goals and targets of the 2023 Agenda for Sustainable Development. A/RES/71/313," 2017. https://docs.un.org/A/RES/71/313 (accessed May 21, 2025)

[13] M. Moore, *Creating Public Value – Strategic Management in Government*. Cambridge: Harvard University Press, 1995.

[14] F. Beştepe and S. Ö. Yildirim, "Acceptance of IoT-based and sustainability-oriented smart city services: A mixed methods study," *Sustain Cities Soc*, vol. 80, article 103794, May 2022, https://doi.org/10.1016/j.scs.2022.103794

[15] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," *MIS Quarterly*, vol. 36, no. 1, p. 157 – 178, March 2012, https://doi.org/10.2307/41410412

[16] R. K. Jena, "Exploring Antecedents of Peoples' Intentions to Use Smart Services in a Smart City Environment: An Extended UTAUT Model," *Journal of Information Systems*, vol. 36, no. 1, pp. 133–149, March 2022, https://doi.org/10.2308/ISYS-2020-050

[17] A. Habib, D. Alsmadi, and V. R. Prybutok, "Factors that determine residents' acceptance of smart city technologies," *Behaviour & Information Technology*, vol. 39, no. 6, pp. 610–623, June 2020, https://doi.org/10.1080/0144929X.2019.1693629

[18] S. Lin, X. Liang, B. Zhang, F. Xing, and G. Peng, "Investigating Users Attitudes and Perceptions Towards the Usage of Smart City Apps," in *Distributed, Ambient and Pervasive Interactions. HCII 2020.* N. Streitz and S. Konomi, Eds. *Lecture Notes in Computer* Science, vol. 12203. Cham: Springer, 2020, pp. 415–430, https://doi.org/10.1007/978-3-030-50344-4_30

[19] T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in *Proc. of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, New York, NY, USA: ACM, June 2011, https://doi.org/10.1145/2037556.2037602

[20] T. Dirsehan and L. van Zoonen, "Smart city technologies from the perspective of technology acceptance," *IET Smart Cities*, vol. 4, no. 3, pp. 197–210, September 2022, https://doi.org/10.1049/smc2.12040

[21] J. Choi, "Enablers and inhibitors of smart city service adoption: A dual factor approach based on the technology acceptance model," *Telematics and Informatics*, vol. 75, article 101911, December 2022, https://doi.org/10.1016/j.tele.2022.101911

[22] T. Wang, C.-H. Jung, M.-H. Kang, and Y.-S. Chung, "Exploring determinants of adoption intentions towards Enterprise 2.0 applications: an empirical study," *Behaviour & Information Technology*, vol. 33, no. 10, pp. 1048–1064, October 2014, https://doi.org/10.1080/0144929X.2013.781221

[23] J. Ye, J. Zheng, and F. Yi, "A study on users' willingness to accept mobility as a service based on UTAUT model," *Technol Forecast Soc Change*, vol. 157, article 120066, August 2020, https://doi.org/10.1016/j.techfore.2020.120066

[24] T. B. Jørgensen and B. Bozeman, "Public Values: An Inventory," *Adm Soc*, vol. 39, no. 3, pp. 354–381, May 2007, https://doi.org/10.1177/0095399707300703

[25] M. Moore, "Creating Public Value: The Core Idea of Strategic Management in Government," *International Journal of Professional Business Review*, vol. 6, no. 1, article e219, February 2021, https://doi.org/10.26668/businessreview/2021.v6i1.219

[26] B. Bozeman, *Public Values and Public Interest: Counterbalancing Economic Individualism*. Georgetown University Press, 2007.

[27] T. Meynhardt, "Public value - Value creation in the eyes of society," in *Public Value: Deepening, Enriching, and Broadening the Theory and Practice*, A. Lindgreen, N. Koenig-Lewis, M. Kitchener, J. Brewer, M. Moore, and T. Meynhardt, Eds., London: Routledge, 2019, pp. 5–39.

[28] N. Faulkner and S. Kaufman, "Avoiding Theoretical Stagnation: A Systematic Review and Framework for Measuring Public Value," *Australian Journal of Public Administration*, vol. 77, no. 1, pp. 69–86, March 2018, https://doi.org/10.1111/1467-8500.12251

[29] K. Karunasena and H. Deng, "Testing and Validating a Conceptual Framework for Evaluating the Public Value of e-Government using Structural Equation Modelling," in *ACIS 2010 Proceedings*, 2010, article 13.

[30] G. Walsham, "Are We Making a Better World with Icts? Reflections on a Future Agenda for the IS Field," *Journal of Information Technology*, vol. 27, no. 2, pp. 87–93, 2012, https://doi.org/10.1057/jit.2012.4

[31] A. Lorenz, N. Madeja, and C. Leyh, "The Influence of Citizen Involvement on the Intention to Use Smart Traffic Management Applications," in *ICIS 2023 Proceedings*, 2023.

[32] A. Lorenz, N. Madeja, and C. Leyh, "Hearing the People: Citizen-Centric Requirements Engineering for the Development of Smart Traffic Management Applications," in *AMCIS 2024 Proceedings*, 2024.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 739–745
DOI: 10.15439/2025F1909
ISSN 2300-5963 ACSIS, Vol. 43

Towards Game Level Generation Through LLM and GAN

Filip Martinović, Danijel Mlinarić, Juraj Dončević, Agneza Krajna, Ivica Botički
0009-0000-1347-4377, 0000-0002-5248-7223, 0000-0001-5221-6848, 0000-0001-8304-5463, 0000-0002-7378-3339
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{filip.martinovic, danijel.mlinaric, juraj.doncevic, agneza.krajna, ivica.boticki}@fer.hr

***Abstract*—This paper tackles the challenge of adaptive level generation in video games, focusing on generating content that aligns with player skill. A key limitation of procedural content generation (PCG) is achieving semantic control. Specifically, generating levels of varying difficulty with limited training data. To address this problem, we propose a hybrid approach combining Large Language Models (LLMs) and Generative Adversarial Networks (GANs). An LLM is used to generate a diverse, difficulty-labeled dataset of Snake game levels, which are validated with A* pathfinding to ensure playability. These levels serve as training data for GANs that are able to efficiently generate new levels. The system is evaluated through user study and playability metrics. Results show that the LLM-assigned difficulty labels correlate strongly with human perception. The achieved playability is 87% for easy levels and 36% for hard levels. Our findings demonstrate that the hybrid LLM-GAN approach enables scalable and semantically controlled content generation, balancing quality, adaptability, and computational efficiency.**

***Index Terms*—procedural content generation, generative adversarial networks, large language models, video games, playability, adaptive systems.**

I. Introduction

THE greatest advances in the software industry today are due to the ability to adapt to the personal needs and preferences of individual users. Functionalities such as Netflix's recommendation system or Duolingo's customizable lessons have given companies an edge over the competition. In recent years, new large language models (LLMs) have also been developed to answer users' questions. Similar principles are being explored in the development of video games, particularly through the automatic generation of content based on players' skills and behavior [1].

A key challenge in adaptive content generation is semantic control, i.e. the ability to create content with specific, interpretable characteristics (such as difficulty level) that align with human perception and intent, rather than producing statistically similar outputs. Traditional procedural content generation methods face challenges. They either rely on manually created rules that limit creativity or require extensive datasets that may not exist for specific domains. And even when successful, they often lack the semantic understanding to produce content with certain meaningful characteristics.

This paper continues research into adaptive software systems [2,3], extending from dynamic software updating to adaptive content generation. Using the game Snake as a controlled environment, we investigated three approaches for AI-driven level generation: standalone GANs, standalone LLMs and a novel hybrid LLM-GAN approach.

Our research specifically examines whether modern AI approaches can provide reliable semantic control for difficulty adaptation. Initial experiments with GANs trained on limited data (15 levels) demonstrated severe overfitting and poor generalization. Pure LLM generation showed excellent quality control through semantic understanding but proved computationally expensive for real-time applications. This motivated our hybrid approach: using LLMs to bootstrap diverse training datasets for GAN training, combining the semantic precision of language models with the computational efficiency of adversarial networks. We validated the semantic control capabilities through user study, demonstrating a strong correlation between LLM-assigned difficulty labels and human perception. The focus is to determine whether the proposed hybrid pipeline can achieve the benefits of both approaches while mitigating their individual limitations.

II. Related Work

The field of procedural content generation (PCG) for video games [4] has made significant progress through the application of artificial intelligence techniques. This section presents the main approaches to game level generation: traditional PCG methods and their scalability limitations, generative model applications, and playability assurance techniques that ensure generated content meets functional requirements.

A. Procedural Content Generation

Togelius et al. [5] define procedural content generation (PCG) as the algorithmic creation of game content with limited or indirect user input. Some games such as Minecraft[1] use

This work was co-funded by the European Union through the National Recovery and Resilience Plan (NPOO), under the project *AutoEvolve – Automated Evolution of Software*, project code: NPOO.C3.2.R3-I1.05.0073.

[1] https://www.minecraft.net/

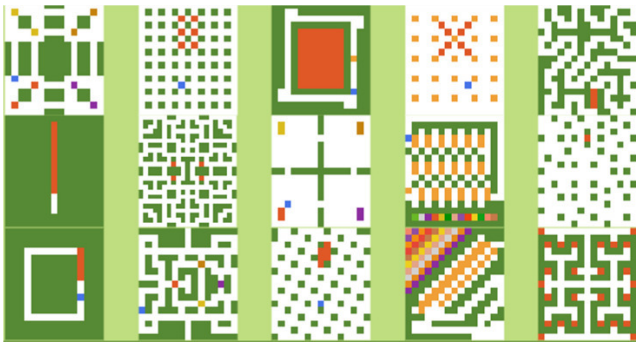


Fig. 1 Pre made levels for the Snake game. Green tiles represent walls, blue tiles represent the player's initial position, orange tiles represent box-es that can be moved by the player and all other colors of the tiles represent different types of food

PCG to create seemingly infinite worlds. Games in the roguelike and dungeon crawler genres, such as Spelunky[2], use it to create unpredictable levels that ensure each playthrough is unique. The challenge that arises from this approach is to manage the generation process to create meaningful and entertaining levels. This also requires substantial effort from the developer, who must set detailed restrictions to ensure playability.

B. Generative Models

Generative models are a class of machine learning algorithms designed to generate new data samples that resemble a given dataset. These models learn the underlying distribution of the data and can generate realistic outputs such as images, text, etc.

GANs [6] use the adversarial training process where the generator produces synthetic samples while the discriminator tries to distinguish real data from the generated ones. During the training process, the generator and discriminator play a mini-max game in which the goal for the generator is to produce realistic samples. Transformers [7], on the other hand, have introduced the mechanism of self-attention, which allows them to efficiently capture long-range dependencies. Although they are mainly used for text generation, they have also proven successful in image generation [8].

C. Generative Models for Level Creation

Recent research has investigated the use of machine learning (ML) for procedural content generation [9]. Generative Adversarial Networks (GANs) [6] have emerged as powerful tools for creating game levels that resemble human-designed content. As Volz et al. [10] have shown, GANs can capture the essence of existing Super Mario Bros levels while introducing new meaningful variations. Their approach uses a Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [11] to optimize the latent space of the network, allowing the creation of levels with certain desired properties, such as the wall density or the number of enemies. However, this method relies heavily on large datasets of existing levels and is constrained by the design patterns present in the original corpus, potentially limiting novelty.

The CESAGAN (Conditionally Embedded Self-Attention GAN) framework [12] builds on basic GAN architectures and incorporates explicit constraints into the generation process. This model encodes specific requirements, such as the number of walls or enemies, into feature vectors that guide the generation process. Bootstrapping techniques were used to solve the problem of a lack of data in the game-level datasets. This involves iteratively adding newly generated levels that meet the playability criteria, such as the correct number of tile types and the existence of a path to all goals, to the training dataset. This approach has significantly increased the percentage of playable levels while reducing the number of duplicates.

D. Playability

One of the main problems encountered with PCG is ensuring that created levels can be completed by the players. There are several approaches to solving this problem from different perspectives. One innovative approach combines classification with explainability techniques [13]. After identifying unplayable levels, models such as SHAP (SHapley Additive exPlanations) [14] or Integrated Gradients [15] assign importance values to each coordinate in the level, highlighting areas that are responsible for the unplayability. These values lead to targeted changes to problematic areas, while the overall structure and aesthetics of the level are preserved.

Zhang et al. [16] used GAN to generate Zelda levels and then used Mixed Integer Linear Programming (MILP) to repair unplayable levels. MILP provides a mathematical approach to level repair by representing levels as graphs in which the nodes correspond to tiles (walls, empty spaces, players, keys, doors or enemies) and the edges represent the connections between tiles. The playability problem is formulated as a network flow problem with sources (e.g. players), targets (e.g. doors, keys) and obstacles (e.g. walls). The solution minimizes the number of changes required to make a level playable while satisfying constraints related to flow preservation and obstacle avoidance. While this approach guarantees playability, the complexity of implementation could increase significantly for games with dynamic elements such as moving opponents.

Cooper and Sarkar [17] have used specialized repair agents to identify and repair unplayable areas. These agents can perform otherwise forbidden actions (e.g., walking through walls), but at a high cost. By identifying expensive actions on an agent's path to the goal, the system can make targeted modifications to the level, such as converting walls into empty spaces.

Jain et al. [18] used autoencoders [19] to generate new levels based on an existing corpus of Super Mario Bros levels. Their model compresses the level designs into a lower-dimensional latent space and then reconstructs them, effectively learning the basic patterns and structures that compose valid

[2] https://www.spelunkyworld.com/

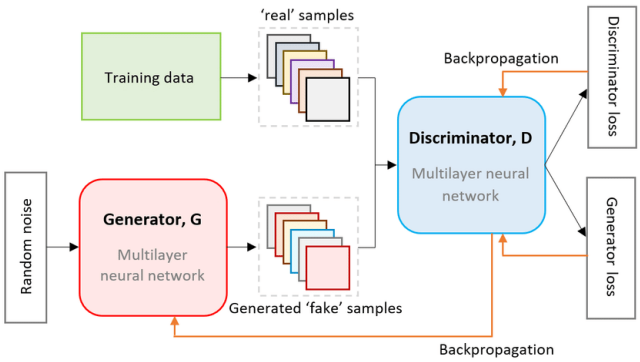


Fig. 2 Architecture of GAN [24]

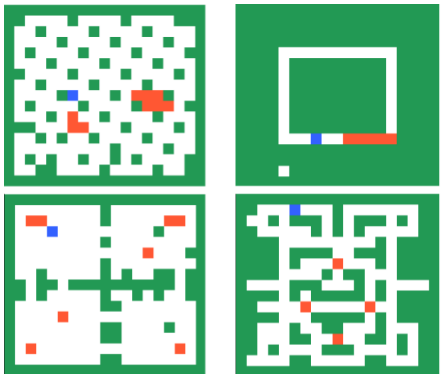


Fig. 3 Levels generated by GAN on initial dataset

levels. They have also used autoencoders to repair unplayable levels by running them through the model, which leverages its learned representations to reconstruct missing or problematic elements while maintaining overall coherence.

III. Hybrid Approach to Procedural Content Generation

This section describes our experimental framework for developing an adaptive Snake game that dynamically generates levels tailored to player skill. Our approach addresses a fundamental challenge in procedural content generation: balancing quality, controllability, and computational efficiency. While LLMs excel at generating semantically meaningful content with explicit difficulty control, they are computationally expensive for real-time applications. Conversely, once trained, GANs can efficiently generate content, but typically lack semantic understanding and require large datasets. Our hybrid pipeline combines these approaches by using LLMs to generate diverse, difficulty-labelled training data, then training specialized GAN models on this data for efficient generation at runtime. The system operates in two distinct phases. In the training phase, we use structured prompts to generate levels of controllable difficulty, validate their playability, and train separate GAN models for easy and hard difficulty levels.

A. Initial Dataset

One of the biggest challenges in training generative models is the lack of high-quality, diverse training data. The initial dataset was extracted from the 15 pre-made levels of the Google Snake Game Level Editor mod [20] (in Fig. 1).

However, this initial dataset of 15 levels presents a significant limitation for training GAN, which typically requires hundreds or thousands of training samples to learn meaningful data distributions and avoid overfitting. To address data scarcity, we used LLM to generate additional training data, as detailed in the section IV.A.

The pre-made levels consist of different types of tiles such as walls, crates, players and different types of food. To simplify the generation process, we have focused on 4 types of tiles that are essential for the game: walls, empty spaces, food and the player's starting position. All the different food types have been combined into one and all other tile types have been treated as empty spaces.

Individual pre-made level images are available on GitHub [21]. The images were processed by extracting tiles based on their colour. The levels can be displayed in different ways. Inspired by the game levels in The Video Game Levels Corpus (VGLC) [22], we opted for the 2D grid representation. A specific character symbol is used for each tile type. In our simplified levels, "W" stands for walls, "-" for empty areas, "F" for food and "S" for snake head, as shown in Fig. 4.

To encode the levels for training with GANs, each tile type is encoded in a separate channel, so that the levels were represented using four channels. Since the pre-made levels have different sizes, further preprocessing was required. In particular, all levels were padded with empty spaces up to a size of 32x32 tiles.

B. GAN Model

The architecture of the GAN model is based on the model from Arjovsky et al. [23]. In this design, the discriminator consists of blocks of strided convolutional layers followed by batch normalization layers and LeakyReLU activation layers. The generator uses transposed convolutional layers followed by batch normalization and ReLU layers. The last transposed convolutional layer is followed by the hyperbolic tangent function (*tanh*), which maps the outputs to the range [-1,1]. The architecture is adapted to 32x32 size images by removing the deepest layer of both the discriminator and the generator. The architecture of the entire system is shown in Fig. 2.

```
WWWWWWWWWWWWWWWWWWWW
W------WWWWW------W
W-A---S---------A-W
W--WW--WWWWW--WW--W
W--WW--WWWWW--WW--W
W----A-------A----W
W------WWWWW------W
WW-WW-WWWWWWW-WW-WW
WW-WW-WWWWWWW-WW-WW
WW-WW-WWWWWWW-WW-WW
W------WWWWW------W
W----A-------A----W
W--WW--WWWWW--WW--W
W--WW--WWWWW--WW--W
W-A-------------A-W
W------WWWWW------W
WWWWWWWWWWWWWWWWWWWW
```

Fig. 4 Example of pre-made level in 2D format

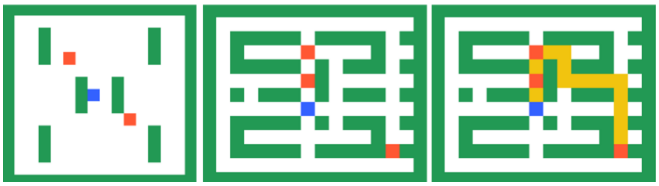


Fig. 5 a) Easy level generated by LLM, b) hard level generated by LLM and c) hard level generated by LLM with shortest path found with A* algorithm

After the training process, the generator can accurately replicate levels in the training dataset, as shown in Fig. 3. However, since only a limited amount of training data is available, the model tends to overfit. It replicates training samples instead of generating varied new levels. Furthermore, to generate levels with the desired difficulty, a larger, balanced dataset with difficulty labels is required.

C. LLM for Training Set

Instead of designing new levels manually, large language models (LLMs) can be used for this task. To generate new levels, the LLM must first be given clear instructions on how to generate them. It must also be given some levels as a reference. In most cases, text prompts are used as input to the model and the output is usually in the form of text. When generating levels with LLMs, we explored two primary output formats: 2D textual grids and structured representation. As level examples, we passed premade levels in 2D format, as shown in Fig. 4. The prompt to the model was as follows: "I have attached game levels for the game Snake. 'W' represents walls and '-' represents empty spaces. Generate more similar levels".

We found that the text outputs, especially for larger levels, often contained errors, such as an incorrect number of rows or an incorrect number of elements in a row. While most of these errors could be corrected manually, our goal is to automate the process. To generate high-quality, structured game layers, we combined GPT-4o and the OpenAI API Structured Outputs feature [25] that allowed us to capture the output in JSON format. To shape the whole process as a pipeline, we used the LangChain [26] framework in combination with a Pydantic-based output parser. Instead of generating free-form grids, we used pydantic.BaseModel to define a detailed schema to explicitly describe the components of a game level. The schema contains fields with the grid size (height and width of the level), a list of wall coordinates, the position of the snake head and a list of food coordinates. The output parser allowed us to validate each generated level and automatically discard levels that did not exactly match the schema. A typical prompt asked the model to generate a solvable Snake level of a given size with specific design constraints (the food must not be placed on the walls, the level must be solvable). The output was automatically parsed and validated. Inconsistencies or missing fields triggered errors, allowing us to discard or regenerate the sample in question. This approach provided a clean, normalized dataset suitable for conversion to structured tensors for GAN training. It also gave us the opportunity to generate new levels under certain conditions. The difficulty of the levels is particularly interesting. With carefully designed prompts, it is possible to generate easier or harder levels, as can be seen in Fig. 5. Easy levels have fewer walls, and food is generated closer to snake's head, while harder levels have more complex wall structures and food is harder to reach. In some cases, however, there is no clear difference in the difficulty of the generated levels.

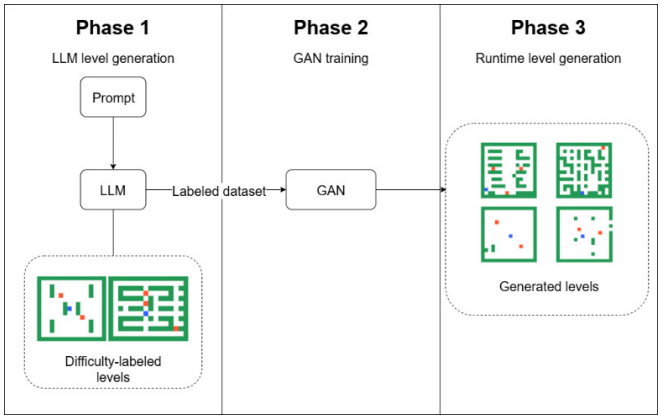


Fig. 6 System architecture of the hybrid LLM-GAN level generation pipeline

Even though there are multiple methods that can check the playability and even repair levels (section II.C), we only calculated the shortest path from the snake's head to all the food items. It can be calculated using the A* algorithm [27], adapted for multiple goals, where additional movement constraints apply: the snake cannot immediately turn back (return to the direction it just came from). The snake's ability to grow incrementally was not considered and it was treated as a single tile. The Manhattan distance was used as a heuristic function. Knowing the shortest path allows comparison of levels from the unbiased perspective of a "perfect player". We track metrics such as path length, the number of turns, the number of tight turns (around walls), and the longest straight path. We also track metrics related to the level itself, such as the percentage of walls, the number of dead ends, the number of "chokepoints" (narrow passages with walls on both sides). Regardless of these metrics, the difficulty depends significantly on the placement of the food.

IV. Experiment Setup

To create a substantially larger dataset for GAN training and validate our hybrid approach, we employed a comprehensive three-phase experimental framework. This methodology enables systematic comparison of the cost and quality of pure LLM generation, the effectiveness of hybrid GAN training, and the validation of semantic control through a user study, providing quantitative analysis of the trade-offs between the different approaches. The system architecture is shown in Fig. 6.

A. Generating Dataset for GAN Training

We used levels generated by an LLM in text format, as described in section III.A, and input them to LLM as examples. When selecting the examples, focus was to ensure that there

TABLE I

MODEL COMPARISON BETWEEN LEVEL DIFFICULTY

Model *Difficulty*	**Max (Avg) Playability Rate (PR)**	**Computational cost**	**Generation time**	**Semantic control**
GAN *Hard*	36 (24*) %	2h training duration	< 1s	Limited
GAN *Easy*	87 (74*) %	2h training duration	< 1s	Limited
LLM *Hard*	66 (100**) %	~705 tokens per level	~30s	Excellent
LLM *Easy*	100 %	~412 tokens per level	~30s	Excellent

*Average of top 5 hyperparameter configurations
**After validation and regeneration of failed levels

is a clear difference between hard and easy levels, like more open spaces and food closer to the player for easier levels, and more narrow passages and food placed in more difficult places for harder levels.

We then generated 50 easy and 50 difficult levels with structured few-shot prompts. Each generation started with a prompt with 3569 tokens for easy levels and 8495 tokens for difficult levels. The model's responses varied in complexity, averaging 412 tokens for easy levels and 705 tokens for difficult levels. Due to the nature of generative models that not all levels were viable. Some were unplayable or logically broken and had to be discarded, resulting in a higher number of tokens spent. As this process immediately generated 50 playable easy levels, there were 17 unplayable hard levels that we discarded. The total number of tokens used to generate all 100 levels was 617131. We can see that the generation process, while successful, consumes a lot of resources and is not practical when a big set of levels is needed.

B. GAN Training

With the generated dataset, we trained two separate GAN models: one for easy and one for difficult levels. Each model was trained[3] with a corresponding subset of 50 levels.

All levels were pre-processed as described in III.A. The training data was normalized to the range [-1,1] to align with the *tanh* activation in the generator's output layer, as mentioned in section III.B.

We explored different values for the learning rate, the stack size, the dimension of the latent vector, and the number of times the discriminator is trained before training the generator during each epoch. The playability rate (PR) was used as the optimization target. It is defined as number of solvable levels per 100 generated levels. The models were trained for 2000 epochs.

C. User Study for Semantic Control Validation

To evaluate the perceived difficulty of the LLM generated game levels, we conducted a user study involving seven participants. Each participant was asked to rate 100 level

[3]Training was conducted on NVIDIA GeForce GTX 1050 for approximately 4 hours. We used Optuna framework [28] for hyperparameters and monitored convergence through playability rate (PR) metric.

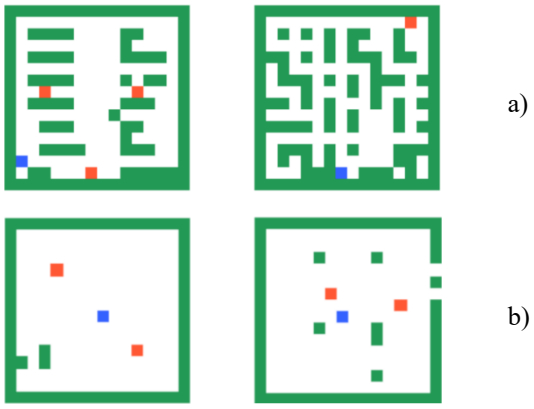


Fig. 7 Playable levels generated by GAN models trained on: a) "hard" dataset, b) "easy" dataset

images on a scale from 1 (very easy) to 5 (very difficult). The dataset consisted of 50 "easy" and 50 "hard" levels generated by a large language model in IV.A. However, users were not informed of these labels. The difficulty levels were randomly shuffled and presented to all users in the same order to ensure consistency. After collecting the individual ratings, we calculated the average rating for each level and analyzed its Pearson correlation with various structural features extracted from the levels described in III.C. This enabled us to identify which features most strongly influence the difficulty perceived by the users and the correspondence between the difficulty assigned as input to the LLM and the user ratings.

V. RESULTS

We present our findings in two main parts: first, validation of LLM semantic control through a user study, showing a strong correlation between assigned and perceived difficulty; second, the analysis of GAN performance when trained on LLM-generated datasets, revealing significant differences between easy and hard level generation complexity.

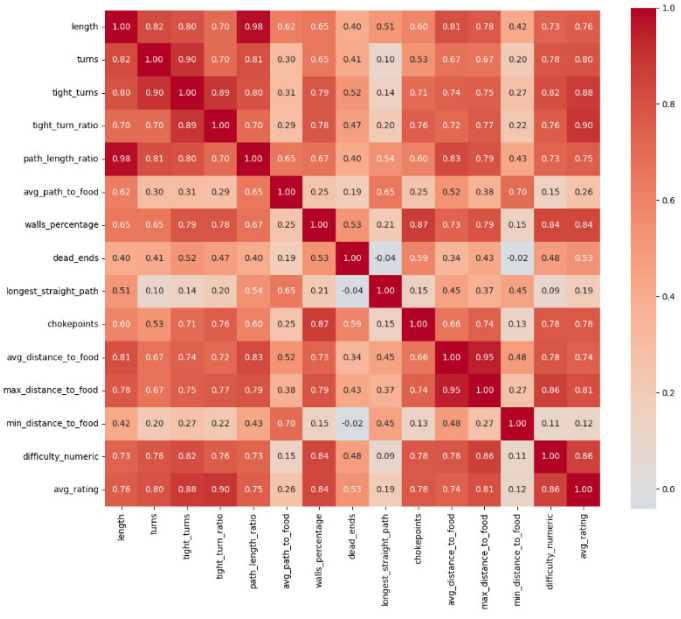


Fig. 8 Correlation matrix between structural level features and user difficulty ratings (n=7 participants, 100 levels)

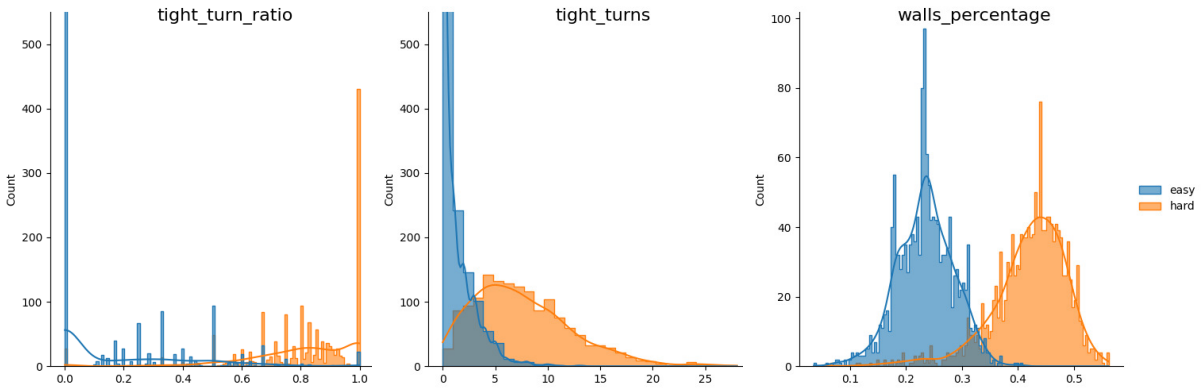


Fig. 9 Distribution comparison of key metrics between GAN-generated “easy” and “hard" levels

A. LLM Generated Levels

Based on the correlation (Fig. 8), we have identified several features that exhibit a strong relationship with user perceived difficulty. The most prominent is the *tight turn ratio*, which has the highest positive correlation with average user ratings, suggesting that levels requiring frequent sharp changes of direction are consistently rated as more difficult. Other features such as *total number of turns* and *percentage of walls* also correlate strongly with higher difficulty ratings, suggesting that spatial complexity and environmental constraints play a central role in perceived challenge. However, several features are highly intercorrelated, most notably *tight turns* and *turns*, as well as *average* and *maximum distance to food*, leading to redundancy and suggesting that a more compact feature set could be used without significant information loss. Interestingly, the binary *difficulty label* originally assigned to the level generation prompt ("easy" vs. "hard") shows a strong correlation with the actual user ratings.

Results show that the generated levels are highly consistent with players' perceptions, which strengthens confidence in the quality of the generated levels and can serve as a useful basis for further automation. However, a larger-scale study would be needed to confirm these findings across diverse player populations. It is important to note that our analysis is based on correlation rather than causation, further controlled experiments are needed to confirm which features directly influence perceived difficulty.

B. GAN Generated Levels

We have selected the five best performing hyperparameter configurations for each model (Easy GAN and Hard GAN) based on the playability rate (PR). For each of these configurations, we also identified the top three training checkpoints that produced the highest quality outputs. From each of these checkpoints, we generated 100 playable level samples, resulting in a diverse evaluation set. We then computed metrics described in IV.C for these generated levels. As shown in Table I, the model trained on the “hard” dataset generates much less playable levels than the one trained on the easy dataset. The significant difference in playability rates (87% vs 36%) suggests that difficult levels require more complex spatial reasoning that our current GAN architecture struggles to capture. This indicates that difficulty affects not only player perception but also computational generation complexity. Similar to the results in A, some features, such as number of *tight turns*, *tight turn ratio* and *percentage of walls*, show a clear difference between the generated easy and hard levels[4], as shown in Fig. 9. Models are able to generate completely new playable levels with distinct features as shown in Fig. 7.

VI. Conclusion

Our research explored different approaches to procedural content generation (PCG) of game level s. We found that, while efficient for generating a large number of levels, GANs require big and quality datasets. LLMs, on the other hand, produce high quality levels with very little input, but require a lot of resources to generate a larger amount of levels. We proposed a hybrid approach by combining the flexibility of large language models (LLMs) with the efficiency of generative adversarial networks (GANs). LLMs proved effective in generating structured and semantically rich level layouts, particularly when guided by schema-constrained prompting. However, due to high computational costs, LLMs are best suited for generating curated datasets. In contrast, GANs, once trained, can quickly generate large volumes of content but exhibit limited generalization when data is scarce. Our user study confirmed that LLM-labeled difficulty levels align well with player perceptions, and that certain structural features (e.g., tight turns, wall density) correlate strongly with perceived difficulty. These insights can guide future systems toward real-time, skill-adaptive content generation.

References

[1] G. N. Yannakakis and J. Togelius, ‘Experience-Driven Procedural Content Generation’, *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 147–161, Jul. 2011, doi: 10.1109/T-AFFC.2011.6.

[2] D. Mlinarić, ‘Extension of dynamic software update model for class hierarchy changes and run-time phenomena detection’, info:eu-repo/semantics/doctoralThesis, University of Zagreb. Faculty of Electrical Engineering and Computing. Department of Applied

[4]Complete results can be found at: https://fer-autoevolve.github.io/results/

Computing, 2020. Accessed: May 26, 2025. [Online]. Available: https://urn.nsk.hr/urn:nbn:hr:168:087042

[3] D. Mlinarić, J. Dončević, M. Brčić, and I. Botički, 'Revolutionizing Software Development: Autonomous Software Evolution', in 2024 47th MIPRO ICT and Electronics Convention (MIPRO), May 2024, pp. 224–228. doi: 10.1109/MIPRO60963.2024.10569871.

[4] N. Shaker, J. Togelius, and M. J. Nelson, Procedural Content Generation in Games. in Computational Synthesis and Creative Systems. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-42716-4.

[5] J. Togelius, E. Kastbjerg, D. Schedl, and G. N. Yannakakis, 'What is procedural content generation? Mario on the borderline', in *Proceedings of the 2nd International Workshop on Procedural Content Generation in Games*, in PCGames '11. New York, NY, USA: Association for Computing Machinery, Jun. 2011, pp. 1–6. doi: 10.1145/2000919.2000922.

[6] I. J. Goodfellow *et al.*, 'Generative Adversarial Networks', Jun. 10, 2014, *arXiv*: arXiv:1406.2661. doi: 10.48550/arXiv.1406.2661.

[7] A. Vaswani et al., 'Attention Is All You Need', Jun. 12, 2017, arXiv: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[8] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', Oct. 22, 2020, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[9] A. Summerville *et al.*, 'Procedural Content Generation via Machine Learning (PCGML)', *IEEE Trans. Games*, vol. 10, no. 3, pp. 257–270, Sep. 2018, doi: 10.1109/TG.2018.2846639.

[10] V. Volz, J. Schrum, J. Liu, S. M. Lucas, A. Smith, and S. Risi, 'Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network', May 02, 2018, *arXiv*: arXiv:1805.00728. doi: 10.48550/arXiv.1805.00728.

[11] N. Hansen, S. D. Müller, and P. Koumoutsakos, 'Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES)', *Evol. Comput.*, vol. 11, no. 1, pp. 1–18, Mar. 2003, doi: 10.1162/106365603321828970.

[12] R. Rodriguez Torrado, A. Khalifa, M. Cerny Green, N. Justesen, S. Risi, and J. Togelius, 'Bootstrapping Conditional GANs for Video Game Level Generation', in *2020 IEEE Conference on Games (CoG)*, Aug. 2020, pp. 41–48. doi: 10.1109/CoG47356.2020.9231576.

[13] M. Bazzaz and S. Cooper, 'Guided Game Level Repair via Explainable AI', Nov. 04, 2024, *arXiv*: arXiv:2410.23101. doi: 10.48550/arXiv.2410.23101.

[14] S. M. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 02, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[15] N. Kokhlikyan *et al.*, 'Captum: A unified and generic model interpretability library for PyTorch', Sep. 16, 2020, *arXiv*: arXiv:2009.07896. doi: 10.48550/arXiv.2009.07896.

[16] H. Zhang, M. C. Fontaine, A. K. Hoover, J. Togelius, B. Dilkina, and S. Nikolaidis, 'Video Game Level Repair via Mixed Integer Linear Programming', Oct. 13, 2020, *arXiv*: arXiv:2010.06627. doi: 10.48550/arXiv.2010.06627.

[17] S. Cooper and A. Sarkar, 'Pathfinding Agents for Platformer Level Repair'.

[18] R. Jain, A. Isaksen, and C. Holmg, 'Autoencoders for Level Generation, Repair, and Recognition'.

[19] G. E. Hinton and R. R. Salakhutdinov, 'Reducing the Dimensionality of Data with Neural Networks', *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[20] 'Google Snake Mods'. Accessed: Apr. 04, 2025. [Online]. Available: https://googlesnakemods.com/v/4/

[21] 'DarkSnakeGang/GoogleSnakeLevelEditor: Level Editor Mod for Google Snake'. Accessed: Apr. 08, 2025. [Online]. Available: https://github.com/DarkSnakeGang/GoogleSnakeLevelEditor

[22] A. J. Summerville, S. Snodgrass, M. Mateas, and S. Ontañón, 'The VGLC: The Video Game Level Corpus', Jul. 03, 2016, *arXiv*: arXiv:1606.07487. doi: 10.48550/arXiv.1606.07487.

[23] M. Arjovsky, S. Chintala, and L. Bottou, 'Wasserstein GAN', Dec. 06, 2017, *arXiv*: arXiv:1701.07875. doi: 10.48550/arXiv.1701.07875.

[24] C. Little, M. Elliot, R. Allmendinger, and S. S. Samani, 'Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study', Dec. 03, 2021, arXiv: arXiv:2112.01925. doi: 10.48550/arXiv.2112.01925.

[25] 'Structured Outputs - OpenAI API'. Accessed: May 24, 2025. [Online]. Available: https://platform.openai.com

[26] 'LangChain'. Accessed: May 22, 2025. [Online]. Available: https://www.langchain.com/langchain

[27] P. E. Hart, N. J. Nilsson, and B. Raphael, 'A Formal Basis for the Heuristic Determination of Minimum Cost Paths', *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 2, pp. 100–107, Jul. 1968, doi: 10.1109/TSSC.1968.300136.

[28] 'Optuna - A hyperparameter optimization framework', Optuna. Accessed: May 23, 2025. [Online]. Available: https://optuna.org/

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 747–752
DOI: 10.15439/2025F6805
ISSN 2300-5963 ACSIS, Vol. 43

DivCrypt: A Structured Framework for Validating Cryptographic Implementations

Artur Misztal
0000-0003-2402-937X
AGH University of Krakow, al. Mickiewicza 30, 30-059 Krakow, Poland
Łukasiewicz - AI, ul. Leopolda 31, 40-189 Katowice, Poland
Email: misztal@agh.edu.pl, artur.misztal@ai.lukasiewicz.gov.pl

***Abstract*—This paper introduces the DivCrypt framework, a step-by-step guidance for validation of cryptographic implementations. The core idea is to decompose the evaluated implementation into components across complexity layers and validate each component through a structured five-step process. DivCrypt is not intended to replace existing standards, but rather to serve as a practical audit playbook. It can be easily adopted by evaluators in formal certification processes, researchers conducting security audits, and developers performing internal testing. The framework is also intended to encourage creators of novel, non-standardized cryptographic constructs to publish and maintain DivCrypt-aligned knowledge base. Such contributions, including test vectors and known implementation vulnerabilities, would support not only the application of DivCrypt, but also benefit the broader research and development community.**

I. Introduction

MODERN services increasingly depend on complex, multi-component IT/OT systems, highlighting the critical need to enhance their robustness. To establish trust, the security of these systems should be validated through external, independent security audits, typically performed in accordance with family of standards such as ISO/IEC 15408 [1] or ISA/IEC 62443 [2]. Compliance with such norms may be enforced by law or required in procurement for the defense sector. Given their fundamental role, cryptographic mechanisms are integral to these certification processes, serving as essential trust anchors. Even minor implementation flaws can significantly undermine security of the system, emphasizing the need for consistent and rigorous validation. Consequently, there are standards for both developing [3], [4] and testing [5], [6] of the cryptographic modules.

However, existing norms do not cover the validation of non-standardized cryptographic mechanisms. Notably, widely adopted community or industry tested cryptographic solutions sometimes enable advanced features in modern applications, yet the formal standardization processes can take years[1] resulting in significant gaps in official guidance. Security auditors evaluating novel cryptographic schemes currently rely on limited informal guidelines [9]. As a result, a comprehensive, high-level framework for the consistent validation of cutting-edge cryptography implementations is currently lacking.

[1]For example, the NIST Post Quantum Cryptography competition selected new standards: ML-KEM [7] and ML-DSA [8], after eight years of evaluation.

To address this gap, we introduce a methodology consisting of five sequential steps, grounded in a core principle: decomposing cryptographic solutions into smaller, manageable components. As illustrated in Figure 1, each component undergoes comprehensive security evaluation across four dimensions: *theoretical*, *functional*, *computational*, and *contextual*. The advantages of our framework include:

- **Efficiency.** Defines high-level procedural steps and methodologies without prescribing specific implementation details. The decomposition of the evaluated implementation and reuse of knowledge bases make the framework efficient by avoiding redundant efforts;
- **Inclusivity.** Accessible to a diverse range of users including developers, security auditors, and evaluators - thanks to its simplicity and flexibility in tool selection;
- **Universality.** Applicable to both standardized and emerging cryptographic protocols, including homomorphic encryption and non-interactive zero-knowledge proofs;
- **Estimability.** Once an implementation is decomposed, the total time required for evaluation becomes easy to estimate.

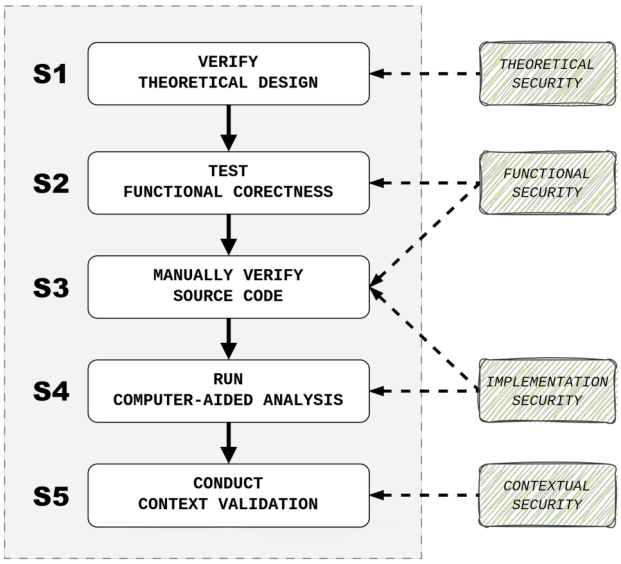


Fig. 1. The five sequential steps of DivCrypt comprehensively address four essential aspects of implementation security.

The absence of standards for novel cryptographic solutions poses significant challenges, especially under regulatory scrutiny of blockchain-based applications. In this domain, there is a growing adoption of zero-knowledge proofs, primarily to enhance privacy and scalability. Our framework is intended to bridge the gap between innovation and compliance, enabling consistent evaluation of cryptographic implementations across industry.

II. DivCrypt Framework

A. Terminology

The DivCrypt framework adopts a structured perspective on cryptographic implementations by organizing them into a hierarchy of four abstraction levels, referred to as *Cryptographic Layers* (CL):

CL1. cryptographic protocols;
CL2. cryptographic schemes;
CL3. cryptographic primitives;
CL4. underlying arithmetic.

We also introduce the notion of *Cryptographic Layer Object* (CLO) which denotes a specific instance at any given layer. For example, CLO can be: TLS v1.3 (CL1), Ed25519 (CL2), SHA-512 (CL3) or prime field (CL4). As a natural extension of CL and CLO, we define the notion of a *Cryptographic Layer Object Implementation* (CLOI), representing the actual implementation of a given CLO. The implementation under evaluation within the DivCrypt framework is referred to as the DivCrypt Target (DCT). DCT is subsequently decomposed into a directed acyclic graph, known as the DCT-graph, where the DCT itself serves as the root node with no parents. An illustrative example of such a decomposition is presented in Figure 2.

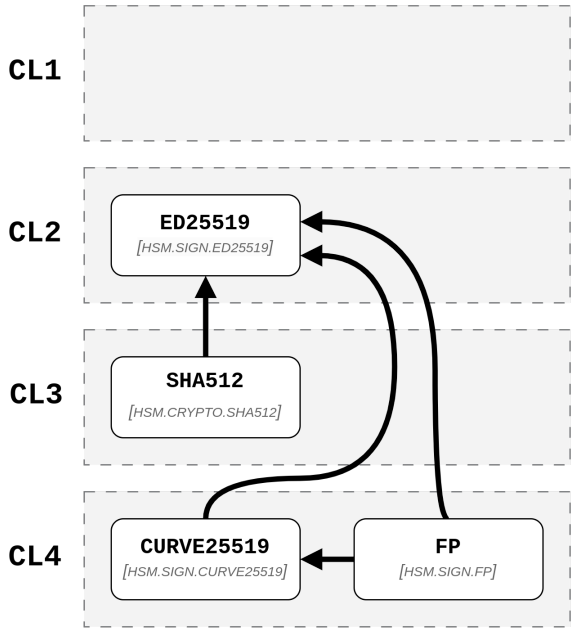


Fig. 2. The DCT-graph of the software implementation of Ed25519 digital signature scheme.

B. Testing procedure

The main idea behind DivCrypt is to conduct methodological analysis of cryptographic primitives, schemes, or protocols by applying divide-and-conquer-like approach, which materializes through the construction of the DCT-graph. The DivCrypt framework incorporates a range of validation techniques to comprehensively address all relevant security aspects of cryptographic algorithm implementations. It consists of three main parts:

1) ***Building the Knowledge Base****:* This is a continuous process of maintaining a database comprising two core elements described in the Section III: the *Cryptographic Validation Card* (CVC) and the *Cryptographic Vulnerability Reference* (CVR). This component is verification-independent and serves as a shared foundation for evaluations.

2) ***Decomposing into DCT-graph****:* By identifying each subcomponent according to its corresponding cryptographic layer, the DCT is decomposed into a directed acyclic graph with four levels of nodes, each visualizing a specific Cryptographic Layer (CL). The DCT-graph can be understood as a graph of dependencies, where edges represent implementation-level reliance. An example DCT-graph for Ed25519 [10] is shown in Figure 2. It is important to note that every node must be identified by its package name, as a DCT may contain multiple implementations of the same primitive (e.g., three different implementations of SHA-256). In Figure 2, package names are enclosed in square brackets. Each package name must be unique to ensure clarity and avoid ambiguity.

3) ***Validating DCT-graph****:* Validation is carried out through five sequential steps described in the Section IV, applied recursively from the leaf nodes to the root, ensuring that each node is verified before its parent. These steps, illustrated in Figure 1, define a strict sequence to follow, while allowing flexibility in how precisely each step is tailored to the context and assurance requirements of the evaluation.

III. Reusable Knowledge Base

During the decomposition of the DCT into the DCT-graph, it is necessary to ensure that every CLO is associated with a corresponding *Cryptographic Validation Card* (CVC) and *Cryptographic Vulnerability Reference* (CVR). The creation and maintenance of both artifacts are verification-independent tasks, performed prior to the validation phase. The purpose of this initial step is to establish and continuously update a knowledge base containing essential information about various CLOs. The CVC (Subsection III-A) captures the detailed specification and structure of CLO, incorporating both official and internally constructed test vectors. The CVR (Subsection III-B) enumerates known sources of security issues, including academic papers, vulnerability disclosures, and implementation-specific attack reports relevant to the examined CLO. The concept of utilizing a reusable knowledge

base to improve the efficiency and consistency of security evaluations has been applied in other contexts - for example, to streamline the complex documentation process for Common Criteria certification [11].

A. Cryptographic Validation Card

[CVC-1] Specifications. Reference to the official CLO specification, including relevant standards and technical documentation.

[CVC-2] Algorithms. A list of algorithms that collectively define the CLO.

[CVC-3] Dependencies. A list of other CLOs from lower cryptographic layers that are directly utilized by the CLO.

[CVC-4] Sources. Collection of useful materials, such as reference implementations and technical analyses.

[CVC-5] Official Test Procedures. References to official testing procedures developed by the authors of the cryptographic standard or proposal. Usually, this is a set of test vectors called Known Answer Tests (KATs).

[CVC-6] Other Test Procedures. Supplementary internal testing strategies aimed at verifying both correctness and robustness against failure scenarios. This can be achieved by designing test vectors that cover edge cases and applying them to both the reference and the evaluated implementation.

B. Cryptographic Vulnerability Reference

[CVR-1] Queries Table. A complete list of search queries used to collect entries for CVR-2, CVR-3, and CVR-4.

[CVR-2] CVE Entries. A list of relevant Common Vulnerabilities and Exposures (CVE) entries associated with the CLO.

[CVR-3] Academic Papers. A curated list of academic publications addressing implementation-specific aspects or vulnerabilities of the CLO.

[CVR-4] Media. A collection of publicly available sources, such as blog posts, technical write-ups, or incident reports, that refer to observed vulnerabilities or behaviors.

IV. DivCrypt's Evaluation Steps

S1. Verify theoretical design

Task. Evaluator shall verify that the DCT is listed in widely accepted standards.

Example. The evaluator confirmed that the Ed25519 signature scheme is included in NIST FIPS 186-5 [12].

Remarks. DivCrypt is not intended to analyze the security of the cryptographic protocols, schemes, or primitives, but only the implementation correctness of standardized ones.

S2. Test functional correctness

Task. Evaluator shall execute the test procedures specified in CVC-5 and CVC-6.

Example. The evaluator developed interfaces in Python for both the tested and reference implementations. A test script was run, invoking procedures described in CVC-5 and CVC-6. The results indicated all tests completed successfully.

Remarks. Testing with predefined or dynamically generated test vectors is the basic method of verifying implementation correctness. The reference implementation serves as a benchmark to validate the behavior of the evaluated cryptographic module.

S3. Source code verification

Task. Evaluator shall perform a comprehensive review of source files, focusing on critical functions and their usage, including input handling and side-channel resistance.

Example. The evaluator manually reviewed the codebase and, using all accessible CVC/CVR artifacts, concluded that the implementation conforms to current best practices and is appropriate for the intended application context.

Remarks. Although time-consuming, manual source code analysis provides deep insight and allows the evaluator to understand the implementation at a level comparable to its developers.

S4. Computer-aided analysis

Task. The evaluator shall apply static and/or dynamic analysis tools to identify implementation flaws.

Example. The evaluator used Wycheproof [13] and Cryptofuzz [14] to check the Ed25519 implementation. Then he applied a generic C++ static analyzer to detect unsafe coding practices.

Remarks. Static analysis does not require execution of the evaluated codebase and includes tools like Wycheproof [13]. Dynamic analysis is performed during execution and includes fuzzers and tools such as Valgrind [15] to detect runtime issues like memory leaks.

S5. Context validation

Task. The evaluator shall examine the interaction of the DCT with its external environment to identify context-related misuse or misconfiguration.

Example. The evaluator used the recursive `grep` command to list function calls and further analyzed the invoking context.

Remarks. Contextual validation ensures that even correctly implemented algorithms are not undermined by improper use. This step often goes beyond the scope of the DCT itself but is essential for assessing overall system security.

V. Case Study: ED25519 Implementation

To validate the implementation of the Ed25519 digital signature scheme [10], we constructed a knowledge base comprising a *Cryptographic Validation Card* and a *Cryptographic Vulnerability Reference* for each component shown in Figure 2. For brevity, Subsection V-A presents CVC of just one component - Ed25519 digital signature scheme. In the Subsection V-B we cover the concise record of the validation process. All future updates to the framework and its knowledge base will be made available in the project's GitHub repository at *https://github.com/arturmisztal/divcrypt*.

A. Cryptographic Validation Card for Ed25519

[CVC-1] Specifications.

- NIST FIPS 186-5 [12].
- RFC 8032 [16].
- EdDSA for more curves [17].
- High-speed high-security signatures [10].

[CVC-2] Algorithms.

- `ed25519-gen`
  - *description:* key generation
  - *input domain:* -
  - *output domain:* $\{0,1\}^{256} \times \{0,1\}^{256}$
- `ed25519-sig`
  - *description:* signature generation
  - *input domain:* message $\mathcal{M}$ of length less than 2^{128} bits
  - *output domain:* $\{0,1\}^{512}$
- `ed25519-ver`
  - *description:* signature verification
  - *input domain:* $\mathcal{M} \times \{0,1\}^{512} \times \{0,1\}^{256}$
  - *output domain:* $\{$`invalid, valid`$\}$

[CVC-3] Dependencies (Table I).

[CVC-4] Sources.

- NIST ACVP EdDSA [18].
- Ed25519 reference implementation [19].
- CRYPTREC Review of EdDSA [20].
- EdDSA signature verification edge cases [21].
- The Provable Security of Ed25519 [22].

[CVC-5] Official Test Procedures.[2]

- For the given set of messages and private keys, the signature is generated with `ed25519-sig` and the expected result is to be correctly verified by `ed25519-ver` [19].
- For the given set of messages and private keys, the signature is generated with `ed25519-sig`, incremented, and the expected result is that the forgery is detected by `ed25519-ver` [19].
- Ed25519 test vectors from RFC 8032 [16].
- EdDSA test procedures located in CCN-STIC 2100 [23].

[CVC-6] Other Test Procedures.[2]

- EdDSA-oriented adaptation of ECDSA test procedures located in SOG-IS Crypto Evaluation Scheme [6] and ISO/IEC 18367 [24].

[2]Although a standalone guide would describe the test procedures in details using pseudocode, this paper provides only a high-level overview due to space constraints.

TABLE I
LIST OF DIRECT `ED25519` DEPENDENCIES.

ALGORITHM	`ed25519-gen`	`ed25519-sig`	`ed25519-ver`
`curve25519-mul`	✓	✓	✓
`curve25519-add`	✗	✗	✓
`sha2-512`	✓	✓	✓
`fp-mul`	✗	✓	✗
`fp-add`	✗	✓	✗

B. Validation Process Record

1) The implementation under evaluation was identified as a portable C implementation of the Ed25519 digital signature scheme.
2) The DCT-graph for the Ed25519 implementation was constructed.
3) The implementation was investigated to ensure that each *Cryptographic Layer Object* (CLO) was uniquely represented. As the hash function was located in a separate file, two package families were created: [HSM.SIGN] and [HSM.CRYPTO] (see Figure 2).
4) The CVC and CVR artifacts were created for the following components:
   - prime field (`fp`),
   - elliptic curve Curve25519 (`curve25519`),
   - hash function SHA512 (`sha2-512`),
   - digital signature scheme Ed25519 (`ed25519`).

   In cases where CVC/CVR artifacts already existed, they only needed to be updated.
5) The five DIVCRYPT steps were performed sequentially for each *Cryptographic Layer Object Implementation* (CLOI). Test procedures (Step S2) were conducted using the Python programming language, while SageMath [25] was used to generate reference test vectors for `fp` and `curve25519`. Computer-aided analysis (Step S4) was performed using the general-purpose static code analysis tool CodeChecker [26].
6) As the evaluation of each CLOI was successful, the final verdict was PASS.

VI. COMPARISION WITH EXISTING STANDARDS

In the domain of systems security, standards can generally be classified into three categories: *theoretical components*, *evaluation criteria*, and *testing guidance*. When narrowing the focus specifically to cryptographic mechanisms, one can readily identify numerous specifications published by National Institute of Standards and Technology [27], whether these define a family of cryptographic solutions [12] or detail a single standardized scheme [7]. In addition, there exist global [3], regional [4] and national [28] standards that specify how and which schemes should be deployed. Unfortunately, when it comes to testing guidance, such documents are almost always tailored to specific schemes within particular application domains. Existing standards that describe testing methodologies for cryptographic implementations are summarized in Table II. According to our case study, we provide a brief overview of the procedures dedicated to the Ed25519 digital signature scheme recommended by each standard.

*a) **ISO/IEC 24759 [3]:*** This document defines the specific test procedures that a laboratory must use to check a module's conformance to the requirements in ISO/IEC 19790 [5]. For digital signatures, its procedures ensure a module can perform critical self-tests, such as the pair-wise consistency test upon key generation and various KATs.

TABLE II
A SELECTION OF ENGLISH-LANGUAGE STANDARDS FOR TESTING CRYPTOGRAPHIC IMPLEMENTATIONS.

YEAR	TITLE	SCOPE
2025	*ISO/IEC 24759: Test requirements for cryptographic modules* [5]	Global
2025	*CCN-STIC 2100: Cryptographic Mechanisms Evaluation Methodology* [23]	Spain
2023	*AEPD Guidelines for the validation of cryptographic systems in data protection processing* [29]	Spain
2022	*EN 17640: Fixed-time cybersecurity evaluation methodology for ICT products (Sections 6.11 and 6.12)* [30]	Europe
2020	*SOG-IS Crypto Evaluation Scheme: Harmonised Cryptographic Evaluation Procedures* [6]	Europe
2019	*FIPS 140-3: Security Requirements for Cryptographic Modules* [31]	USA
2016	*ISO/IEC 18367: Information technology - Security techniques - Cryptographic algorithms and security mechanisms conformance testing* [24]	Global
2015	*ANSSI-CC-CRY-P01: Methods For Carrying Out Cryptographic Analysis And Random Number Evaluations* [32]	France

b) ***CCN-STIC 2100 [23]****:* This document from Spain's National Cryptologic Center is a highly detailed and prescriptive methodology for evaluating specific, CCN-agreed cryptographic mechanisms against defined certification levels. It provides an exhaustive list of conformity tests for EdDSA (see Table 66 in [23]), including specific Validation Tests for signature generation and KATs for signature verification.

c) ***AEPD [29]****:* These guidelines from the Spanish Data Protection Agency provide a framework for assessing cryptographic systems, specifically to ensure they comply with GDPR [33] and effectively protect personal data. Its focus is on the governance and risk management surrounding the cryptosystem rather than detailing technical verification procedures for specific algorithms like EdDSA.

d) ***EN 17640 [30]****:* EN 17640 defines a modular evaluation methodology for ICT products designed to be performed within a fixed time budget, including tasks for basic and extended analysis of cryptographic mechanisms. Its recommendations cover conformance testing with KATs and source code analysis to check for implementation errors.

e) ***SOG-IS [6]****:* These harmonized procedures detail the evaluation tasks for cryptographic mechanisms within the formal SOG-IS scheme, which is used for Common Criteria evaluations in Europe. Crucially, this document defines a complete suite of evaluation tasks specifically for ECDSA-like digital signatures, including conformity testing and analysis of implementation pitfalls related to the underlying elliptic curve cryptography.

f) ***FIPS 140-3 [31]****:* This standard specifies the security requirements for a cryptographic module as a whole, covering areas from physical security to roles and services. It delegates the specific test procedures for approved algorithms like EdDSA to related standards.

g) ***ISO/IEC 18367 [24]****:* This standard provides guidelines for black-box and white-box conformance testing to ensure a cryptographic implementation correctly adheres to its specification. ISO/IEC 18367:2016 is based on conformance testing methods employed by the Japan Cryptographic Module Validation Program (JCMVP) and the NIST Cryptographic Algorithm Validation Program (CAVP).

h) ***ANSSI-CC-CRY-P01 [32]****:* This French procedure defines the formal roles and responsibilities for how approved evaluation labs must conduct cryptographic analyses for the national certification scheme, distinguishing between theoretical and implementation analysis. It specifies that an evaluator should perform analysis of conformity and vulnerability in cryptography implementation.

Many standards such as CCN-STIC 2100 [23], EN 17640 [30] or SOG-IS procedures [6] cover important topics related to cryptographic scheme sub-primitives, implementation pitfalls and usage context. However, a universal, formalized methodology that is equally applicable to both a simple stream cipher like ChaCha20 [34] and a complex non-interactive zero-knowledge protocol zk-STARK [35] is currently lacking, given their significant differences in cryptographic structure, complexity, and maturity. What these diverse cryptographic constructs do share is the need to address four fundamental aspects of implementation security, which DIVCRYPT covers through five dedicated actions (see Figure 1).

VII. CONCLUSION AND FUTURE WORK

This work introduced DIVCRYPT, a structured and layered framework for validating cryptographic implementations. By decomposing cryptographic solutions into layered components and evaluating each through structured steps, DIVCRYPT ensures that both standardized and emerging cryptographic mechanisms are subjected to thorough and context-aware analysis. The framework balances flexibility with rigor, offering evaluators a practical methodology that is both scalable and reusable across implementations. DIVCRYPT addresses a critical gap in current evaluation practices by introducing a reusable knowledge base (CVC and CVR), a clear abstraction hierarchy (CLO, CLOI and DCT-graph), and a comprehensive five-step validation process that covers theoretical design, functional correctness, code-level assurance, computer-aided analysis, and contextual robustness.

Future work should focus on several key directions. First, initiating a broader discussion with the community is essential to evaluate the advantages and limitations of the DIVCRYPT framework. Second, establishing a publicly accessible database of CVC and CVR entries would significantly support shared validation efforts and improve reproducibility. Finally, incorporating formal verification methods potentially as an optional step should be considered, as this would further strengthen the rigor and reliability of the validation process where applicable.

Acknowledgment

The author would like to thank Paweł Topa (AGH University) and Dariusz Rogowski (Łukasiewicz – AI) for critically reviewing the manuscript, as well as the anonymous reviewers for their constructive feedback.

References

[1] International Organization for Standardization, "ISO/IEC 15408: Information security, cybersecurity and privacy protection - Evaluation criteria for IT security," International Organization for Standardization, Tech. Rep., 2022. [Online]. Available: https://www.commoncriteriaportal.org

[2] International Society of Automation, "ISA/IEC 62443: Security for industrial automation and control systems," IEC Central Office, Tech. Rep., 2019. [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards

[3] International Organization for Standardization, "ISO/IEC 19790: Security Requirements for Cryptographic Modules," International Organization for Standardization, Tech. Rep. ISO/IEC 19790, 2025. [Online]. Available: https://www.iso.org/standard/82423.html

[4] SOG-IS Crypto Working Group, "Crypto Evaluation Scheme: Agreed Cryptographic Mechanisms," Senior Official Group Information Security Systems, Tech. Rep., 2023. [Online]. Available: https://www.sogis.eu/documents/cc/crypto/SOGIS-Agreed-Cryptographic-Mechanisms-1.3.pdf

[5] International Organization for Standardization, "ISO/IEC 24759: Test Requirements for Cryptographic Modules," International Organization for Standardization, Tech. Rep. ISO/IEC 24759, 2025. [Online]. Available: https://www.iso.org/standard/82424.html

[6] SOG-IS Crypto Working Group, "Crypto Evaluation Scheme: Harmonised Cryptographic Evaluation Procedures," Senior Official Group Information Security Systems, Tech. Rep., 2020. [Online]. Available: https://www.sogis.eu/documents/cc/crypto/202203-hep-draft16.pdf

[7] National Institute of Standards and Technology, "Module-Lattice-Based Key-Encapsulation Mechanism Standard," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST FIPS 203, Aug. 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.203.pdf

[8] ——, "Module-Lattice-Based Digital Signature Standard," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST FIPS 204, Aug. 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.204.pdf

[9] "ZKDocs." [Online]. Available: https://www.zkdocs.com/

[10] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B.-Y. Yang, "High-speed high-security signatures," *Journal of Cryptographic Engineering*, vol. 2, no. 2, pp. 77–89, Sep. 2012. [Online]. Available: http://link.springer.com/10.1007/s13389-012-0027-1

[11] D. Rogowski, "Software Implementation of Common Criteria Related Design Patterns," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. Kraków: IEEE, 2013, pp. pages 1135–1140. [Online]. Available: https://annals-csis.org/Volume_1/pliks/210.pdf

[12] L. Chen, D. Moody, A. Regenscheid, and A. Robinson, "NIST FIPS 186-5: Digital Signature Standard (DSS)," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, Tech. Rep. NIST FIPS 186-5, Feb. 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.186-5.pdf

[13] "Wycheproof," May 2025, original-date: 2016-11-08T20:56:25Z. [Online]. Available: https://github.com/C2SP/wycheproof

[14] "Cryptofuzz," Jun. 2025, original-date: 2024-11-25T19:31:28Z. [Online]. Available: https://github.com/MozillaSecurity/cryptofuzz

[15] "Valgrind." [Online]. Available: https://valgrind.org/

[16] S. Josefsson and I. Liusvaara, "Edwards-Curve Digital Signature Algorithm (EdDSA)," RFC Editor, Tech. Rep. RFC8032, Jan. 2017, issue: 8032 Num Pages: 60 Series: Request for Comments Published: RFC 8032. [Online]. Available: https://www.rfc-editor.org/info/rfc8032

[17] D. J. Bernstein, S. Josefsson, T. Lange, P. Schwabe, and B.-Y. Yang, "EdDSA for more curves," 2015, published: Cryptology ePrint Archive, Paper 2015/677. [Online]. Available: https://eprint.iacr.org/2015/677

[18] "ACVP EdDSA Algorithm JSON Specification." [Online]. Available: https://pages.nist.gov/ACVP/draft-celi-acvp-eddsa.html

[19] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B.-Y. Yang, "Ed25519 software." [Online]. Available: https://ed25519.cr.yp.to/software.html

[20] S. D. Galbraith, "CRYPTREC Review of EdDSA," Mathematics Department, University of Auckland, Auckland, New Zealand, Tech. Rep. CRYPTREC EX-3003-2020, 2020. [Online]. Available: https://www.cryptrec.go.jp/exreport/cryptrec-ex-3003-2020.pdf

[21] "novifinancial/ed25519-speccheck," Apr. 2025, original-date: 2020-07-28T21:44:25Z. [Online]. Available: https://github.com/novifinancial/ed25519-speccheck

[22] J. Brendel, C. Cremers, D. Jackson, and M. Zhao, "The Provable Security of Ed25519: Theory and Practice," in *2021 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, May 2021, pp. 1659–1676. [Online]. Available: https://ieeexplore.ieee.org/document/9519456/

[23] Centro Criptologico Nacional, "CCN-STIC 2100: Cryptographic Mechanisms Evaluation Methodology," 2025. [Online]. Available: https://oc.ccn.cni.es/en/types-of-certification/crytpologic-certification/criteria-and-methodologies

[24] International Organization for Standardization, "ISO/IEC 18367: Information technology - Security techniques - Cryptographic algorithms and security mechanisms conformance testing," International Organization for Standardization, Tech. Rep., 2016. [Online]. Available: https://www.iso.org/standard/62286.html

[25] "SageMath," Jun. 2025. [Online]. Available: https://github.com/sagemath/sage

[26] "CodeChecker," Jul. 2025. [Online]. Available: https://github.com/Ericsson/codechecker

[27] "National Institute of Standards and Technology," Apr. 2025, last Modified: 2025-06-04T08:37-04:00. [Online]. Available: https://www.nist.gov/

[28] Federal Office for Information Security, "BSI TR-02102-1: Cryptographic Mechanisms - Recommendations and Key Lengths," 2025. [Online]. Available: https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Technische-Richtlinien/TR-nach-Thema-sortiert/tr02102/tr02102_node.html

[29] Spanish Data Protection Agency, "Guidelines for the validation of cryptographic systems in data protection processing," 2023. [Online]. Available: https://www.aepd.es/guides/guidelines-validation-cryptographic-systems-data-protection-processing.pdf

[30] The European Committee for Standardization, "EN 17640: Fixed-time cybersecurity evaluation methodology for ICT products," CEN and CENELEC, Tech. Rep. EN 17640, 2022.

[31] National Institute of Standards and Technology, "NIST FIPS 140-3: Security Requirements for Cryptographic Modules," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST FIPS 140-3, Apr. 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.140-3.pdf

[32] French Cybersecurity Agency, "Methods For Carrying Out Cryptographic Analysis And Random Number Evaluations," 2015. [Online]. Available: https://cyber.gouv.fr/sites/default/files/2022-08/ANSSI-CC-CRY-P-01-Modalites-pour-la-realisation-des-analyses-crypto_v003_EN%5B3%5D.pdf

[33] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

[34] D. J. Bernstein, "ChaCha, a variant of Salsa20," 2008. [Online]. Available: https://cr.yp.to/chacha/chacha-20080128.pdf

[35] E. Ben-Sasson, I. Bentov, Y. Horesh, and M. Riabzev, "Scalable, transparent, and post-quantum secure computational integrity," Cryptology {ePrint} Archive, Tech. Rep. 2018/046, 2018. [Online]. Available: https://eprint.iacr.org/2018/046

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 753–758
DOI: 10.15439/2025F8280
ISSN 2300-5963 ACSIS, Vol. 43

NPDP programming for RISC multi-core processors

Mateusz Gruzewski, Marek Palkowski
West Pomeranian University of Technology in Szczecin
ul. Zolnierska 49, 71-210 Szczecin, Poland
e-mail: mpalkowski@zut.edu.pl

***Abstract*—In recent years, parallel architectures have become ubiquitous due to advancements in AI and cloud computing. However, parallel processing is not limited to x86 CISC CPUs and advanced graphics cards, GPUs; it also includes computations on ARM-based and RISC-V devices. In recent years, ARM processors have been adapted to incorporate an increasing number of cores. Today, mobile devices feature at least eight execution units, typically divided into energy-efficient and performance-oriented groups. RISC-based parallel processors are also integrated on development boards supported by the Linux kernel. In this article, we tested our NPDP Benchmark Suite for non-serial polyadic dynamic programming, primarily in the field of computer algorithms and bioinformatics, to evaluate the performance of the RISC processors under study, as well as code locality and cache efficiency. The benchmark consists of 10 kernels written in C++ and OpenMP. In the Android environment, we used the JAVA NDK (Native Development Kit) to port the application. For Apple machines, we used a port to OpenMP for parallelization. For the RISC-V native Linux environment, we applied the native Linux setup for efficient execution. Finally, we summarized the article and outlined future work.**

I. Introduction

RISC-based parallel processors, including ARM and RISC-V architectures, are becoming increasingly important in modern computing, particularly in mobile, embedded, and energy-efficient systems. Despite their growing adoption and multicore capabilities, their performance under complex, irregular workloads remains less explored compared to traditional x86-based platforms.

Nonserial Polyadic Dynamic Programming (NPDP) is an advanced class of dynamic programming techniques characterized by:

- Nonserial dependencies: subproblems do not follow a strict linear order;
- Polyadic recurrences: subproblems may depend on more than two others;
- Dynamic behavior: the dependency structure changes depending on input data or runtime conditions.

Such problems arise in domains like RNA folding, sequence alignment, and optimal polygon triangulation. Their nontrivial dependency patterns pose challenges for both parallel execution and efficient memory access.

Despite progress in compiler optimizations, limited attention has been paid to the performance of NPDP workloads on modern RISC-based architectures. In this work, we evaluate the execution-time performance of a 10-kernel NPDP benchmark suite compiled with OpenMP on several RISC platforms, including Apple ARM chips, Android smartphones, and a RISC-V development board. We apply three polyhedral compilers—Pluto, Traco, and Dapt—for automatic loop transformation and parallelization.

II. NPDP Benchmark Suite

The Non-serial Polyadic Dynamic Programming (NPDP) benchmark suite [1] consists of the following ten kernels:

1) Nussinov Algorithm - Predicts RNA secondary structures by identifying the optimal folding pattern.
2) Zuker Algorithm – An extension of RNA folding prediction that considers free energy minimization.
3) Smith–Waterman (SW) Algorithm – Performs local sequence alignment to identify similar regions between two nucleotide or protein sequences.
4) Needleman–Wunsch (NW) Algorithm – Conducts global sequence alignment for comparing entire sequences.
5) Smith–Waterman Algorithm for Three Sequences (SW3D) – Extends the original algorithm to align three sequences simultaneously.
6) Counting Algorithm – Used in RNA folding studies to count specific structural configurations.
7) McCaskill's Algorithm (Mcc) – Calculates the partition function and base-pairing probabilities in RNA secondary structures.
8) Maximum Expected Accuracy (MEA) Algorithm – Predicts RNA secondary structures based on maximizing expected accuracy.
9) Knuth's Algorithm – Determines the optimal binary search tree, minimizing the search time for a set of keys.
10) Optimal Polygon Triangulation – Divides a polygon into triangles to minimize the total edge length or another cost function.

III. Polyhedral optimization for NPDP loop nests

The polyhedral model represents loop nests as polyhedra with affine loop bounds and schedules. This model provides a foundation for advanced loop transformations and the analysis of data dependences. By harnessing the power of the polyhedral model, compilers can automatically optimize loops, enhance performance (especially in terms of locality with loop tiling), and exploit parallelism (particularly with loop skewing for NPDP codes) [2].

Nonserial Polyadic Dynamic Programming (NPDP) applications are generally easy to parallelize using a loop transformation technique known as skewing [3]. Skewing is a transformation applied to nested loops to restructure the iteration space,

ensuring that the dependent computations can be executed in parallel. It is particularly useful for NPDP problems where dependencies form a non-trivial pattern.

Consider a standard DP recurrence with dependencies on previous iterations:

$$DP(i,j) = f(DP(i-1,j), DP(i,j-1))$$

The corresponding loop structure might be:

```
for (i = 1; i <= N; i++) {
  for (j = 1; j <= N; j++) {
    DP[i][j] = f(DP[i-1][j], DP[i][j-1]);
  }
}
```

This loop cannot be directly parallelized because each iteration `(i, j)` depends on previous iterations. By applying the loop skewing transformation:

$$i' = i, \quad j' = i + j$$

we transform the loop structure into:

```
for(j' = 2; j' <= 2N; j'++) {
  lb = max(1, j'-N);
  ub = min(N, j'-1);
  for(i = lb; i <= ub; i++) {
    int j = j' - i;
    DP[i][j] = f(DP[i-1][j], DP[i][j-1]);
  }
}
```

Thus, NPDP problems, despite their complex dependencies, can be efficiently parallelized using loop skewing [3]. However, parallelization alone is not sufficient. Optimal cache utilization is also crucial, and scheduling blocks in NPDP remains challenging due to non-uniform dependencies [4], [5], [6].

NPDP problems, such as the Nussinov algorithm [7], can often be represented in the polyhedral model, enabling advanced loop transformations as performed by compilers like Pluto [8], Dapt [9], and Traco [5].

The polyhedral compilers Pluto [8], Dapt [9], and Traco [5] are based on affine transformation frameworks (ATF), space-time tiling, and tile correction, respectively. Pluto excels in generating well-balanced affine schedules, but for NPDP codes, the framework can address a set of affine equations to tile all loop nests [5]. ATF is unable to tile the Nussinov or SW kernels, and it cannot parallelize the tiled code of the Mcc kernel [1]. The Traco compiler generates 3D tiles using the transitive closure of the dependence graph of the union of loop dependences. For some tasks from the NPDP benchmark suite, it is not possible to compute the exact transitive closure of the dependency graph; in such cases, an over-approximation is applied [1]. Further transformations for 3D-tiling of NPDP codes were implemented in the Dapt compiler by dividing the iteration space into timed parallel spaces. Dapt addresses irregularities in code obtained in Traco, providing a comprehensive solution to optimize and refine the generated code [10].

Loop tiling, also referred to as loop blocking or loop partitioning, is a widely used optimization to improve cache efficiency [11]. It restructures loops into smaller blocks (tiles), reducing cache misses and improving memory access locality. In polyhedral compilation, tiling is often applied in combination with other transformations, such as skewing, to facilitate both locality and parallelism.

Ultimately, the compilers generate cache-efficient, parallelized code using the same underlying approaches—primarily the polyhedral model and its supporting library, ISL (Integer Set Library)—for both dependency analysis and code generation. This unified approach ensures that the code generation aligns with the dependency structure and optimizations applied.

Along with the original NPDP benchmark codes, the repository https://github.com/markpal/NPDP_Bench also includes OpenMP code generated by polyhedral compilers, which we adapted for execution on RISC-based devices.

IV. Experimental study

The experimental evaluation was conducted across three distinct RISC-based hardware platforms: Apple ARM-based laptops (M2 Pro, M3, and M4 Pro), a high performance RISC-V development board (Banana Pi BPI F3 SpacemiT K1), and an Android-based mobile device (Samsung Galaxy A55). All experiments focused on executing NPDP kernels compiled with OpenMP parallelisation support [12].

To ensure consistent and platform-specific testing, two separate benchmarking projects were developed. The first project, designed for macOS and the RISC-V platform, was responsible for executing the benchmarks and recording performance results to structured output files. The second project was implemented in Android Studio and dedicated to the Android operating system. It included a complete test execution scenario, with background task management and result storage integrated into the application logic.

Across all platforms, the NPDP computational kernels remained identical. Only the system-specific integration layer differed, adapting the same computational logic to the requirements of each operating system.

A. ARM-based Apple Platforms

The first experimental setup involved Apple machines equipped with ARM-based processors: M2 Pro, M3, and M4 Pro. These platforms feature heterogeneous multi-core architectures combining high-performance and energy-efficient cores. All experiments were conducted natively on macOS. Specifically, the M2 Pro system ran macOS Sequoia 15.1.1, the M3 system used macOS Sequoia 15.2, and the M4 Pro machine operated on macOS Sequoia 15.3. Since the default Apple Clang compiler does not provide support for OpenMP, the LLVM toolchain was installed via Homebrew, along with the necessary runtime library (libomp). The benchmarks were compiled using the following command:

```
/opt/homebrew/opt/llvm/bin/clang++
  -fopenmp \
  -I/opt/homebrew/opt/llvm/include \
  -L/opt/homebrew/opt/llvm/lib \
  -o test test.cpp
```

This configuration allowed for the successful compilation and execution of OpenMP-parallelized code on macOS. Tests were conducted in terminal shells with controlled background activity to minimize external interference. While we initially considered evaluating performance on iPhones, it was not possible to execute OpenMP code due to iOS restrictions. The operating system lacks support for dynamic linking against the OpenMP runtime, and instead promotes the use of proprietary concurrency models such as Grand Central Dispatch [13].

B. RISC-V Platform

The second experimental platform was based on the RISC-V architecture. We employed the Banana Pi BPI-F3 board [14], featuring an eight-core SpacemiT K1 [15], [16] processor with 16 GB LPDDR4 RAM and 128 GB of eMMC storage. The system ran Bianbu Linux 1.0rc1 (codename mantic) [17], a lightweight RISC-V–oriented distribution based on Ubuntu 23.10, designed for embedded and development use cases. The board was configured in a headless mode and accessed remotely via SSH over Ethernet.

Due to limited support for the RISC-V Vector Extension (RVV) in available prebuilt toolchains, we compiled a custom cross-compilation toolchain targeting rv64gcv on an external x86-64 machine. The NPDP benchmark suite was compiled with OpenMP and vectorization enabled, using the following command:

```
riscv64-unknown-linux-gnu-g++ \
  -march=rv64gcv -mabi=lp64d -O3 \
  -fopenmp -ftree-vectorize \
  -fopt-info-vec-optimized \
  -o test test.cpp
```

The compiled binaries were transferred to the Banana Pi board and executed under Bianbu Linux. This setup enabled the assessment of OpenMP-parallelized NPDP kernels on a native RISC-V platform with partial RVV support and a modern, minimal Linux runtime environment.

C. ARM-based Mobile Platform

The final test environment was a Samsung Galaxy A55 smartphone running Android 13. The device is powered by an ARM-based octa-core processor, comprising four Cortex-A78 cores (2.7 GHz) [18] and four Cortex-A55 cores (2.0 GHz) [19]. The NPDP benchmark suite was ported using the Android Native Development Kit (NDK) [20], allowing native C++ code to be executed within the Android runtime environment. The OpenMP-enabled kernels were compiled into shared libraries and invoked via a Java Native Interface (JNI) wrapper.

To minimize interference from the graphical interface and system events, the benchmarks were executed in the background using a dedicated background thread. This ensured isolated execution, independent of the UI thread. Results were written to files in the internal storage and subsequently retrieved for further analysis. This setup enabled reliable measurement of kernel performance under realistic conditions on mobile hardware, although variability due to thermal management and background activity could not be entirely eliminated. All measurements were repeated multiple times to ensure consistency.

D. Execution Time Constraints and Omitted Results

In all presented Tables 1-5, certain entries are marked with the symbol "–". These indicate cases in which a particular kernel execution was terminated or skipped due to exceeding predefined time limits. To maintain practical runtime and result comparability, thresholds were set at 20 minutes for input size 2200, and 40 minutes for larger sizes (5000 and 10000). If a kernel exceeded these limits, execution was interrupted and the result omitted from the final report.

This limitation was observed across all platforms and test configurations. Algorithms such as *Knuth*, *McCaskill*, *Triangulation*, and *Zuker* frequently exhibited prohibitive execution times, especially in the original and Traco-generated code variants. Additionally, computationally complex kernels such as *SW3D* and *MEA* consistently demonstrated exceedingly long runtimes and were therefore excluded from measurement.

As a result of preliminary testing, the input sizes used for RISC-V and Android platforms were deliberately limited to 500, 1000, and 2200. This decision was made to ensure feasibility on resource-constrained devices. Moreover, certain algorithms such as *Knuth* and *Zuker* were excluded entirely from tests on these two platforms due to previously observed stability and performance issues during execution, including excessive runtimes or failures. These exclusions were applied consistently across both platforms as a conservative measure to preserve experimental integrity.

The exclusion of the Knuth and Zuker algorithms was driven by consistent and reproducible performance issues. The Zuker kernel, which involves complex free-energy minimization and deeply nested loop structures, incurred high computational overhead. On several platforms, particularly those using Traco-generated variants, execution times exceeded predefined limits for larger input sizes. The Knuth algorithm, in turn, triggered instability on resource-constrained RISC-V systems, including segmentation faults and abnormal termination, likely due to its memory access patterns and recursive structure. These issues rendered reliable benchmarking infeasible and led to the omission of these kernels from the reported results.

It is important to note that all benchmarked algorithms were based on identical NPDP kernel implementations. The only platform-specific differences resided in the build system, integration method, and runtime environment.

TABLE I: Execution times of algorithms on M2 Pro (in seconds)

	Org			Traco			Pluto			Dapt		
Algorithm	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**
Nussinov	1.10	35.61	362.59	0.53	4.72	36.28	0.24	5.83	56.14	0.31	3.24	23.89
Counting	4.38	85.47	1010.36	0.67	10.41	135.36	0.94	10.34	97.47	0.65	11.72	135.67
Knuth	1.65	45.62	-	2.50	41.46	-	0.33	8.50	-	0.26	7.64	-
Mcc	7.96	195.77	-	1.56	31.55	-	3.40	162.98	-	1.21	43.64	-
Triang	2.74	99.51	-	0.93	14.87	-	0.57	24.22	-	0.69	23.97	-
Zuker	392.47	-	-	43.91	-	-	15.17	-	-	16.47	-	-
NW	5.81	249.22	2299.86	0.96	62.34	500.14	2.04	71.42	569.67	1.09	69.79	576.01
SW	6.26	267.26	-	1.20	61.78	505.01	1.79	72.21	585.32	1.13	72.06	576.97

TABLE II: Execution times of algorithms on M3 (in seconds)

	Org			Traco			Pluto			Dapt		
Algorithm	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**
Nussinov	1.13	34.40	336.34	0.60	6.41	52.73	0.29	6.43	67.48	0.58	5.54	40.64
Counting	4.22	57.26	739.27	1.12	14.10	198.04	1.21	12.08	135.17	0.97	13.75	197.44
Knuth	1.94	51.16	-	1.70	26.49	-	0.35	10.46	-	0.38	11.57	-
Mcc	1.14	8.48	-	0.25	1.63	-	0.64	2.76	-	0.29	1.91	-
Triang	3.37	178.10	-	0.92	25.25	-	0.72	40.42	-	1.02	50.82	-
Zuker	288.26	-	-	54.67	-	-	25.71	-	-	32.68	-	-
NW	6.81	351.05	2483.51	1.50	67.36	678.42	1.59	115.30	1192.05	2.06	96.08	857.60
SW	8.37	494.12	-	2.11	83.54	864.38	1.55	104.24	1087.03	1.87	79.62	830.81

TABLE III: Execution times of algorithms on M4 Pro (in seconds)

	Org			Traco			Pluto			Dapt		
Algorithm	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**	**2200**	**5000**	**10000**
Nussinov	0.92	20.32	244.97	0.49	3.28	21.11	0.20	2.75	28.44	0.29	2.56	17.85
Counting	4.54	61.70	718.30	0.62	7.09	84.69	1.07	12.08	112.15	0.54	6.38	80.93
Knuth	1.61	56.41	-	1.59	25.15	-	0.21	6.98	-	0.20	6.72	-
Mcc	7.63	360.85	-	1.21	56.16	-	4.57	276.38	-	0.95	52.85	-
Triang	4.74	126.24	-	0.48	14.32	-	0.93	17.60	-	0.64	15.90	-
Zuker	238.54	-	-	23.16	-	-	12.16	-	-	11.57	-	-
NW	11.06	606.88	-	1.12	87.72	490.37	2.04	98.99	516.79	1.21	131.91	768.41
SW	12.89	599.46	-	1.19	182.67	1195.24	2.06	97.16	594.67	1.31	176.46	981.69

TABLE IV: Execution times of algorithms on Risc-V (in seconds)

	Org			Traco			Pluto			Dapt		
Algorithm	**500**	**1000**	**2200**	**500**	**1000**	**2200**	**500**	**1000**	**2200**	**500**	**1000**	**2200**
Nussinov	0.28	5.26	98.36	0.19	1.34	9.35	0.07	0.74	20.63	0.11	0.74	8.89
Counting	0.64	4.77	68.33	0.15	0.70	11.9	0.30	1.44	11.35	0.14	0.72	10.89
Knuth	-	-	-	-	-	-	-	-	-	-	-	-
Mcc	9.42	17.52	272.59	1.20	3.33	51.96	1.69	8.54	100.66	1.78	4.04	100.99
Triang	5.12	42.38	490.66	1.12	7.71	71.36	1.17	8.65	105.52	1.02	7.60	98.06
Zuker	-	-	-	-	-	-	-	-	-	-	-	-
NW	1.99	26.77	382.56	0.37	5.47	106.47	0.91	9.58	159.87	0.37	6.30	131.24
SW	1.98	26.37	381.60	0.39	5.30	104.66	0.40	6.52	124.30	0.38	5.97	124.12

E. Result Discussion

Tables I–V present execution times for all kernels across four code variants (Org, Traco, Pluto, Dapt) and five hardware platforms. Overall, the results confirm that polyhedral compilation significantly improves performance, and that Dapt frequently produces the fastest code across most tested configurations.

On Apple ARM platforms (M2–M4), the Dapt-generated code consistently outperforms both the original and other compiler variants. For example, on the M4 Pro (Table III), Dapt executes the Nussinov kernel in 17.85s for input size 10000, while the original version requires 244.97s. This trend is visible across most kernels and input sizes. However, it is worth noting that Apple systems are based on a heterogeneous architecture not explicitly designed for HPC-like parallel workloads. Their strong performance stems primarily from advanced chip design, high memory bandwidth, and aggressive power management at the cost of being proprietary and expensive.

Although Dapt yields strong results overall, certain kernels like Counting show comparable or even better performance with Traco on selected platforms. This may be due to Traco's tile correction approach occasionally producing layouts with better cache fit for specific iteration spaces [10]. While Dapt generally provides more consistent tiling across irregular kernels, localized cache effects may favor Traco in select scenarios.

The RISC-V platform, represented by the low-cost Banana Pi BPI-F3 board, completed all tests but with significantly

TABLE V: Execution times of algorithms on Samsung Galaxy A55 (in seconds)

	Org			Traco			Pluto			Dapt		
Algorithm	**500**	**1000**	**2200**	**500**	**1000**	**2200**	**500**	**1000**	**2200**	**500**	**1000**	**2200**
Nussinov	0.14	1.17	13.11	0.15	2.18	17.16	0.19	0.87	4.13	0.20	0.83	7.27
Counting	0.38	2.83	36.07	0.22	1.76	10.15	0.33	2.83	20.45	0.16	1.18	8.63
Knuth	-	-	-	-	-	-	-	-	-	-	-	-
Mcc	0.42	3.59	72.35	0.31	2.28	22.37	0.28	2.37	29.69	0.23	1.50	16.76
Triang	1.05	8.04	127.95	0.80	5.80	43.78	0.89	3.74	37.60	0.82	3.42	35.75
Zuker	-	-	-	-	-	-	-	-	-	-	-	-
NW	0.88	7.16	111.91	0.49	2.75	27.49	0.67	2.99	30.64	0.82	2.91	28.98
SW	1.00	8.09	124.30	0.55	3.13	27.36	0.78	3.04	32.49	0.79	2.97	34.10

higher execution times. This reflects both early-stage toolchain maturity and the limited performance of the tested hardware. RISC-V remains promising due to its openness and Linux compatibility, and future work will explore more capable boards.

The Android platform (Samsung Galaxy A55) exhibits intermediate results. Despite using a modern ARM-based SoC, the performance is limited by the operating system, which does not prioritize background computation. For instance, thermal throttling and restricted OpenMP threading in the ART environment reduce the potential gains from polyhedral optimizations. Android also introduces unpredictable performance variability due to dynamic task scheduling and user-space constraints. As a result, OpenMP kernels run slower and less deterministically compared to Linux-based systems.

While the Dapt compiler provides robust performance across all platforms, the relative gains vary depending on the architecture. Moreover, some kernels (e.g., Knuth, Zuker, SW3D, MEA) were excluded due to excessive runtimes or instability, especially in Traco or original versions. This was consistent across all environments and input sizes.

In summary, the results highlight the benefits of polyhedral compilation techniques for NPDP workloads and expose the current disparity between platform classes. These results are in line with earlier work on multi-core systems, where OpenMP-based parallelization and loop restructuring also led to notable performance improvements in geospatial workloads [21]. Apple systems deliver the highest raw performance but are closed and costly. RISC-V is significantly slower on current boards but offers openness, modularity, and full Linux support — key advantages for long-term research. Android devices, while architecturally capable, suffer from system-level limitations that reduce usable compute throughput. These findings justify the need for continued testing on new RISC-V platforms and more refined kernel-level optimization.

V. Conclusion

When comparing research conducted on RISC processors, it is essential to consider factors such as cost, licensing model, and the intended application of the processing unit itself. In terms of raw performance, the fastest processors currently available are undoubtedly the high-end Apple M4 Pro units, although they also come at a significantly higher cost. However, Apple and Cortex processors are based on closed architectures, which limits their utility in development and research contexts. In contrast, RISC-V processors, due to their open-source architecture, offer far greater flexibility and potential for innovation.

Unfortunately, the studied Banana chip performance still lags behind ARM-based solutions. While Linux kernel support exists, substantial work remains to improve overall efficiency. Features such as vector processing are still in the developmental stage, and compilers require significant refinement to reach production-grade maturity. Nonetheless, the open-source nature of the platform and the dedicated efforts of volunteer contributors suggest a promising future—potentially more so than the long-standing but limited Android NDK.

In terms of performance at the same price point, it is currently easier to utilize older CISC CPUs or other types of RISC-based GPUs. While multi-board computers like Banana, Leeche, or other RISC-V developer boards are becoming more widespread, especially among emerging manufacturers from countries affected by U.S. sanctions, their performance capabilities are still maturing and not yet competitive with mainstream alternatives. However, the performance of RISC-V may improve in the future—the question is how rapidly this improvement will occur. We also plan to evaluate emerging high-performance RISC-V platforms, including 64-core development boards that are now becoming commercially available [22].

In future work, we will extend our research framework to include other polyhedral benchmark suites and a detailed analysis of the energy consumption on these devices.

References

[1] M. Palkowski and W. Bielecki, “NPDP benchmark suite for the evaluation of the effectiveness of automatic optimizing compilers,” *Parallel Computing*, vol. 116, p. 103016, Jul. 2023. doi: 10.1016/j.parco.2023.103016. [Online]. Available: https://doi.org/10.1016/j.parco.2023.103016

[2] S. Verdoolaege, “Integer set library - manual,” www.kotnet.org/~skimo/isl/manual.pdf, 2011, accessed on: 2024-01-11.

[3] L. Liu, M. Wang, J. Jiang, R. Li, and G. Yang, “Efficient nonserial polyadic dynamic programming on the cell processor.” in *IPDPS Workshops*. Anchorage, Alaska: IEEE, 2011, pp. 460–471.

[4] R. T. Mullapudi and U. Bondhugula, “Tiling for dynamic scheduling,” in *Proceedings of the 4th International Workshop on Polyhedral Compilation Techniques*, S. Rajopadhye and S. Verdoolaege, Eds., Vienna, Austria, Jan. 2014.

[5] M. Palkowski and W. Bielecki, “Parallel tiled Nussinov RNA folding loop nest generated using both dependence graph transitive closure and loop skewing,” *BMC Bioinformatics*, vol. 18, no. 1, p. 290, 2017. doi: 10.1186/s12859-017-1707-8

[6] V. K. Tchendji, F. I. K. Youmbi, C. T. Djamegni, and J. L. Zeutouo, "A parallel tiled and sparsified Four-Russians algorithm for Nussinov's RNA folding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–12, 2022. doi: 10.1109/tcbb.2022.3216826
[7] R. Nussinov *et al.*, "Algorithms for loop matchings," *SIAM Journal on Applied mathematics*, vol. 35, no. 1, pp. 68–82, 1978.
[8] U. Bondhugula *et al.*, "A practical automatic polyhedral parallelizer and locality optimizer," *SIGPLAN Not.*, vol. 43, no. 6, pp. 101–113, Jun. 2008. doi: 10.1145/1379022.1375595
[9] W. Bielecki and M. Poliwoda, "Automatic parallel tiled code generation based on dependence approximation," in *Parallel Computing Technologies*, V. Malyshkin, Ed. Cham: Springer International Publishing, 2021, pp. 260–275.
[10] M. Palkowski and M. Gruzewski, "Time and energy benefits of using automatic optimization compilers for NPDP tasks," *Electronics*, vol. 12, no. 17, p. 3579, Aug. 2023. doi: 10.3390/electronics12173579. [Online]. Available: http://dx.doi.org/10.3390/electronics12173579
[11] J. Xue, *Loop Tiling for Parallelism*. Norwell, MA, USA: Kluwer Academic Publishers, 2000. ISBN 0-7923-7933-0
[12] OpenMP Architecture Review Board, "OpenMP application program interface version 5.2," https://www.openmp.org/specifications, 2021, accessed on: 2023-10-22.
[13] Apple Inc., "Grand central dispatch (gcd)," https://developer.apple.com/documentation/dispatch, 2009, accessed: 2025-05-05.
[14] Banana Pi Team, "Banana pi bpi-f3 documentation," https://docs.banana-pi.org/en/BPI-F3/BananaPi_BPI-F3, 2025, accessed: 2025-05-05.
[15] SpacemiT, "Spacemit key stone k1 – octa-core 64-bit risc-v ai cpu," https://www.spacemit.com/en/key-stone-k1/, 2024, accessed: 2025-05-05.
[16] Banana Pi Team, "Spacemit k1 8-core risc-v chip brief," https://docs.banana-pi.org/en/BPI-F3/SpacemiT_K1, 2024, accessed: 2025-05-05.
[17] SpacemiT, "Bianbu linux," https://bianbu.spacemit.com/en/, 2024, accessed: 2025-05-05.
[18] Arm Ltd., "Arm cortex-a78 processor," https://developer.arm.com/Processors/Cortex-A78, 2020, accessed: 2025-05-05.
[19] ——, "Arm cortex-a55 processor," https://developer.arm.com/Processors/Cortex-A55, 2017, accessed: 2025-05-05.
[20] Google LLC, "Android ndk – native development kit," https://developer.android.com/ndk, 2024, accessed: 2025-05-05.
[21] B. Bylina, J. Potiopa, M. Klisowski, and J. Bylina, "The impact of vectorization and parallelization of the slope algorithm on performance and energy efficiency on multi-core architecture," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 25. PTI, 2021. doi: 10.15439/2021F68 pp. 283–290.
[22] Sophgo, "Sg2042 risc-v 64-core soc," 2025, accessed on: 2024-07-13. [Online]. Available: https://en.sophgo.com/sophon-u/product/introduce/sg2042.html

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 759–764
DOI: 10.15439/2025F2264
ISSN 2300-5963 ACSIS, Vol. 43

Bias in Classical Life Tables Under Censoring: A Comparative Study With Kaplan-Meier Estimation and Actuarial Estimation Using Real and Simulated Data

Lubomír Seif†, Ondřej Vít‡
†ORCiD: 0009-0003-7444-9425
‡ORCiD: 0009-0003-7317-5856
Department of Statistics and Probability
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
†Email: lubomir.seif@vse.cz
‡Email: ondrej.vit@vse.cz

Lubomír Štěpánek[1, 2, 3]
ORCiD: 0000-0002-8308-4304
[1]Department of Statistics and Probability
[2]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
Email: lubomir.stepanek@vse.cz
&
[3]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, 12000 Prague, Czech Republic
Email: lubomir.stepanek@lf1.cuni.cz

***Abstract*—Classical life tables assume fully observed lifespans and ignore censoring, which can bias survival estimates. In this study, we compare the classical life table with two censoring-aware approaches: the Kaplan-Meier estimator applied to simulated censored lifetimes, and the actuarial estimator assuming uniform censoring within intervals.**

Using mortality data for the Czech Republic (from year 2021), we show that both the mentioned alternative methods yield systematically lower survival and life expectancy estimates, compared to life tables-based approach. The actuarial estimator is the most conservative, and we provide a formal proof that it underestimates survival relative to Kaplan-Meier.

These differences have practical implications for actuarial pricing and longevity risk. We advocate incorporating survival analysis techniques into actuarial workflows when dealing with incomplete or censored data.

***Index Terms*—life tables, survival analysis, censoring, Kaplan-Meier estimator, actuarial estimator, life expectancy overestimation, life expectancy underestimation, actuarial science**

I. Introduction

LIFE tables are a cornerstone of actuarial science and demographic research, offering standardized tools to assess mortality and survival in a given population. Based on aggregated counts of deaths and population exposures at each age, these tables enable the computation of key quantities such as age-specific death probabilities and life expectancy. In standard construction, however, life tables operate under the assumption of fully observed lifespans and do not incorporate censoring the fact that, in many contexts, we do not observe the complete lifespan for every individual.

Although this assumption holds in retrospective national mortality data, it becomes increasingly problematic in settings involving survey-based samples, incomplete registration systems, or real-time monitoring, where censoring is inherent in the data structure. Censoring occurs naturally when the full path of survival is not observed, either due to loss to follow-up, study end, or delayed entry. Right-censoring means that we do not know when an individual would experience the event of interest because the monitoring period ended before the event could occur.

The distinction could be nontrivial. As noted by Slud [1], the classical construction of life tables, based on central death rates and complete cohorts, fails to account for right-censoring, which can bias survival estimates if present but ignored. Missov and Vaupel [2] further explore the impact of censoring on mortality measures, showing that even small amounts of right-censoring can influence life expectancy and related metrics. Modern survival analysis, in contrast, is equipped to handle censored observations through estimators such as Kaplan-Meier, Nelson-Aalen, and parametric survival models.

In this study, we revisit classical life table methods using mortality data from the Human Mortality Database (HMD) [3] for the Czech Republic in 2021. We begin by constructing a traditional life table and estimating survival probabilities and life expectancy under the assumption of complete data.

This research was supported by the grant no. F4/36/2025 which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

We then introduce synthetic censoring at the aggregate level and evaluate its effect on survival metrics. Next, we simulate individual-level lifetimes with right-censoring and estimate the survival function [4] using the Kaplan-Meier estimator, a non-parametric method that accommodates incomplete observation [5], [6].

In addition, we implement the actuarial estimator of the survival function, which assumes uniformly distributed censoring within fixed intervals and is widely used when exact death times are not available [7]. We analytically demonstrate that this estimator yields lower survival estimates than Kaplan-Meier under the same data, and we confirm this property empirically.

By comparing all three approaches, i.e. the classical life table, Kaplan-Meier, and actuarial estimator, we quantify the extent of deviation in survival curves and life expectancy measures due to censoring and method choice. Finally, we discuss the implications of these differences and possible underestimation of the survivals for actuarial pricing, reserving, and longevity risk assessment.

II. Preliminaries

Life tables are traditionally constructed using a deterministic sequence of transformations, starting from observed age-specific death rates or death probabilities. The standard components of a complete life table include the probability of dying between ages x and $x+1$ (denoted as q_x), the number of survivors at age x (marked as l_x), the number of deaths at age x (d_x), the person-years lived between ages x and $x+1$ (here denoted as L_x), the total person-years remaining beyond a given age x (T_x), and the remaining life expectancy at age x (marked as e_x). These definitions are extensively documented in actuarial and demographic literature, for instance [8], [9], [10].

The death probability q_x is typically calculated from empirical data as

$$q_x = \frac{D_x}{N_x}, \tag{1}$$

where D_x is the number of deaths observed at age x, and N_x is the number of individuals exposed to risk at that age.

Starting from a radix $l_0 = 100,000$, the number of survivors at each age is recursively defined as

$$l_{x+1} = l_x(1 - q_x). \tag{2}$$

From this, the number of deaths is given by

$$d_x = l_x q_x, \tag{3}$$

and the number of person-years lived between ages x and $x+1$ is typically approximated by

$$L_x = l_x - \frac{1}{2}d_x, \tag{4}$$

assuming that deaths are uniformly distributed within each age interval. The cumulative future person-years T_x and the corresponding life expectancy e_x are then computed as

$$T_x = \sum_{k=x}^{\omega} L_k, \qquad e_x = \frac{T_x}{l_x}, \tag{5}$$

respectively, where ω denotes the maximum age in the life table.

III. Methods

In this section, we implement the above methodology using mortality and exposure data from the Human Mortality Database (HMD) for the Czech Republic in 2021. The resulting classical life table will serve as the benchmark for later comparison with methods that incorporate synthetic censoring.

All computations and visualizations were performed using the R language [11] with the packages `ggplot2` [12], `dplyr` [13], and `survival` [14], [15].

A. Classical Life Table Construction

Using mortality and exposure data from the Human Mortality Database for the Czech Republic in 2021, we constructed a complete life table following the standard actuarial approach. The key steps involved computing age-specific probabilities of death q_x, the number of survivors l_x, the number of deaths d_x, person-years lived L_x, cumulative person-years T_x, and remaining life expectancy e_x, always considering age x. The radix was set at $l_0 = 100{,}000$, and the last observed age $\omega = 109$ was treated as an absorbing state with $q_{109} = 1$, meaning that all remaining individuals are assumed to die within the interval $\langle 109, 110)$, as is standard in demographic practice, [8], [9], and is feasible even for research purposes, including this proceeding.

Part of the resulting table for ages 0–9 is shown in Tables I and II. It illustrates the decreasing pattern of life expectancy with age and the consistent structure of death probabilities and exposures.

Fig. 1 presents the curve of remaining life expectancy e_x as a function of age. The smooth and decreasing trajectory reflects the typical aging pattern in low-mortality populations.

B. Incorporating Synthetic Censoring

To assess how censoring could influence life table calculations even when working with aggregated population-level data, we construct a synthetic scenario in which a portion of the population at each age is assumed to be right-censored. This mimics conditions observed in longitudinal surveys, clinical studies, or incomplete registration systems, where not all individuals' times of death are observed. Since the dataset Human Mortality Database (HMD) does not include (or even

TABLE I
First 10 observations of HMD data (Czech Republic, 2021)

Age	Deaths	Exposure
0	246	110751.7
1	15	109581.9
2	12	112244.1
3	13	114184.9
4	11	113540.5
5	6	112191.5
6	18	111140.5
7	14	109107.7
8	8	108884.5
9	3	109051.5

TABLE II
FIRST 10 OBSERVATIONS OF THE CONSTRUCTED CLASSICAL LIFE TABLE (CZECH REPUBLIC, 2021)

Age	q_x	l_x	d_x	L_x	T_x	e_x
0	0.002221	100000.00	222.12	99888.94	7699599	76.996
1	0.000137	99777.88	13.66	99771.05	7599710	76.166
2	0.000107	99764.22	10.67	99758.89	7499939	75.177
3	0.000114	99753.56	11.36	99747.88	7400180	74.185
4	0.000097	99742.20	9.66	99737.37	7300433	73.193
5	0.000054	99732.54	5.33	99729.87	7200695	72.200
6	0.000162	99727.20	16.15	99719.13	7100965	71.204
7	0.000128	99711.05	12.79	99704.66	7001246	70.215
8	0.000073	99698.26	7.33	99694.60	6901542	69.224
9	0.000028	99690.93	2.74	99689.56	6801847	68.229

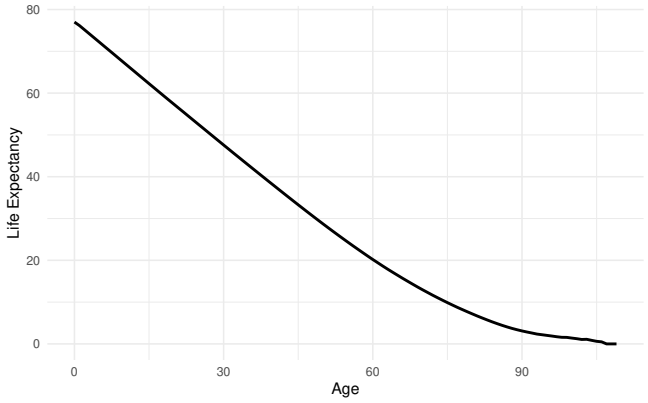


Fig. 1. Remaining life expectancy (in years) at age x based on classical life table construction.

consider) censoring, the synthetic, i.e. simulated censoring is the only way how to investigate its impact.

In our synthetic setup, we assume a fixed censoring probability c_x at each age. Specifically, we define $c_x = 0.10$, i.e., 10 % of the exposure at each age is assumed to be censored. This means that instead of observing N_x person-years fully at risk, we effectively observe only $(1 - c_x)N_x$ as uncensored exposure, while $c_x N_x$ is assumed to be censored.

Following the framework suggested by Slud [1] and the analytical concerns raised in Missov and Vaupel [2], we adjust the death probability coming from formula (1) calculation accordingly,

$$\tilde{q}_x = \frac{D_x}{(1 - c_x)N_x}, \tag{6}$$

where $\tilde{q}_x$ represents the adjusted probability of death under censoring. All subsequent life table quantities $\tilde{l}_x, \tilde{d}_x, \tilde{L}_x, \tilde{T}_x, \tilde{e}_x$ from formulae (2, 3, 4, 5) are computed using this modified death probability $\tilde{q}_x$, replicating the structure of the classical method.

We emphasize that this adjustment does not imply individual-level censoring but rather a proxy for its effect at the population level. The goal is to examine whether such synthetic censoring, even when introduced into aggregate mortality data, leads to meaningful differences in key survival metrics such as life expectancy.

In the next subsection, we present the life table results under the synthetic censoring assumption and compare them to the classical baseline.

C. Comparison of Results under Censoring

Fig. 2 shows the comparison between the classical and synthetically censored life tables in terms of remaining life expectancy. As expected, the life expectancy under synthetic censoring is systematically lower in all ages. This is a direct consequence of inflating the estimated mortality probabilities when only partial exposures are assumed to be observed. In other words, if classical life tables-based approach assumes non-censored framework when probability of death at given age x is, using formula (1), equal to $q_x = \frac{D_x}{N_x}$, while the death probability under censoring is (6), $\tilde{q}_x = \frac{D_x}{(1-c_x)N_x}$, obviously is $\tilde{q}_x > q_x$ for any $c_x > 0$, i.e., whenever there is nonzero censoring we are not aware of, the probability of death is in fact a bit higher and, consequently, the true remaining life expectancy is a bit lower.

When we assume that data are not censored we might get falsely optimistic classical estimate used in life tables. When incorporating censoring we can always see the downward shift in estimated life expectancy curve which might be closer to the reality.

In further sections we discuss methods for estimating survival function which is then used to model life expectnancy. We present the actuarial estimator and the Kaplan-Meier estimator.

IV. ACTUARIAL (LIFETABLE) ESTIMATOR

One of the approaches for estimating the survival function, i.e., a probability that an object will survive longer than a specified time point [4], is the actuarial or life table estimator [7]. This method is suitable for interval-censored or grouped data where exact times of death or censoring are not known. It is commonly used in demography and actuarial science when data are available in age or time intervals.

Let the time axis be partitioned into intervals $\langle t_{j-1}, t_j)$. For each interval, we define

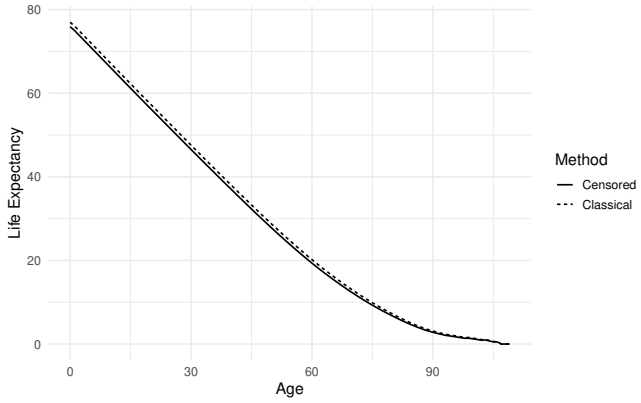


Fig. 2. Comparison of remaining life expectancy between classical life table and synthetic censoring scenario ($c_x = 10\,\%$).

- r_j as the number of individuals entering interval j,
- d_j as the number of deaths observed in interval j,
- c_j as the number of right-censored observations in interval j.

Assuming uniform censoring within the interval, the effective number at risk is approximated by

$$r'_j = r_j - \frac{c_j}{2}, \tag{7}$$

clearly saying that at a midpoint of the given interval (j), there is just half the censoring ($\frac{c_j}{2}$) of the entire interval (c_j). The conditional survival probability in interval j is then

$$p_j = \frac{r'_j - d_j}{r'_j} = 1 - \frac{d_j}{r'_j}. \tag{8}$$

The actuarial estimate of the survival function $\hat{S}(t)$ is obtained as the cumulative product,

$$\hat{S}_A(t) = \prod_{j:t_j \le t} p_j, \tag{9}$$

since, if an individial survived up to interval j, they had to survive also up to intervals $j-1, j-2, \ldots, 1$. Thus, we have to mutually multiply the interval-related probabilities p_j.

This estimator is generally more conservative in estimating survival than Kaplan-Meier, especially when censoring is heavy or uneven. The assumption of uniformly distributed censoring within intervals can lead to a downward bias in estimated survival compared to methods that use exact event times.

V. Kaplan-Meier Estimator

The Kaplan-Meier estimator is a non-parametric method for estimating the survival function $S(x)$, which represents the probability that a subject survives beyond age x [5], [6]. It is particularly well-suited for handling right-censored data, where the exact time of death is not observed for some individuals.

Let $t_1 < t_2 < \cdots < t_k$ denote the distinct observed failure times in the sample, d_j the number of events (deaths) between time points t_{j-i} and t_j, and r_j the number of individuals at risk just before t_j. Then the Kaplan-Meier estimator is given by

$$\hat{S}_{KM}(t) = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{r_j}\right). \tag{10}$$

This step function decreases only at observed event times and remains constant in intervals where no events occur. Censored observations reduce the risk set r_j at subsequent time points but do not directly contribute to d_j.

Once the survival function is estimated, the life expectancy at birth e_0 can be computed as the area under the curve (which is not differentiable but is continuous, though),

$$e_0^{KM} = \int_0^\infty \hat{S}(x)\, dx, \tag{11}$$

which in practice is approximated as the sum over the discrete age grid,

$$e_0^{KM} \approx \sum_x \hat{S}(x), \tag{12}$$

where $\hat{S}(x)$ is an estimate of survival function, modeled using actuarial estimate $\hat{S}_A$ or Kaplan-Meier estimate $\hat{S}_{KM}$, as introduced before. This provides a consistent estimator of the mean remaining lifetime even in the presence of censoring, unlike classical life table methods that assume fully observed (non-censored) data.

We compare the actuarial estimator with the Kaplan-Meier estimator in the next section and quantify their divergence in terms of survival function and life expectancy.

VI. The Actuarial Estimator Underestimates Survival Compared to Kaplan-Meier Estimator

We now provide an analytical justification for why the actuarial estimator $\hat{S}_A(t)$ is systematically lower than the Kaplan-Meier estimator $\hat{S}_{KM}(t)$ under the same data.

Recall that in Kaplan-Meier from formula (10), the survival function decreases only at exact failure times,

$$\hat{S}_{KM}(t_j) = \hat{S}_{KM}(t_{j-1}) \cdot \left(1 - \frac{d_j}{r_j}\right), \tag{13}$$

where r_j is the number at risk just before time point t_j.

By contrast, the actuarial estimator assumes that censoring is uniformly distributed within interval $\langle t_{j-1}, t_j)$, so the effective number at risk is adjusted to

$$r'_j = r_j - \frac{c_j}{2}, \tag{14}$$

and the survival estimate is

$$\hat{S}_A(t_j) = \hat{S}_A(t_{j-1}) \cdot \left(1 - \frac{d_j}{r'_j}\right). \tag{15}$$

As comes from formulae (13, 15), terms $1 - \frac{d_j}{r_j}$ and $1 - \frac{d_j}{r'_j}$ are the key. Because $r'_j = r_j - \frac{c_j}{2} < r_j$ in most practical cases (i.e., if any censoring $c_j > 0$ occurs before the end of the interval), we have

$$\frac{d_j}{r'_j} > \frac{d_j}{r_j}, \tag{16}$$

which implies

$$1 - \frac{d_j}{r'_j} < 1 - \frac{d_j}{r_j}, \tag{17}$$

for every time point $t_j \le t$, and therefore, multiplying the terms from previous formula (17), we get

$$1 - \frac{d_j}{r'_j} < 1 - \frac{d_j}{r_j} \qquad \forall t_j \in \{t_1, t_2, \ldots, t\}$$

$$\prod_{j:t_j \le t} \left(1 - \frac{d_j}{r'_j}\right) < \prod_{j:t_j \le t} \left(1 - \frac{d_j}{r_j}\right)$$

$$\hat{S}_A(t) \overset{(8,9,10)}{<} \hat{S}_{KM}(t). \tag{18}$$

This inequality propagates multiplicatively, so the entire survival curve under the actuarial method lies below the Kaplan-Meier curve, if any censoring occurs. Hence, the actuarial estimator systematically underestimates survival compared to Kaplan-Meier, especially when censoring is substantial.

This property also directly impacts life expectancy estimates derived as area under the survival curve.

VII. Estimation with Simulated Censoring

To analyze the effect of censoring using a statistically rigorous survival model, we simulated individual-level lifetimes from the classical life table probabilities q_x and applied a 10 % right-censoring mechanism. We then estimated the survival function using the Kaplan-Meier estimator.

In parallel, we computed the actuarial survival function based on interval-level data and uniform censoring assumptions. The corresponding life expectancy was obtained as the discrete area under the actuarial survival curve,

$$e_0^A = \sum_x \hat{S}_A(x), \tag{19}$$

where $\hat{S}_A(x)$ denotes the actuarial survival estimate at the beginning of age x. This summation provides a consistent approximation of the expected remaining lifetime under the actuarial model.

Fig. 3 shows the comparison between the Kaplan-Meier estimate, actuarial estimate and the classical life table survival function $S(x) = l_x/l_0$, as the first two come from formulae (10, 9). The Kaplan-Meier curve lies systematically below the classical curve, indicating shorter estimated survival due to the presence of censoring. The curve estimated by the actuarial estimate lies below both curves. Fig. 4 shows estimated survival for ages above 75 years. The differences between the estimates, and, consequently, between the treatment of the censoring, are more clear in this figure.

In terms of overall life expectancy at birth[1], the classical method based on l_x gave $e_0^{LT} = 77.50$ (95% CI $[77.43, 77.57]$), while the Kaplan-Meier estimator yielded $e_0^{KM} = 76.54$ (95% CI $[76.46, 76.63]$) and the actuarial estimate gave $e_0^A = 75.97$ (95% CI $[75.90, 76.04]$), respectively. The differences here are quiet obvious. Between classical method and Kaplan-Meier estimate, the difference is approximately 0.96 years. Between classical estimate and actuarial estimate, the difference is 1.53 years and when comparing Kaplan-Meier and actuarial estimate the difference shows 0.57 years.

VIII. Discussion

The results of our study provide a comprehensive assessment of how censoring affects survival estimation using three different approaches: the classical life table, the Kaplan-Meier estimator, and the actuarial survival estimator. All methods were applied consistently to the same underlying mortality data for the Czech Republic in 2021, using synthetic (simulated) censoring.

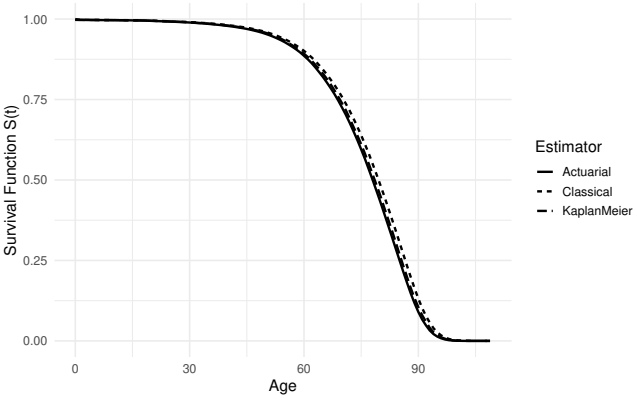


Fig. 3. Comparison of survival functions: Kaplan-Meier estimate, actuarial estimate and life table estimate.

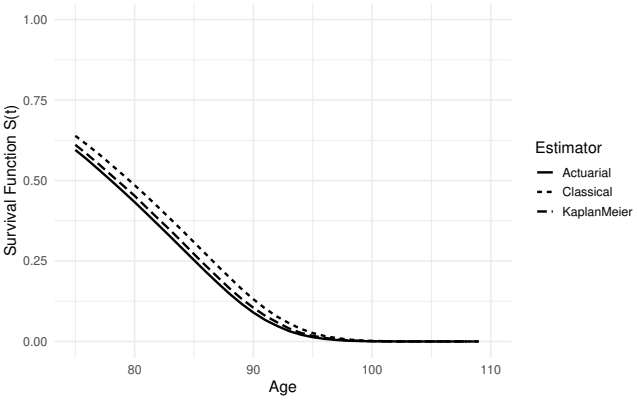


Fig. 4. Comparison of survival functions for ages 75+ years.

Our analyses revealed that the assumptions of no censoring in life table estimates might lead to falsely optimistic estimates. Incorporating censoring could bring the estimate closer to the reality. The classical life table estimate of life expectancy at birth was 77.50 years (95% CI $[77.43, 77.57]$), while the Kaplan-Meier estimator applied to simulated censored lifetimes yielded 76.54 years (95% CI $[76.46, 76.63]$), a nearly one-year difference. When the actuarial estimator was applied under uniform censoring assumptions (75.97 with 95% CI $[75.90, 76.04]$), life expectancy was lower still, confirming analytically and empirically that this method is the most conservative among the three.

From a demographic perspective, this highlights the critical importance of censoring-aware methodologies, especially in datasets where full lifespan observation is implausible. Ignoring censoring can lead to biased survival estimates and misleading conclusions about population longevity.

In an actuarial context, these differences carry tangible

[1]To report variability of the simulated estimates, 95% confidence intervals are constructed using the percentile bootstrap, i.e., using the 0.025-th and 0.975-th quantiles of the bootstrapped distribution of each estimator. Confidence intervals are reported in square brackets, preceded by the prefix "95% CI".

consequences. Life expectancy estimates directly influence pricing, reserving, and solvency calculations in life insurance and pension products. A lower estimated life expectancy typically leads to higher premiums for annuity products and lower reserves for life insurance, potentially introducing pricing inaccuracies if censoring is not handled correctly. Conversely, underestimating survival in products with longevity risk exposure may lead to systematic underreserving, e.g. two people with the same product but different life expectancy should lead to higher reserves for the person with higher probability of death. But when censoring is not accounted for these reserves could calculated to lower volume leading to underreserving. In other words when calculating death insurance under falsely optimistic estimates, lower probability of death (higher life expectancy) could lead to cheaper insurance policies and vice versa.

These findings reinforce the need to integrate non-parametric survival estimators such as Kaplan-Meier into actuarial workflows when individual-level or censored data are available. Even when working with grouped data, awareness of the limitations of life table and actuarial approximations is essential.

IX. Future Outlooks

This study opens several directions for further research. First, extending the analysis to various censoring levels (e.g., censoring of 5 %, 20 %, 30 % of available data) would reveal how estimator performance changes with increasing data incompleteness. Second, stratifying results by subgroups such as sex or region could uncover heterogeneity in censoring effects, which is relevant in pricing and reserving. Third, comparing classical and non-parametric estimators with parametric survival models, such as Weibull or Gompertz, may enhance accuracy where model assumptions are plausible. Finally, quantifying how survival differences (e.g., in life expectancy) translate into financial impacts, such as annuity prices or reserves, would strengthen the actuarial relevance of this work.

X. Conclusion

This study demonstrates that classical life tables, while effective under complete observation, can yield upwardly biased survival estimates in the presence of censoring. We showed that both synthetic censoring-like adjustments to exposure and non-parametric survival methods yield systematically lower survival curves and life expectancy values.

Among the three methods considered, the actuarial estimator produced the lowest survival estimates due to its structural assumption of uniform censoring. The Kaplan-Meier estimator, by leveraging precise event and censoring times, offered a more accurate and interpretable estimate under realistic data conditions.

Although the absolute difference in life expectancy across methods may seem modest, around one year at birth, such discrepancies can significantly impact actuarial outcomes. For annuity products, underestimating survival leads to financial strain due to underestimated liabilities, while for life insurance, overestimation may affect competitive pricing and risk pooling.

We conclude that integrating survival analysis techniques, particularly those accounting for censoring, is essential for reliable longevity modeling. These tools not only improve the statistical fidelity of survival estimates but also support more robust decision-making in insurance, demography, and public health.

References

[1] Slud, Eric V. *Actuarial Mathematics and Life-Table Statistics*. University of Maryland, 2006. Available at https://www.math.umd.edu/~slud/s470/BookChaps/01Book.pdf.

[2] Missov, Tiago V., and James W. Vaupel. “How much can we trust life tables? Sensitivity of mortality measures to right-censoring.” *Palgrave Communications* 1 (2015): 1–6. https://doi.org/10.1057/palcomms.2015.49.

[3] HMD. Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Available at www.mortality.org (data downloaded on 22.5.2025).

[4] Štěpánek, L., Habarta, F., Malá, I., Marek, L. (2024). Non-parametric comparison of survival functions with censored data: A computational analysis of greedy and Monte Carlo approaches. *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 39, 725–730. http://dx.doi.org/10.15439/2024F223

[5] Kaplan, E. L., & Meier, P. "Nonparametric estimation from incomplete observations." (1958). *Journal of the American Statistical Association*, 53(282), 457–481. https://doi.org/10.1057/palcomms.2015.49.

[6] Štěpánek, L., Habarta, F., Malá, I., Marek, L. (2020). Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test. *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, 21, 453–460. http://dx.doi.org/10.15439/2020F198

[7] Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, 89(2), 232–238. https://doi.org/10.1038/sj.bjc.6601118

[8] Social Security Administration. *Definitions of Life Table Functions*. SSA, (2016). https://www.ssa.gov/oact/Downloadables/LifeTableDefinitions.pdf

[9] Office for National Statistics. *Guide to calculating national life tables*. ONS, (2019). https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/methodologies/guidetocalculatingnationallifetables

[10] Furman, Edward. *Actuarial Mathematics*. York University, (2020). https://edfurman.info.yorku.ca/files/2020/09/MATH3280-2020-lec2.pdf

[11] R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[12] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[13] Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

[14] Therneau T (2024). _A Package for Survival Analysis in R_. R package version 3.6-4, <https://CRAN.R-project.org/package=survival>.

[15] Terry M. Therneau, Patricia M. Grambsch (2000). _Modeling Survival Data: Extending the Cox Model_. Springer, New York. ISBN 0-387-98784-3.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 765–770
DOI: 10.15439/2025F7678
ISSN 2300-5963 ACSIS, Vol. 43

An RL Agent to Find Minimum Energy in a Tensegrity Representing a Cell

Mustafa Shah
*D*epartment of Electronics,
Quaid-e-Azam University,
Islamabad, Pakistan
mustafamohmand59@gmail.com

Arsenio Cutolo
*D*ept. of Structures for Engineering and Architecture
University of Napoli Federico II
Naples, Italy
arsenio.cutolo@unina.it

Muddasar Naeem
*U*niversità Giustino Fortunato
Benevento Italy
m.naeem@unifortunato.eu

Muhammad Waris
*D*epartment of Electronics,
Quaid-e-Azam University,
Islamabad, Pakistan
mwaris.22411012@ele.qau.edu.pk

Musarat Abbas
*D*epartment of Electronics,
Quaid-e-Azam University,
Islamabad, Pakistan
mabbas@qau.edu.pk

***Abstract*—Understanding the mechanical behavior of cells is a complex challenge at the crossroads of physics, biology, and engineering. The cytoskeleton which is a dynamic network of filaments which helps cells maintain shape, move, and respond to their environment. Tensegrity structures, made of interconnected tensile and compressive elements, offer a compelling way to model these internal forces. In this work, we use Reinforcement Learning (RL) to simulate and optimize cellular mechanics. We propose an RL framework where an agent learns to minimize the total mechanical energy of tensegrity-based cell models by adjusting node positions. We consider diverse shapes from simple shapes like lines and triangles, to more complex shapes like cell-like geometries. Our approach shows that RL can effectively model mechanical adaptations in cells and opens the door to intelligent, bio-inspired simulations. This work bridges biophysics, AI, and structural mechanics, offering new ways to predict and understand how cells respond to mechanical stress.**

***Index Terms*—Reinforcement Learning, Tensegrity, Cell Mechanics, Structural Optimization, Deep Learning, PPO, Cytoskeleton**

I. Introduction

CELLS are not static entities; they are dynamic, mechanically active systems that continuously adapt their internal architecture in response to various environmental and biological cues [14]. One of the most compelling frameworks for modeling the mechanical integrity of cells is the tensegrity model, which represents the cytoskeletal network as a balance of tensile and compressive forces [1], [4]. Tensegrity—short for "tensional integrity"—describes a structural system composed of isolated struts (under compression) suspended within a web of tensile elements (like cables or filaments). This principle reflects the way real cells maintain their shape, transmit forces, and react to mechanical stress.

The tensegrity-based approach allows one to capture the interplay between cytoskeletal elements such as actin filaments, microtubules, and intermediate filaments, providing a robust platform to simulate how cells respond to external mechanical stimuli and internal biochemical signals [3]. Beyond biology, tensegrity models have also found significant applications in the design and development of soft robotic systems and bioinspired structural materials that mimic cellular mechanics and adaptability [15].

With the rapid advancement of artificial intelligence and machine learning, particularly in the realm of reinforcement learning (RL), new opportunities have emerged to understand and control dynamic systems in ways that were previously unfeasible [8]. RL algorithms allow agents to interact with complex environments and learn control strategies through trial and error [13].

In this work, we propose a novel integration of RL with tensegrity physics to simulate a biologically inspired cell model. An RL agent is tasked with iteratively adjusting the positions of the nodes within a tensegrity structure to achieve a configuration with minimal mechanical energy, a proxy of structural equilibrium. Previous work in robotics has demonstrated the efficacy of such approaches in enabling tensegrity-based systems to learn locomotion and adaptive behaviors [6], [17]. By extending these concepts to cellular modeling, we aim to bridge the gap between mechanobiology and intelligent control systems.

The rest of the paper is organised as follow: Section II presents related work where we review the exisitng work in cell mechanics as well as use of RL to solve various problems. The section III is about our system model and methodology were we present in details our approach and tensegrity structure visualizations using RL. We outlined the results of experiments in section IV, followed by a discussion in section V and conclusion in section VI.

II. Related Work

In this section, we report the relevant literature to highlight the limitation of existing work and contribution of our work.

The work in [18] proposes an interpretable machine learning framework of cell mechanics from protein images and neural

networks (NN) for prediction of traction forces from a single focal adhesion protein field. Agnostic and physics-constrained approaches learn interpretable rules for prediction. A DCell is proposed in [20] that is a visible NN embedded in the hierarchical structure of 2,526 subsystems and encompasses an eukaryotic cell. DCell is able to simulates cellular growth with reasonable accuracy after training on millions of genotypes. The proposed setup provides a foundation to decode the drug resistance, genetics of disease, and synthetic life.

An interesting work is done in [22] to develop the safe RL approaches specially for environments where non rewarded states are also important. Authors considered different scenario to validate their approaches and use of safe RL methods could have useful contribution in cell mechanics. RL methods are also used in important medical issues to provide assistance to patients with disabilities [19], [21]. These works demonstrate a useful application of RL algorithms but the such formulation of tensegrity representing a cell is work to be done.

Moreover, Deep RL (DRL) is utilized in [9] to infer collective cell behaviours and cell–cell interactions in tissue morphogenesis from 3D time-lapse images. Hierarchical DRL is applied to investiage cell migrations from the images with an ubiquitous nuclear label. The hierarchical DRL method HDRL reveals a modular, multiphase organization of cell movement to Caenorhabditis elegans embryogenesis. A hybrid RL model is proposed in [12] to manage process control efficiently. A probabilistic knowledge graph mdoel is developed characterizing the science and risk-based understanding of quantifying inherent stochasticity and biomanufacturing process mechanisms.

In our work, we investigate how reinforcement learning can be applied to better understand the ways in which biological cells adapt their shapes and internal structures in response to mechanical forces. The proposed work formulate a tensegrity model representing a cell into an RL setup and then design a RL agent using proximal policy optimization to find minimum energy in a tensegrity representing a cell.

III. Methodology

This section presents the methodology used in our work. The first step is to model the cell as a tensegrity structure—a network of compressive elements (rods) and tensile elements (cables)—to capture how real cells maintain mechanical balance through tension and compression [1].

To explore this model computationally, we employ a Reinforcement Learning (RL) agent that learns to adjust node positions in the tensegrity structure to reduce the total mechanical energy, guiding it toward a stable, low-energy configuration. We use the Proximal Policy Optimization (PPO) algorithm due to its robustness in continuous control tasks [2].

By integrating biophysics, artificial intelligence, and structural mechanics, this approach contributes to the understanding of how cellular behavior emerges from physical constraints. Such methods also support the design of soft robotic systems that mimic cellular mechanics [5].

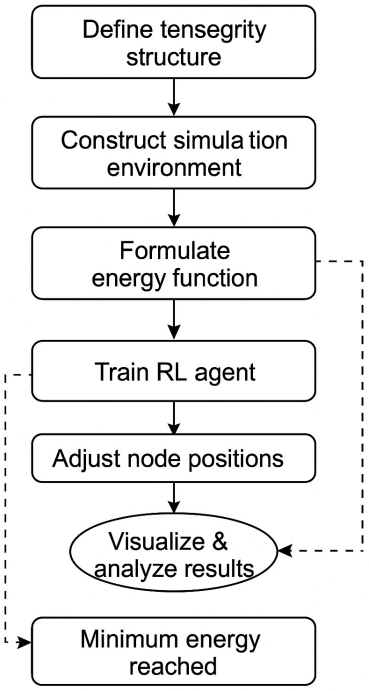


Fig. 1: Flowchart outlining the proposed approach.

The mechanical energy of the tensegrity structure is computed based on edge deformations relative to rest length:

$$E = \sum_{(i,j)\in \text{edges}} k_{ij} \cdot (||x_i - x_j|| - l_{ij}^0)^2$$

where x_i, x_j are 2D positions of nodes, l_{ij}^0 is the rest length, and k_{ij} is the stiffness constant (set to 1). This energy reflects internal stress.

1) Mechanical Energy Decomposition: To distinguish tension and compression, we decompose the total energy:

$$E_{\text{total}} = \sum_{(i,j)\in\mathcal{C}} k_{ij}^{(c)} \left(\|x_i - x_j\| - l_{ij}^0\right)^2 + \sum_{(i,j)\in\mathcal{T}} k_{ij}^{(t)} \left(\|x_i - x_j\| - l_{ij}^0\right)^2 \quad (1)$$

Here, $\mathcal{C}$ and $\mathcal{T}$ are sets of compressive and tensile elements, with different stiffness coefficients $k_{ij}^{(c)}$, $k_{ij}^{(t)}$, respectively.

2) Gradient of Energy with Respect to Node Position: Although PPO is model-free, internal forces can be approximated by the energy gradient:

$$\frac{\partial E}{\partial x_i} = \sum_{j\in\mathcal{N}(i)} 2k_{ij}(\|x_i - x_j\| - l_{ij}^0) \cdot \frac{(x_i - x_j)}{\|x_i - x_j\|} \quad (2)$$

This gradient represents the net force acting on node i, offering a physical interpretation of movement.

3) Proximal Policy Optimization (PPO) Loss: We use PPO's clipped surrogate loss [2]:

$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right] \tag{3}$$

with

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$

where $\hat{A}_t$ is the advantage estimate. This formulation ensures stable policy updates.

4) Energy Convergence Metric: To evaluate learning stability, we compute average energy change over the last N steps:

$$\Delta E = \frac{1}{N} \sum_{t=T-N+1}^{T} |E_t - E_{t-1}| \tag{4}$$

A small ΔE signals convergence to a stable, low-energy configuration.

5) RL Formulation:

- **State Space:** A flattened vector of 2D node coordinates representing the structure.
- **Action Space:** Small shifts $(\Delta x, \Delta y)$ applied to movable nodes.
- **Reward Function:** Negative of total mechanical energy.
- **Termination:** Fixed-length episodes without early stopping.

6) Implementation Steps:

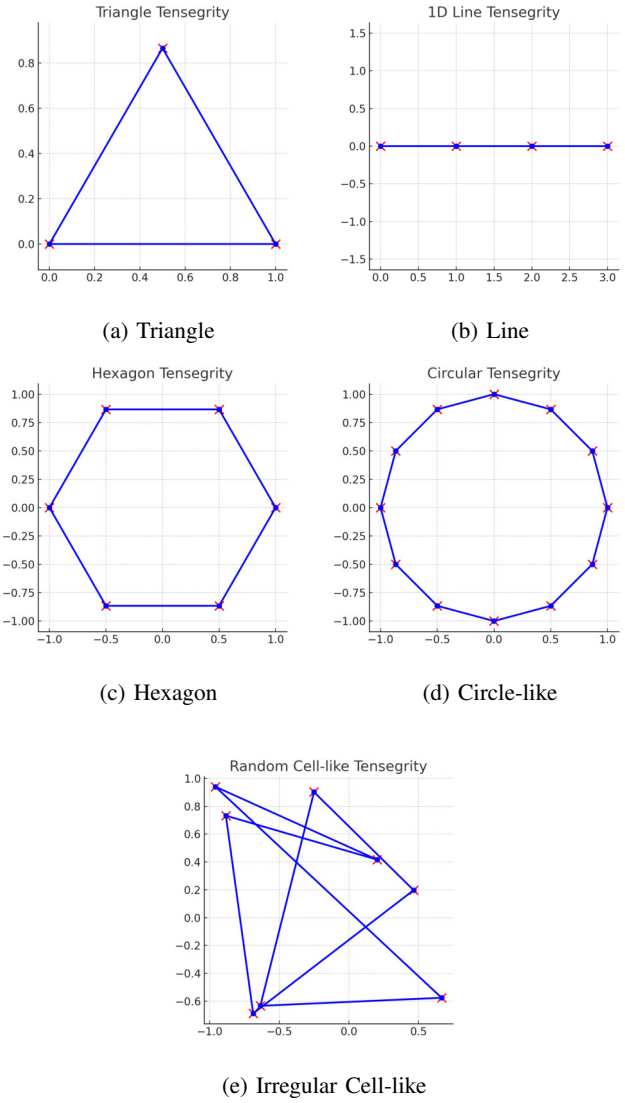
- Model the cell as a tensegrity structure of nodes and rods/cables [3].
- Develop a Python simulation environment to update node positions.
- Define an energy function to evaluate mechanical stress.
- Use PPO (via Stable-Baselines3) to train the agent [2].
- Train the agent to iteratively reduce system energy.
- Visualize learning and adaptations of the structure.
- Analyze biological relevance of learned behaviors [5].
- Explore applications in mechanobiology and soft robotics [15].

A. Reproducibility and Future Extensions

Each structure is defined by nodes and edges (cables/rods). A custom Python environment updates positions using actions from the agent. The state is a flattened list of node coordinates, while actions apply perturbations. The reward is based on internal potential energy. All components are modular and can be adapted for future extensions, including 3D structures and additional biophysical constraints.

B. Tensegrity Structures Visualization

We consider diverse tensegrity structures, from simple to complex, in our work as shown in Figure 2. Each configuration is simulated and visualized using Python, enabling us to observe how reinforcement learning (RL) guides the system toward more stable, low-energy states. These structures ranging from basic geometries to biologically inspired shapes demonstrate the adaptability of RL in modeling cellular mechanics [6], [7].



(a) Triangle (b) Line

(c) Hexagon (d) Circle-like

(e) Irregular Cell-like

Fig. 2: Various tensegrity structures used in the study.

Each of the structures illustrated in Figure 2 represents a unique scenario in our simulations. Figure 2a shows a simple triangular tensegrity, serving as a basic case to validate the RL agent's energy minimization capability in a constrained geometry. Figure 2b presents a linear configuration, ideal for examining the agent's behavior in elongated systems. The symmetric hexagonal form in Figure 2c introduces more complexity, resembling biological structures. Figure 2d illustrates a radially balanced, circular tensegrity, while Figure 2e presents an irregular, cell-like morphology—a step toward realistic cellular modeling.

Across all cases, the RL agent effectively minimized system energy, reinforcing the concept that even simple tensegrity

models can provide meaningful insights into cellular biomechanics.

C. Towards Realistic Cell-like Tensegrity Structures

We demonstrate a biologically inspired tensegrity system that reflects the intricate architecture of real cells, in Figure 3. This advanced model incorporates elements resembling cytoskeletal components such as actin filaments and microtubules, which are critical for maintaining cell shape and enabling mechanical responsiveness. Such structures have been central to tensegrity-based modeling frameworks in cellular biomechanics [4], [5]. By simulating these complex geometries, we move closer to capturing the adaptive, dynamic nature of living cells.

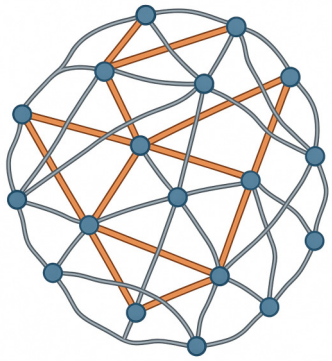


Fig. 3: A biologically inspired tensegrity model representing a complex cell.

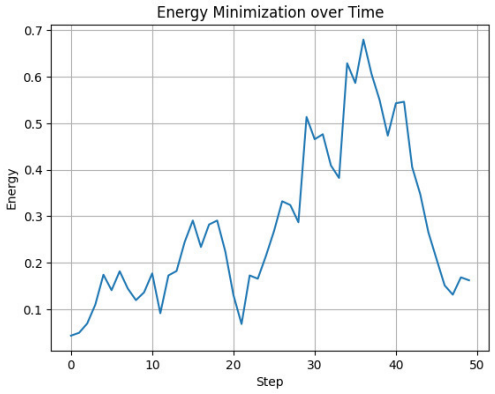
IV. Results

This section presents the experimental results that we performed while considering different structures.

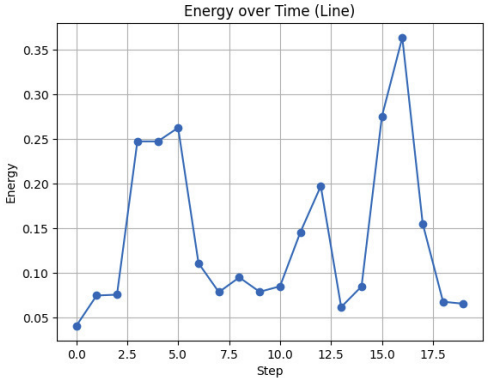
Our experiments began with basic triangular tensegrity structures as shown in Figure 2a, where the RL agent consistently achieved noticeable reductions in total mechanical energy over training episodes. These early successes demonstrated the agent's capacity to learn stable configurations, even in minimal geometric setups.

As we progressed to more complex shapes such as hexagonal (Figure 2c) and circular configurations (Figure 2d), the RL agent adapted well, maintaining structural coherence while reducing energy levels. Notably, the hexagonal structures, which resemble symmetrical cell geometries, converged faster toward low-energy states. The circular models, simulating soft boundary conditions like those found in real cell membranes, further confirmed the agent's ability to generalize across geometries.

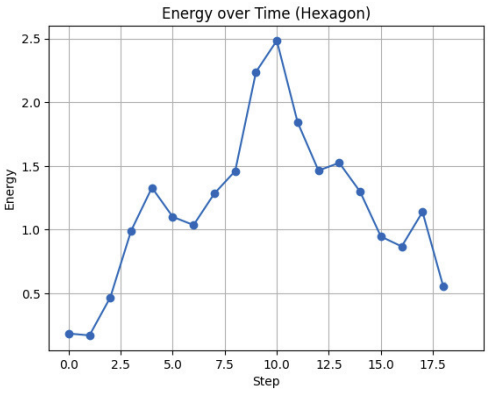
Visual analysis of the final tensegrity forms and energy plots validated the model's effectiveness. These findings are consistent with earlier research where RL-based policies achieved adaptive control and energy-efficient locomotion in soft tensegrity robots [7], [11].



(a) Triangle tensegrity



(b) Line tensegrity



(c) Hexagon tensegrity

Fig. 4: Energy of the tensegrity structures over training steps. (a) Triangle tensegrity (b) Line tensegrity (c) Hexagon tensegrity.

For tensegrity structures to train the RL agent, we employed the Proximal Policy Optimization (PPO) algorithm and used a reward function based on the system's mechanical energy. Specifically, the agent receives a negative reward equal to the sum of squared differences between the current and rest lengths of all links, effectively encouraging configurations that minimize internal tension and compression. The environment state is defined as a flat vector of 2D coordinates representing the positions of all nodes. This simple representation provides

the agent with sufficient spatial information to act. During early training, exploration is naturally handled through PPO's stochastic policy updates, allowing the agent to try unstable configurations and gradually converge toward energy-efficient ones.

Each experiment was trained over 100,000 timesteps, typically spanning 500 to 1,000 episodes depending on episode length. This duration provided the agent with enough interaction to consistently learn optimal or near-optimal configurations. For example, in the triangular and hexagonal tensegrities, the policy reliably minimized energy within the first 60,000 timesteps, while irregular cell-like shapes required more training to stabilize. We observed that once the energy starts decreasing steadily, subsequent steps fine-tune the positions rather than cause drastic shifts. Preliminary robustness checks indicate the learned policies remain effective when tested on slightly perturbed initial states.

While our current results are mostly qualitative, they offer clear and reproducible patterns of energy reduction across varied structures. In future work, we plan to include statistical metrics such as average final energy, convergence rate, and baseline comparisons with classical physics-based solvers. We are also extending our approach to 3D tensegrity models and exploring more expressive RL methods like Soft Actor-Critic (SAC) to improve learning in high-dimensional, nonlinear cellular geometries.

V. Discussion and Future Work

This study has demonstrated the feasibility of using reinforcement learning to manipulate tensegrity-based cell models toward stable and energy-efficient states. It provides a promising step in merging biophysical modeling with artificial intelligence, especially for systems that rely on structural tension and compression, like real cells. Despite its success in 2D structures, our framework opens the door to deeper biological relevance through future improvements as explained below:

- **Adding biological constraints**, such as anchor points representing extracellular matrix contacts or internal pressure mimicking cytoplasmic resistance [4].
- **Scaling to 3D tensegrity systems**, which can better replicate the spatial complexity of cytoskeletal networks found in living cells [10].
- **Incorporating viscoelasticity**, allowing simulation of how cells respond over time to mechanical stimuli and not just at equilibrium.
- **Integrating with real data**, such as microscopy-derived cell geometries, to improve biological validity and potentially guide experimental modeling [5].

Finally, we can summarize that as artificial intelligence continues to make significant contributions in the medical and healthcare domains, its integration into biological modeling is becoming increasingly relevant. Studies have shown how AI-driven approaches enhance diagnostic precision in areas such as skin cancer and COVID-19 detection [23], as well as in risk management [22]. In a similar spirit, our work demonstrates the potential of reinforcement learning for understanding complex biomechanical behavior at the cellular level. This further supports the growing trend of leveraging AI not only for diagnosis but also for predictive and mechanistic modeling of biological systems.

These enhancements would bring the relevant research community closer to a useful tool that can not only simulate the structure but also predict and interpret cellular mechanical behavior in health and disease.

In the future work, we will extend our work to more complex topologies and we will also consider including external constraints like cell-substrate interaction or intracellular pressure.

VI. Conclusion

In this work, we have presented a novel integration of tensegrity mechanics with reinforcement learning to explore how AI methods, specifically RL agent can autonomously discover stable, low-energy configurations of cell-like structures. By learning control strategies that minimize mechanical energy, the agent mimics cellular adaptation behaviors in silico. This framework lays the groundwork for more intelligent models of mechanobiology—capable of interpreting, simulating, and even designing cellular responses. With further development, it may contribute to advances in synthetic biology, bio-inspired robotics, and our theoretical understanding of cellular mechanics.

Acknowledgment

This research has been partially supported by the European Union - Next Generation EU through the Project of National Relevance "Innovative mathematical modeling for cell mechanics: global approach from micro-scale models to experimental validation integrated by reinforcement learning", financed by European Union-Next-GenerationEU-National Recovery and Resilience Plan-NRRP-M4C1-I 1.1, CALL PRIN 2022 PNRR D.D. 1409 14-09-2022—(Project code P2022MXCJ2, CUP F53D23010080001) granted by the Italian MUR.

References

[1] D. E. Ingber, "Cellular tensegrity: defining new rules of biological design that govern the cytoskeleton," *Journal of Cell Science*, vol. 104, pp. 613–627, 1993.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017.

[3] R. E. Skelton and M. C. de Oliveira, *Tensegrity Systems*, Springer, 2009.

[4] D. E. Ingber, "Tensegrity I. Cell structure and hierarchical systems biology," *Journal of Cell Science*, vol. 116, no. 7, pp. 1157–1173, 2003.

[5] Y. Chen, X. Wang, L. Yang, et al., "Machine Learning for Mechanobiology: From Cells to Tissues," *Annual Review of Biomedical Engineering*, vol. 25, pp. 377–406, 2023.

[6] C. Paul and M. Spröwitz, "Reinforcement learning for tensegrity robot locomotion," *Frontiers in Robotics and AI*, vol. 5, p. 77, 2018.

[7] K. Caluwaerts, J. Despraz, A. Işçen, A. Sabelhaus, V. SunSpiral, and R. D. Beer, "Design and control of compliant tensegrity robots through simulation and hardware validation," *Journal of The Royal Society Interface*, vol. 11, no. 98, 2014.

[8] Naeem, M., Coronato, A., Ullah, Z., Bashir, S. and Paragliola, G., 2022. Optimal user scheduling in multi antenna system using multi agent reinforcement learning. Sensors, 22(21), p.8278.

[9] Wang, Z., Xu, Y., Wang, D., Yang, J. and Bao, Z., 2022. Hierarchical deep reinforcement learning reveals a modular mechanism of cell movement. Nature machine intelligence, 4(1), pp.73-83.
[10] V. SunSpiral, A. Agogino, D. Atkinson, and A. Moored, "Tensegrity based probes for planetary exploration: Entry, descent and landing (EDL) and surface mobility analysis," *NASA Ames Research Center*, 2013.
[11] M. Ahn, J. Lee, S. Lee, and Y. Lee, "Policy gradient reinforcement learning for locomotion control of tensegrity robots," *International Conference on Ubiquitous Robots*, 2019.
[12] Zheng, H., Xie, W., Wang, K. and Li, Z., 2022. Opportunities of Hybrid Model-based Reinforcement Learning for Cell Therapy Manufacturing Process Control. arXiv preprint arXiv:2201.03116.
[13] D. Pathak, P. Agrawal, A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," *Proceedings of the 34th International Conference on Machine Learning*, 2017.
[14] Naeem, M., Fiorino, M., Addabbo, P. and Coronato, A., 2024. Integrating Artificial Intelligence Techniques in Cell Mechanics. Annals of Computer Science and Information Systems, 41, pp.111-116.
[15] M. Pagitz, M. Pellegrino, and A. Daynes, "Bio-inspired tensegrity soft robots," *Bioinspiration & Biomimetics*, vol. 7, no. 4, 2012.
[16] Fiorino, M., Naeem, M., Ciampi, M. and Coronato, A., 2024. Defining a metric-driven approach for learning hazardous situations. Technologies, 12(7), p.103.
[17] J. Kim, Y. Kim, and J. Ryu, "Bio-inspired reinforcement learning control for soft tensegrity robot locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3937–3944, 2022.
[18] Schmitt, M.S., Colen, J., Sala, S., Devany, J., Seetharaman, S., Caillier, A., Gardel, M.L., Oakes, P.W. and Vitelli, V., 2024. Machine learning interpretable models of cell mechanics from protein images. Cell, 187(2), pp.481-494.
[19] Ismail, A., Naeem, M., Khalid, U.B. and Abbas, M., 2025. Improving adherence to medication in an intelligent environment using reinforcement learning. Journal of Reliable Intelligent Environments, 11(1), pp.1-10.
[20] Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R. and Ideker, T., 2018. Using deep learning to model the hierarchical structure and function of a cell. Nature methods, 15(4), pp.290-298.
[21] Ismail, A., Naeem, M., Syed, M.H., Abbas, M. and Coronato, A., 2024. Advancing Patient Care with an Intelligent and Personalized Medication Engagement System. Information, 15(10), p.609.
[22] Fiorino, M., Naeem, M., Ciampi, M. and Coronato, A., 2024. Defining a metric-driven approach for learning hazardous situations. Technologies, 12(7), p.103.
[23] Qayyum H., Rizvi S.T.H., Naeem M., Khalid U.b., Abbas M., and Coronato A., "Enhancing Diagnostic Accuracy for Skin Cancer and COVID-19 Detection: A Comparative Study Using a Stacked Ensemble Method," *Technologies*, vol. 12, no. 9, p. 142, 2024. doi: https://doi.org/10.3390/technologies1209014210.3390/technologies12090142

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 771–776
DOI: 10.15439/2025F0792
ISSN 2300-5963 ACSIS, Vol. 43

AI Software Security for Smart Environment in a Dynamically Changing Knowledge Management Strategy

Anna Sołtysik-Piorunkiewicz
0000-0002-7935-1377
University of Economics in Katowice
ul. 1 Maja 50, 40-287
Katowice, Poland
Email: anna.soltysik-piorunkiewicz@uekat.pl

Małgorzata Pańkowska
0000-0001-8660-606X
University of Economics in Katowice
ul. 1 Maja 50,
40-287 Katowice, Poland
Email: malgorzata.pankowska@uekat.pl

Stanisław Stanek
0000-0001-9939-4350
Emeritus of General Tadeusz Kościuszko Military University of Land Forces,
Wrocław, Poland
Email: swstanek@gmail.com

Karsten Böhm
0000-0002-2950-7433
FH Kufstein Tirol
Andreas Hofer Str. 7
6330 Kufstein, Austria
Email: karsten.boehm@fh-kufstein.ac.at

Mariusz Żytniewski
0000-0003-2170-1191
University of Economics in Katowice
ul. 1 Maja 50,
40-287 Katowice, Poland
Email: zyto@ue.katowice.pl

***Abstract*—The theory of smart environments assumes a diversity of human-supporting solutions that interact with each other and are oriented towards applying artificial intelligence. Concepts such as digital twin, software agent, and machine learning can aid the creation of smart environments by supporting the design and simulation of device behaviors. This article aimed to conduct research addressing the question how does multi-agent software impact smart grid security and show the practical exemplification of supporting smart grid security with multi-agents system. The paper's research findings discussed the AI-based knowledge management model in the context of the GRAI framework usage, which stands for Generative, Receptive Artificial Intelligence, and presents a proposal of an updated model that takes into account recent developments regarding GenAI and its consequences for KM. Based on the conducted research, an algorithm for collecting information from devices was proposed and used in a simulation based on the Isolation Forest algorithm.**

***Index Terms*—Digital Twin, Smart Grid, Internet of Things, Internet of Intelligence, Immersive Technology, Smart Environment**

I. Introduction

THE development of the Internet of Things (IoT) [5], and Internet of Intelligence (IoI) theory [26] points to a wide range of smart solutions in everyday life. Concepts such as smart city, smart home, smart grids, and others allow for the application of information solutions in user support and provide the capability to process large amounts of data. In the context of smart systems, perceiving them as distributed architecture systems requires appropriate management, monitoring, and security. Smart environments can encompass a variety of smart solutions, including smart grids, smart homes, and smart cities. The melding of these solutions is an important area of research. Exploring ways to integrate them represents a key direction in the development of smart-environment theory.

Smart environments refer to the multiple application domains where real-time Internet of Things (IoT) data processing allows for improved decision-making. These are ecosystems in which humans, conventional computer systems, and IoT devices are combined through cyber-physical services. Smart environments represent any physical environment augmented with a collection of embedded systems elaborating heterogeneous data and interacting with people.

Typically, smart cities aim at providing applications such as smart transportation, smart grids, smart surveillance, to enhance urban living standards by IoT, AI, digital twins, augmented reality, and the metaverse. A smart city typically comprises smart homes, smart health, smart roads, smart parking, and smart people. Similarly, smart health spaces like smart hospitals, smart gyms, or smart retirement homes are aimed at continuous patient monitoring and therapy provision. The term smart home is used to describe homes with home applications, illumination, heating, air conditioning, television, computers, audio and video entertainment systems, security systems, and camera schemes. Smart building is to monitor the energy and water utilization, as well as emissions and waste products. A smart environment can be made up of users' wearable and handheld devices, and several smart personal spaces. Smart agriculture assists farmers in minimizing wastage and improving productivity because the systems support irrigation and crop field monitoring. A smart traffic light

system can monitor green light times and automatically switch lights to avoid traffic jams. Smart environments encompass a large set of technologies, including computer networking, wireless communication, computational infrastructures, sensor networks, and algorithm design (Fig 1).

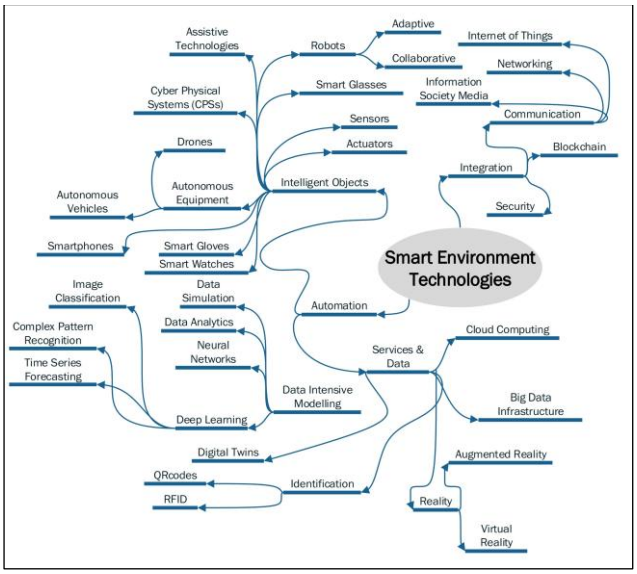


Fig 1. Smart environment technologies

The critical outcome of smart environment applications and execution is the generation of large-scale data sets. This large amount of data requires significant processing through artificial intelligence and machine learning algorithms. The term ‘smart’ was originally an acronym standing for “Self-Monitoring, Analysis and Reporting Technology”, but smart technology is now defined as much by its capacity to monitor and track its user activities, preferences, and regularities [29].

In this paper, the impact of multi-agents on the smart grid is detailed by the assumptions of smart grid security, adaptability, and user-friendliness factors. The following research questions were formulated:

- RQ1: How does multi-agent software impact smart grid security?
- RQ2: How to implement the multi-agent in a smart environment with current AI software?

The theoretical background was presented in Section 2. The research began with a preliminary study of the literature review, which is presented in Section 3. The empirical part of the study will demonstrate how a multi-agent system can be applied to enhance the security of a home grid network. The practical exemplification of the usage of multi-agent software was presented in Section 4.

The topic presented here is becoming increasingly actual and relevant in dynamically changing environments with its threats, which are related to, e.g., climate change, the energy crisis, or limited resources. The presented findings are an itemization of the factors that might be useful for further work, academic researchers, or business practitioners.

II. Theoretical Background

A. Multi-Agents software implementation

The areas of the Smart Environment theory, which include the smart grid theories discussed here, require the analysis of many aspects in the design and construction of modern computer systems. As the authors indicate [26], the Internet of Intelligence (IoI) requires defining the research area, the technologies in which the solution is being developed, its architecture, and identifying research gaps. Based on the research methodology, the studies undertaken in the article relate to Smart Grid/Energy in terms of application areas, Artificial Intelligence, Digital Twin, Immersive Technology in terms of considered technologies, and Modeling Intelligence and Security and Privacy in terms of the defined gaps in research discussed in the article.

From a technological perspective, the research considerations have been focused on multi-agent systems, which allow for the creation of intelligent and autonomous systems. These have the ability not only to automate processes but also to act on behalf of users within information systems in the context of Digital Twin (DT) theory. The use of multi-agent systems in the context of the Internet of Intelligence allows for a range of benefits such as the implementation of well-defined communication standards between agents and the creation of a system architecture within the framework of DT theory.

The features of multi-agent systems presented here suggest various possible scenarios for using these systems during the development of Internet of Intelligence systems.

B. Smart grid systems and security

One of the areas of implementing the IoI theory is smart grid systems. A smart grid (SG), also called smart electrical/power grid, intelligent grid, intelligrid, future grid, intergrid, or intragrid, is an enhancement of the 20th-century power grid. The traditional power grids are generally used to carry power from a few central generators to a large number of users or customers. In contrast, the SG uses two-way flows of electricity and information to create an automated and distributed advanced energy delivery network.

Due to the importance of the energy sector for the safety and uninterrupted operation of systems and devices, the search for solutions that safeguard such systems is a very important element in the development of IoI theory.

The Smart Grid term encapsulates the convergence of remote monitoring and control technologies with communication technologies, renewable generation, and analytics capabilities so that electric, natural gas, and water grid infrastructures can continuously provide information for operational and strategic management. The smart grids’ security challenges cover assurance of reliability and resilience of energy supply power, as well as safety, possible expansion and improvement plans, and efficiency of energy provision [2].

The ISO standards on smart grid security are as follows: ISO/IEC DIS 27019 Information technology Security techniques, Information security controls for the energy utility industry [38], ISO/IEC 30101:2014 Information technology,

Sensor networks: Sensor network and its interfaces for smart grid system [39].

C. Multi-Agent software in smart-grid system implementation

Within the framework of IoI theory, it becomes necessary to develop solutions that operate autonomously and are capable of collecting data, processing it, and making decisions regarding the behaviour of devices. One such solution is multi-agent systems.

Multi-agent systems (MASs) are particularly well suited for studying Smart Grid system implementation, such as a collection of at least two or more (artificial) agents, where the word agent can be succinctly described as an artificial entity that can act in its environment [1]. It is implicitly assumed that they are computational by design, that is, a decision about an action "can be broken down into primitive operations that can be implemented in a physical device" [7]. There exist several studies taking a MAS approach to studying community-based markets for electricity energy trading.

In the article [5], the authors presented a three-element architecture of the smart grid system, outlining elements such as Perception, Network, Service and Application Layer. In terms of possible applications of multi-agent systems, in addition to the levels presented, multi-agent systems allow for the division of the network layer into communication with IoT devices and communication between agents. The primary role of multi-agent systems is to provide control functions and assist communication in smart grid systems, particularly in microgrid systems [8]. Such a division of communication in smart grid systems relates to the security postulates of the systems, as discussed in the study [6]. Multi-agent systems, by their nature, resemble the concept of P2P systems used in smart grid theory [9], [11], but they have several additional features. These include the ability to define an agent's artificial intelligence mechanisms, dedicated communication standards that ensure the design patterns of the communication being conducted, agent mobility, and negotiation mechanisms. In the case of smart grids, the application of auction systems in the area of energy sharing [12], [13] can be facilitated by agents in the construction of auction systems, supported by agents. In the present article, the multi-agent system is examined within the context of a digital twin, wherein individual agent entities simulate the functioning of the network using real-world data. The simulation was developed based on the proposed algorithm and uses one of the machine learning models

III. Materials and Methods

This article used the literature analysis method. The literature review of the impact of multi-agents on the smart grid is detailed by the assumptions of smart grid security, adaptability, and user-friendliness factors. The data set of publications in the Web of Science was examined. The first query in the topic "Smart grid and security" referred to 5.888 publications discovered between 1900 and 2023. The topic was studied mostly by topics connected with the power grid, renewable energy resources, and the Internet of Things.

The previous studies and the state-of-the-art of smart grids concerning the surveys in different contexts. The most cited research about the security of smart grid by Fang et al. [3] showed the smart grid survey till 2011 as a new and improved power grid and showed the security features in IoT applications of smart grid, smart transportation, and smart cities, to demonstrate how fog/edge computing based IoT should be implemented in real world. Ellabban, Abu-Rub, and Blaabjerg showed the status, prospects, and enabling technology of renewable energy resources [4]. Lin et al. showed the architecture, enabling technologies, security and privacy, and applications of the Internet of Things [5]. Yan et al highlighted the major motivations, requirements, and challenges in smart grid communication infrastructures [6], as well as Mo et al. showed the vision of a secure smart grid infrastructure [33], and S. Sridhar et al. [34]. Deilami et al. presented the new real-time smart load management (RT-SLM) control strategy for coordinating the charging of multiple plug-in electric vehicles (PEVs) in a smart grid system performed by a utility energy distribution system [32]. The examples of the most cited studies were connected with smart-grid security issues, i.e., security technology for smart-grid networks, malicious data attacks on the smart grid with intrusion detection techniques for cyber-physical systems, and classifications and applications of physical layer security techniques for confidentiality.

A. The Literature Review Method

The research methodology was based on the literature review and research questions RQ1 and RQ2. The research questions were based on the research gap present in the previous research studies. The following query was formulated to verify the previous findings described as RQ1: Security aspects query: "Multi-Agents" AND "Smart Grid" AND "Security". Data collected from the Web of Science databases were used. The research findings were evaluated and discussed.

B. Research Results

The research results of RQ1 presented multidisciplinary publications in the field of security multi-agent systems for improving smart grid implementation. There are some features of multi-agent software usage and its role in smart grid systems that have been studied. Based on a review of the literature, the features are presented as research results based on literature review. Multi-agent software affects the security of smart grids, especially in relation to their expansion and modification. It cooperates with other solutions, which is necessary in the context of an intelligent network that is a system of systems with critical infrastructure. As the security solutions used in traditional IT networks are not sufficient for the system of systems [41], in the case of smart grids, such features of multi-agent systems as decentralized decision-making, adaptability, substitutability, communication, learning, mechanisms of trust, cooperation, resilience, and self-healing

gain in importance. As a result, multi-agent systems provide an adequate basis for the construction of security that is required in system of systems and critical infrastructure. Answering the RQ2 the example of studies influence of multi-agents software usage on a smart grid system was presented in Table I.

TABLE I.
INFLUENCE OF MULTI-AGENT SOFTWARE USAGE ON A SMART GRID SYSTEM

Smart Grid Feature	Multi-Agent Influence	Source
Adaptive robotics	Modular systems enable flexible adaptation in production. Human–robot interactions via mobile devices.	[15], [17]
Additive manufacturing	Possibility to individualize products in cost-effective, small batches.	[24]
Automated Home Energy Management	Secure Automated Home Energy Management in Multi-Agent Smart Grid Architecture	[19]
Cybersecurity	Methodologies allowing for automatic detection and response to cyberattacks; adaptive computational intelligent systems continuously evolving; a distributed multi-agent scheme to enhance the cybersecurity of smart power grids.	[28]
Decentralized Intrusion Prevention	Decentralized Intrusion Prevention (DIP) Against Coordinated Cyberattacks on Distribution Automation Systems	[30]
Demand forecasting	The MAS (Multi-Agent System) Model allows the VPP (Virtual Power Plant) to know in advance the amount of power and is based on artificial neural networks (ANNs)	[20]
Energy management	Smart grid management in prosumer communities, trading of energy	[13], [14], [16]
Demand response	Consumers can reduce their energy consumption through load curtailment, shift their energy consumption over time, or generate and store energy at certain times to provide the grid with more flexibility.	[21], [23]
User-Friendliness Environment	An effective modeling and simulation environment, in terms of generality, user-friendliness, modeling flexibility, interoperability with GIS datasets, and computational efficiency	[22]

The AI-Knowledge Management Process of implementing the agents in a smart environment could be discovered in future research. Nowadays, the RQ2 answer is based on the description of the AI-Knowledge Management model for agent tools implementation . The background is based on the revised SECI model, with the machine (AI agent) in the knowledge management process that can play an active or a passive role. The active role would generate an output or a response, while the passive role could be compared to listening and adapting/rebuilding the internal (knowledge) representation. Consequently, the four areas would each be split into a human perspective and a machine perspective, leading to eight fields of action in the new GRAI model [25]. More recent versions of LLM-based systems, such as OpenAI ChatGPT or Google Gemini, allow the human user to add additional relevant materials into the conversation, e.g., using the "memory function" of those systems. This context relates to the interaction field as it helps the machine to identify a more precise context with a domain-specific focus in the dialogue. This way, a general conversation can be leveraged toward a specific direction by the human user by providing externalized information to the machine. Another application use case that is emerging rather frequently here is the use of GenAI in retrieval-oriented task using Retrieval Augmented Generation (RAG), which combines an initial search query with specific document collections (externalized information) to derive a more specific search results when the original query is augmented by the generated content from the externalized information [26]. In contrast to the use case of information retrieval in enterprise-specific information sources (enterprise search), in the field of KM, the use of GenAI could introduce a bias in the results that originates from the foundation models used. However, these effects might be similar to the contextual integration of a human user and could be counteracted by carefully selecting and adopting the right foundation model [27]. In the case of a smart grid security project, the design of architecture, algorithms and simulation studies are becoming more important.

The example described in the paper illustrates RQ1 and RQ2 answers.

IV. PRACTICAL EXEMPLIFICATION

As detailed earlier, IoI and IoT underpin Smart Grid oversight and can be leveraged to strengthen the operation of these networks. Multi-agents could support Smart Grid Security, the robotisation of processes, and system adaptiveness. Authors' prior research into Smart Grid security devised a multi-agent system architecture to reinforce its security [18]. The resulting algorithm was calibrated to gauge the Grid's performance in terms of energy drawn by connected devices. In the example shown below, an algorithm for collecting information from IoT devices is described, the process in which it was used is presented, and a machine learning algorithm is employed to evaluate the collected data. The proposed algorithm was validated via a simulation conducted under the Digital Twin paradigm on the Jade platform. One possible application of IoT and Smart grid solutions is to support the monitoring of user and property security. The subsequent analysis concerns the Home Area Network (HAN) and Neighbourhood Area Network (NAN), where devices are examined locally, either within a single building or across multiple buildings in a given area. Home security systems, such as burglar alarms, sensors, can be supported by IoT devices and generate events related to the breach of defined rules. A problem encountered in IoT systems is the low performance of devices designed for minimal power consumption. In such solutions, it is crucial to use ML algorithms with low resource demands. For this experiment, we used the Isolation Forest algorithm. It is recognised for its low resource requirements.

The literature review [36], [37] allowed for an attempt at specifying an algorithm to support the operation of the network within the framework of IoI and IoT theory using a digital twin theory. To illustrate the conducted research experiment, elements of the architecture of a prototype using business process modelling and IoTDT-BPMN notation were developed. Fig. 2 presents the process of the system's operation

based on IoTDT-BPMN elements. Each pool constitutes a set of tasks executed by the individual devices used when experimenting.

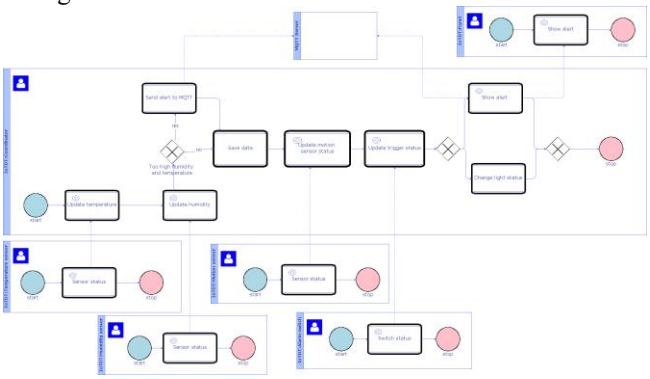


Fig 2. A process in the extended BPMN notation depicting the operation of the simulated system

The whole experiment was developed based on a JADE multi-agent system, which was used to perform the simulation. All the devices were registered in a multi-agent system, and agents' specific behaviours were defined. The implementation uses the IotDTBPMN2Jade algorithm available at GitHub [40]. Activation of the simulation led to the generation of communication between the agents, which resulted in the sending of a notification about a detected event and the development of a knowledge base of alert events. Data from the available sensors was gathered using the developed solution after the simulation phase. The data shown in the figure were collected from an IoT sensor connected to an MQTT server between 26 May 2025 and 8 May 2025. Two pronounced "spikes" in humidity levels can be observed, corresponding to emergency shutdowns of the ventilation system on 2 May 2025 and 4 May 2025. After this initial stage, the sensor readings were used to develop a model based on the Isolation Forest algorithm. One requirement of the algorithm is that the input data must be supplied at regular time intervals. In this project, the raw sensor readings were resampled into five-minute intervals. Since the sensor conserves power by reporting only when its state changes, any missing values were filled using linear interpolation. This ensured that every time slot was represented in the training set. The conducted experiments showed that, with the number of trees set to n_estimators=200 and a contamination level of 0.2 %, the model produced anomaly predictions that closely matched the real-world events illustrated in Fig. 3.

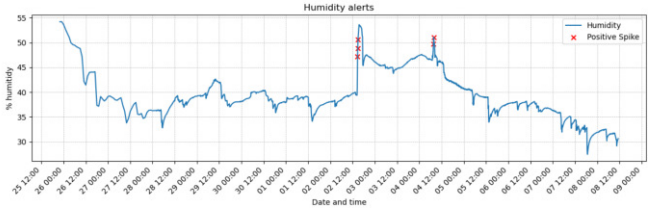


Fig 3. Outcomes of using the ML algorithm (humidity alerts)

The model flagged 5 points as anomalies from the initial 3,617 samples, which is about 0.14%. Increasing the contamination level to 0.5% raised the number of alerts to 12, while lowering it to 0.05% caused it to miss all anomalies. In the subsequent model test, a three-day dataset was prepared containing two events characterised by sudden spikes in humidity levels. The model's results on this simulated data are shown in Fig. 4. The model correctly detected the humidity spikes, indicating that the recommended 50% threshold was exceeded.

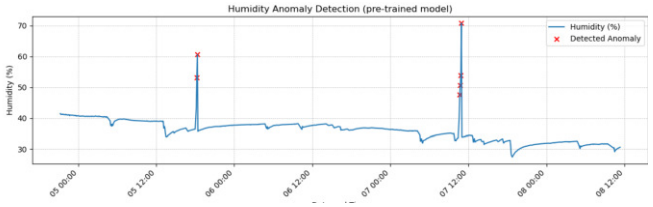


Fig. 4. Humidity anomaly detection

V. Conclusion

The paper showed the studies of theoretical and practical exemplification of multi-agents software security due to RQ1 and RQ2. The proposed research methodology was based on literature and focused on an IoT alarm-monitoring algorithm for smart homes and smart grids, which was described using an expanded BPMN notation and built on the idea of a digital twin. The algorithm specifies how agents evaluate the need to trigger alarms for particular model parameters and may serve to simulate the behaviour of IoT devices within Smart Grid networks. The algorithm's mechanism for gathering parameters from IoT devices offers further extension through the use of a SIEM-Connector concept. This idea involves a service provider continuously monitoring a large number of devices across several networks. Such integration allows real-time detection of threats reported by multiple domestic networks simultaneously. The use of the Isolation Forest algorithm facilitated the model's training and its effective identification of situations that could threaten the system's operational continuity. This was verified by the conducted simulation. The future research should be concerned with the implementation and usage of the AI-based Knowledge Management (AI-KM) models with GenAI and the internalization interaction between the machine agents. Internalization relates strongly to the creation of internal models (or representations) of the knowledge of the outer world of the agent, reflecting the views and beliefs of the agent. The assistance by GenAI might both help the internalization processes of the human agent, but might also be beneficial to proactively build or adapt (digital) representations for the machine agent.

References

[1] Q. Tang, F. R. Yu, R. Xie, A. Boukerche, T. Huang and Y. Liu, "Internet of Intelligence: A Survey on the Enabling Technologies, Applications, and Challenges" in *IEEE Communications Surveys & Tutorials*, vol. 24, no. 3, pp. 1394-1434, third quarter 2022, doi: 10.1109/COMST.2022.3175453

[2] A. Okino Otuoze, A. Masood, R. Larik "Smart grids security challenges: Classification by sources of threats". Journal of Electrical Systems and Information Technology 5(2018) 468-482, https://doi.org/10.1016/j.jesit.2018.01.001

[3] X. Fang, S. Misra, G. Xue and D. Yang, "Smart Grid — The New and Improved Power Grid: A Survey," in *IEEE Communications* Surveys & Tutorials, vol. 14, no. 4, pp. 944-980, Fourth Quarter 2012, doi: 10.1109/SURV.2011.101911.00087

[4] O. Ellabban, H. Abu-Rub, F. Blaabjerg "Renewable energy resources: Current status, future prospects and their enabling technology", Renewable and Sustainable Energy Reviews, vol.39, 748-764, 2014, 10.1016/j.rser.2014.07.113.

[5] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang and W. Zhao, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications," in IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125-1142, Oct. 2017, doi: 10.1109/JIOT.2017.2683200

[6] Y. Yan, Y. Qian, H. Sharif and D. Tipper, "A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges," in IEEE Communications Surveys & Tutorials, vol. 15, no. 1, pp. 5-20, First Quarter 2013, doi: 10.1109/SURV.2012.021312.00034

[7] D. L. Poole, A. K. Mackworth "Artificial intelligence: foundations of computational agents", Second ed., USA, Cambridge University Press; 2017.

[8] M. L Tuballa, M. L Abundo "A review of the development of Smart Grid technologies", Renewable and Sustainable Energy Reviews, Volume 59, 2016, pp. 710-725, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2016.01.011.

[9] J. Kim and Y. Dvorkin, "A P2P-Dominant Distribution System Architecture," in IEEE Transactions on Power Systems, vol. 35, no. 4, pp. 2716-2725, July 2020, doi: 10.1109/TPWRS.2019.2961330.

[10] F. Moret and P. Pinson, "Energy Collectives: A Community and Fairness Based Approach to Future Electricity Markets," in IEEE Transactions on Power Systems, vol. 34, no. 5, pp. 3994-4004, Sept. 2019, doi: 10.1109/TPWRS.2018.2808961.

[11] T. Morstyn and M. D. McCulloch, "Multiclass Energy Management for Peer-to-Peer Energy Trading Driven by Prosumer Preferences," in IEEE Transactions on Power Systems, vol. 34, no. 5, pp. 4005-4014, Sept. 2019, doi: 10.1109/TPWRS.2018.2834472.

[12] P. Olivella-Rosell, G. Viñals-Canal, A. Sumper, R. Villafafila-Robles, B. A. Bremdal, I. Ilieva, S. Ø. Ottesen "Day-ahead micro-market design for distributed energy resources" In: 2016 IEEE International Energy Conference. 2016, p. 1–6. http://dx.doi.org/10.1109/ENERGYCON.2016.7513961.

[13] W. Tushar, B. Chai, C. Yuen, S. Huang, D. B. Smith, H. V. Poor, Z. Yang "Energy Storage Sharing in Smart Grid: A Modified Auction-Based Approach," in IEEE Transactions on Smart Grid, vol. 7, no. 3, pp. 1462-1475, May 2016, doi: 10.1109/TSG.2015.251226.

[14] R. Verschae, T. Kato, T. Matsuyama "Energy management in prosumer communities: A coordinated approach", Energies 2016;9(7). http://dx.doi.org/10.3390/en9070562, URL https://www.mdpi.com/1996-1073/9/7/562.

[15] A. Billard, S. Mirrazavi, N. Figueroa "Learning for Adaptive and Reactive Robot Control. A Dynamical Systems Approach", MIT Press, 2022.

[16] R. Jing, MN. Xie, FX. Wang, LX Chen. "Fair P2P energy trading between residential and commercial multi-energy systems enabling integrated demand-side management", Appl Energy 2020;262:114551.

[17] S. Nolfi "Behavioral and cognitive robotics: An adaptive perspective". Roma, Italy: Institute of Cognitive Sciences and Technologies, National Research Council CNR-ISTC (2021).

[18] T. Kisielewicz, S. Stanek S., M. Żytniewski "A Multi-Agent Adaptive Architecture for Smart-Grid-Intrusion Detection and Prevention", Energis, 2022 15, 4726. https://doi.org/10.3390/en15134726.

[19] C. P. Nguyen, A. J. Flueck "Agent based restoration with distributed energy storage support in smart grids", IEEE Trans. Smart Grid, vol. 3, no. 2, pp. 1029–1038, Jun. (2012).

[20] P. Vytelingum, T. D. Voice, S. D. Ramchurn, A. Rogers, N. R. Jennings "Agent-based micro-storage management for the smart grid" in Proc. 9th Int. Conf. Auto. Agents Multiagent Syst., 2010, pp. 39–46.

[21] G. H. Merabet et al., Applications of multi-agent systems in smart grids: A survey, in Proc. Int. Conf. Multimedia Comput. Syst. (ICMCS), pp. 1088–1094, (2024)

[22] L. Hernández-Callejo et al. A Multi-Agent-System Architecture for Smart Grid Management and Forecasting of Energy Demand in Virtual Power Plants, January 2013, IEEE Communications Magazine 51(1), DOI: 10.1109/MCOM.2013.6400446.

[23] J. R. Vázquez-Canteli, Z. Nagy "Reinforcement learning for demand response: A review of algorithms and modeling techniques", Applied Energy, Volume 235, 2019, pp. 1072-1089, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2018.11.002.

[24] I. Blecic, A. Cecchini, G.A. Trunfio "A Software Infrastructure for Multi-agent Geosimulation Applications" In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, M.L. (eds) Computational Science and Its Applications – ICCSA 2008. ICCSA 2008. Lecture Notes in Computer Science, vol 5072, 2008, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69839-5_28.

[25] K. Böhm, S. Durst "Knowledge management in the age of generative artificial intelligence – from SECI to GRAI", VINE Journal of Information and Knowledge Management Systems Emerald Publishing Limited 2059-5891, 2025, DOI 10.1108/VJIKMS-10-2024-0357.

[26] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu "Evaluation of retrieval-augmented generation: a survey", arXiv, 2024, doi: 10.48550/arXiv.2405.07437.

[27] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui "Retrieval-augmented generation for AI-generated content: a survey", arXiv, 2024,. doi: 10.48550/arXiv.2402.19473.

[28] W. Ben, H. Ines "Preventive mental health care: A complex systems framework environments", Cognitive Systems Research, 84, 101199, 2024, http://doi.org/ 10.1016/j.cogsys.2023.101199

[29] Y. Mo, T. H. -J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, B. Sinopoli "Cyber–Physical Security of a Smart Grid Infrastructure", in Proceedings of the IEEE, vol. 100, no. 1, pp. 195-209, Jan. 2012, doi: 10.1109/JPROC.2011.2161428.

[30] N. Z. Aitzhan, D. Svetinovic "Security and Privacy in Decentralized Energy Trading Through Multi-Signatures, Blockchain and Anonymous Messaging Streams" in IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 5, pp. 840-852, 1 Sept.-Oct. 2018, doi: 10.1109/TDSC.2016.2616861.

[31] J. Appiah-Kubi, C. C. Liu "Decentralized Intrusion Prevention (DIP) Against Co-Ordinated Cyberattacks on Distribution Automation Systems", in IEEE Open Access Journal of Power and Energy, vol. 7, pp. 389-402, 2020, doi: 10.1109/OAJPE.2020.3029805.

[32] S. Deilami, A. S. Masoum, P. S. Moses and M. A. S. Masoum, Real-Time Coordination of Plug-In Electric Vehicle Charging in Smart Grids to Minimize Power Losses and Improve Voltage Profile, in IEEE Transactions on Smart Grid, vol. 2, no. 3, pp. 456-467, Sept. 2011, doi: 10.1109/TSG.2011.2159816.

[33] Y. Mo, T. H. -J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, B. Sinopoli "Cyber–Physical Security of a Smart Grid Infrastructure", in Proceedings of the IEEE, vol. 100, no. 1, pp. 195-209, Jan. 2012, doi: 10.1109/JPROC.2011.2161428.

[34] S. Sridhar, A. Hahn, M. Govindarasu "Cyber–Physical System Security for the Electric Power Grid". in Proceedings of the IEEE, vol. 100, no. 1, pp. 210-224, Jan. 2012, doi: 10.1109/JPROC.2011.2165269.

[35] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, X. Yang "A Survey on the Edge Computing for the Internet of Things," in IEEE Access, vol. 6, pp. 6900-6919, 2018, doi: 10.1109/ACCESS.2017.2778504.

[36] B. Karaduman, B. T. Tezel, G. Kardas, M. Challenger "DSML4JaCaMo: A Modelling tool for Multi-agent Programming with JaCaMo", Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS, Vol. 39, pages 637–642 (2024)

[37] V. Carchiolo, G. Catalano, M. Malgeri, C. Pellegrino, G. Platania, N. Trapani "BPM Tools for Asset Management in Renewable Energy Power Plants", Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 18, pages 645–649 (2019)

[38] ISO/IEC DIS 27019, https://www.iso.org/standard/85056.html

[39] ISO/IEC 30101:2014, https://www.iso.org/standard/53221.html

[40] GitHub, https://github.com/Anakainosis/IotDTBPMN2Jade

[41] F. Aloul, A.R. Al-Ali, R. Al-Dalky, M. Al-Mardini, El-Hajj, W. Smart Grid Security: Threats, Vulnerabilities and Solutions. International Journal of Smart Grid and Clean Energy. 1. 1-6. 10.12720/sgce.1.1.1-6. (2012).

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 777–781
DOI: 10.15439/2025F4849
ISSN 2300-5963 ACSIS, Vol. 43

A New Optimization Method for Evaluating Sobol' Sensitivity Indices

Venelin Todorov*†, Velichka Traneva§, Stoyan Tranev§,
Mihai Petrov§, Slavi Georgiev‡¶, Fatima Sapundzhi‖
*Department of Parallel Algorithms and Machine Learning with a Laboratory in Neurotechnologies,
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
†Centre of Excellence in Informatics and Information and Communication Technologies,
Acad. G. Bonchev Str. Bl. 25A, 1113 Sofia, Bulgaria
‡Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
§Department of Mathematics, Informatics and Physics, Prof. Dr. Asen Zlatarov University – Bourgas,
1 Prof. Yakim Yakimov Blvd., 8010 Bourgas, Bulgaria
¶Department of Applied Mathematics and Statistics, Angel Kanchev University of Ruse,
8 Studentska Str., 7004 Ruse, Bulgaria
‖Department of Communication and computer engineering, Faculty of Engineering,
South-West University "Neofit Rilski", 2700, Blagoevgrad, Bulgaria
Email: venelin@parallel.bas.bg, veleka13@gmail.com, tranev@abv.bg,
mihpetrov@abv.bg, sggeorgiev@math.bas.bg, sggeorgiev@uni-ruse.bg, sapundzhi@swu.bg

***Abstract*—This paper presents an optimization method based on a particular polynomial lattice rule with interlaced factor two for estimating sensitivity indices in global sensitivity analysis, focusing on total, first-order and second-order Sobol indices. A comparison with one of the best available methods the Modified Sobol Sequence and component by component construction polynomial lattice rule have been done. Relative errors for key output quantities are analyzed and compared. Our results show that the proposed optimization method consistently outperforms other methods in accurately estimating both first-order and total-order sensitivity indices, especially for parameters with smaller effects. These findings highlight the strengths and limitations of each method, providing guidance for selecting appropriate stochastic sampling strategies in computational sensitivity analysis.**

I. Introduction

SENSITIVITY analysis (SA) [12], [14], [17], [21] is a fundamental technique in modeling and simulation that investigates how variations in input parameters influence the output of a given system. By identifying which inputs contribute most to uncertainty in the model results, SA provides valuable insights into the robustness and reliability of mathematical models. This approach is especially crucial in complex environmental, financial, and engineering systems, where numerous variables interact in nonlinear ways. SA not only enhances model transparency but also supports decision-making by highlighting the key drivers of system behavior. Quasi Monte Carlo methods (QMC) methods are powerful technique for performing SA [14], [21].

Venelin Todorov was partially supported by the Centre of Excellence in Informatics and ICT under the Grant No BG16RFPR002-1.014-0018, financed by the Research, Innovation and Digitalization for Smart Transformation Programme 2021-2027 and co-financed by the European Union and by BNSF under Projects KP-06-N62/6 "Machine learning through physics-informed neural networks".

The input data for SA has been obtained during runs of a large-scale mathematical model for remote transport of air pollutants - **Uni**fied **D**anish **E**ulerian **M**odel (UNI-DEM) [11], [16], [26]. UNI-DEM is described mathematically [23], [24], [25]) by the following system of PDEs:

$$\frac{\partial c_s}{\partial t} = -\frac{\partial (uc_s)}{\partial x} - \frac{\partial (vc_s)}{\partial y} - \frac{\partial (wc_s)}{\partial z} + \\ + \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s}{\partial z}\right) + \\ + E_s + Q_s(c_1, c_2, \ldots, c_q) - (k_{1s} + k_{2s})c_s, \quad s = 1, 2, \ldots, q.$$

The number of studied pollutants by UNI-DEM determines the number q of equations in the system. It follows the other dimensions of the model:

- c_s - pollutant concentrations,
- u, v, w - wind components along the coordinate axes,
- K_x, K_y, K_z - diffusion coefficients,
- E_s - space emissions,
- k_{1s}, k_{2s} - dry and wet deposit coefficients, respectively ($s = 1, \ldots, q$),
- $Q_s(c_1, c_2, \ldots, c_q)$ - nonlinear functions describing chemical reactions between pollutants.

It is assumed that the mathematical model can be presented by a model function

$$\mathrm{u} = f(\mathrm{x}), \quad \text{where} \quad \mathrm{x} = (x_1, x_2, \ldots, x_s) \in U^s \equiv [0;1]^s \tag{1}$$

is a vector of input parameters with a joint **p**robability **d**ensity **f**unction (p.d.f.) $p(\mathrm{x}) = p(x_1, \ldots, x_s)$.

The concept of Sobol approach is based on a decomposition of an integrable model function f [20]:

$$f(\mathrm{x}) = f_0 + \sum_{\nu=1}^{s} \sum_{l_1<\ldots<l_\nu} f_{l_1\ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu}), \qquad (2)$$

where f_0 is a constant. The representation (2) is referred to as the ANOVA-representation of the model function $f(\mathrm{x})$ if each term satisfy [20]:

$$\int_0^1 f_{l_1\ldots l_\nu}(x_{l_1}, x_{l_2}, \ldots, x_{l_\nu})\mathrm{d}x_{l_k} = 0,\ 1 \le k \le \nu, \nu = 1, \ldots, s.$$

The quantities

$$\mathbf{D} = \int_{U^s} f^2(\mathrm{x})\mathrm{d}\mathrm{x} - f_0^2,\ \mathbf{D}_{l_1\ \ldots\ l_\nu} = \int f^2_{l_1\ \ldots\ l_\nu}\mathrm{d}x_{l_1}\ldots\mathrm{d}x_{l_\nu} \qquad (3)$$

are the so-called total and partial variances, respectively. A similar decomposition holds for the total variance that is represented by the corresponding partial variances: $\mathbf{D} = \sum_{\nu=1}^{s}\sum_{l_1<\ldots<l_\nu}\mathbf{D}_{l_1\ldots l_\nu}$. The main sensitivity measures following the Sobol approach are the so-called Sobol global sensitivity indices [20], [17] defined by

$$S_{l_1\ \ldots\ l_\nu} = \frac{\mathbf{D}_{l_1\ \ldots\ l_\nu}}{\mathbf{D}}, \quad \nu \in \{1, \ldots, s\}. \qquad (4)$$

and the **t**otal **s**ensitivity **i**ndex (TSI) of an input parameter $x_i, i \in \{1, \ldots, s\}$ defined by [20], [17]:

$$S_i^{tot} = S_i + \sum_{l_1 \ne i} S_{il_1} + \sum_{l_1, l_2 \ne i, l_1<l_2} S_{il_1l_2} + \ldots + S_{il_1\ldots l_{s-1}}, \qquad (5)$$

where S_i is called *the main effect (first-order sensitivity index)* of x_i and $S_{il_1\ldots l_{j-1}}$ is the j-th order sensitivity index. The higher-order terms describe the interaction effects between the unknown input parameters $x_{i_1}, \ldots, x_{i_\nu}, \nu \in \{2, \ldots, s\}$ on the output variance. This implies that performing global sensitivity analysis from a mathematical standpoint involves calculating total sensitivity indices (see Equation 5), which—according to the formulations in Equations 3–4—reduces to the evaluation of multidimensional integrals.

II. The Stochastic Approaches

Let us consider the following task for multidimensional integration:

$$S(f) := I = \int_{U^s} f(x)dx.$$

The first algorithm that we are going to use is the modified Sobol sequence based on procedure of shaking.

We will use a method that looks like the stratified symmetrized Monte Carlo [19]. For our Monte Carlo algorithm based on modified Sobol sequence (MCA-MSS-2-S), the original domain of integration is divided into m^s disjoint subdomains with equal volumes $\mathbf{K}_i^s, i = 1, \ldots, m^s$, where m is the number of subintervals used for the partition in each dimension. Here two pseudorandom points are generated. The first ξ_i is generated uniformly distributed inside the subdomain $\mathbf{K}_i^s$ and second $\xi_i^{'}$ is computed to be symmetric to ξ_i according to the central point s_i in $\mathbf{K}_i^s$. The concept in two-dimensional case for MCA-MSS-2-S is illustrated on Figure 1.

The value of the integral can be approximated [6]:

$$I(f) \approx \frac{1}{2m^s}\sum_{i=1}^{m^s}\left[f(\xi_i) + f(\xi_i^{'})\right].$$

It is proved that the algorithm MCA-MSS-2-S has an optimal rate of convergence $(n^{-\frac{1}{2}-\frac{2}{s}})$ for the class of continuous functions with continuous first derivatives and bounded second derivatives [7]. Up to now this method gives one of the best results for estimating sensitivity indices [6].

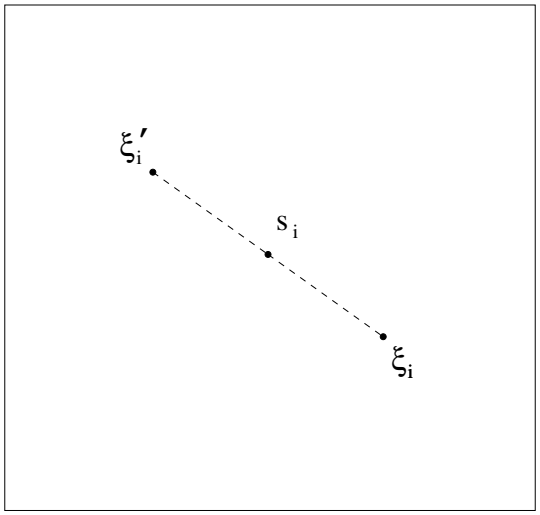


Fig. 1. Generation of a pseudorandom point $\xi_i(\xi_i') \in \mathbf{E}_i^2$.

Now we will use the so-called lattice rules. To introduce rank-1 lattice rule we will use the following formula [22]:

$$\mathbf{x}_k = \left\{\frac{k}{N}\mathbf{z}\right\},\ k = 1, \ldots, N, \qquad (6)$$

where N is an integer, $N \ge 2$, $\mathbf{z} = (z_1, z_2, \ldots z_s)$ is the generating vector and $\{z\}$ denotes the fractional part of z. For the definition of the $E_s^\alpha(c)$ and $P_\alpha(z, N)$ see [22].

In 1959 Bahvalov [1] proved that there exists an optimal choice of the generating vector $\mathbf{z}$:

$$\left|\frac{1}{N}\sum_{k=1}^{N} f\left(\left\{\frac{k}{N}\mathbf{z}\right\}\right) - \int\limits_{[0,1)^s} f(u)du\right| \le cu(s,\alpha)\frac{(\log N)^{\beta(s,\alpha)}}{N^\alpha}, \qquad (7)$$

for the function $f \in E_s^\alpha(c)$, $\alpha > 1$ and $u(s,\alpha), \beta(s,\alpha)$ do not depend on N.

The first generating vector in construction of our lattice rule is the generalized Fibonacci numbers of the corresponding dimension and the method will be called FIBO. FIBO will use the following generating vector [22]:

$$\mathbf{z} = (1, F_n^{(s)}(2), \ldots, F_n^{(s)}(s)), \qquad (8)$$

where we use that $F_n^{(s)}(j) := F_{n+j-1}^{(s)} - \sum_{i=0}^{j-2} F_{n+i}^{(s)}$ and $F_{n+l}^{(s)}$ ($l = 0, \ldots, j-1, j$ is an integer, $2 \leq j \leq s$) is the term of the s-dimensional Fibonacci sequence [22].

Now we will consider a special lattice rule based on the component by component construction method [2], [3], [18]. An optimal generating vector based on the rank-1 lattice rules with prime number of points is obtained using the component by component construction method. More specifically, let us initially set $z_1 := 1$. Then, assuming z_1 remains fixed, and $z_2 \in U^N := z \in \mathbb{N} : 1 \leq z \leq N-1,\ \gcd(z, N) = 1$ is chosen in such a way that a predefined error criterion [5], [15] is minimized in two dimensions. Then, iteratively for $i = 3, \ldots, s$, z_i is selected from U^N in such a manner that it minimizes the predefined error criterion in i dimensions. At the first step of the algorithm s dimensional optimal generating vector $\mathbf{z} = (z_1, z_2, \ldots z_s)$ is generated by the fast component by component construction method described above. On the second step of the algorithm we generate the points of lattice rule by formula $\mathbf{x}_k = \left\{ \frac{k}{N}\mathbf{z} \right\},\ k = 1, \ldots, N$. And at the third and last step of the algorithm an approximate value I_N of the multidimensional integral is evaluated by the formula:

$$I_N = \frac{1}{N} \sum_{k=1}^{N} f\left(\left\{ \frac{k}{N}\mathbf{z} \right\} \right).$$

The steps of working for the method are given on the flowchart on Fig. 2.

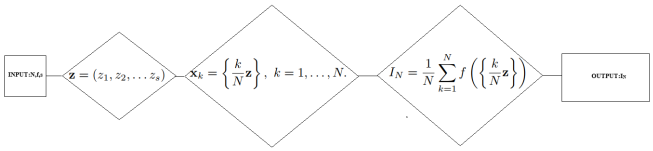


Fig. 2. The flowchart of the algorithm

Now we will continue with our optimization method based on polynomial lattice rule with special generating matrices never used for this computational task before. Let b be prime, let p be a polynomial and $deg\ p = m$ with coefficients in Z_b, and let $q_1, \ldots q_s$ be polynomials and $deg(q_1, \ldots, q_s) \leq m-1$ with coefficients in Z_b. Then we will choose C_j, the j^{th} generating matrix.

Let for $u_1, u_2, \cdots \in Z_b$:

$$\frac{q_j(x)}{p(x)} = \frac{u_1}{x} + \frac{u_2}{x^2} + \ldots.$$

Then set

$$\mathbf{C_j} = \begin{pmatrix} u_1 & u_2 & u_3 & \ldots & u_m \\ u_2 & u_3 & \ldots & \ldots & u_{m+1} \\ u_3 & \ldots & \ldots & \ldots & u_{m+2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ u_m & u_{m+1} & u_{m+2} & \ldots & u_{2m-1} \end{pmatrix} \in Z_b^{m \times m}. \tag{9}$$

The digital net [4] with $C_1, \ldots, C_s$ is named a polynomial lattice point set (PLPS), and a QMC rule using a polynomial lattice point set is named a polynomial lattice rule, where p is the modulus, and $(q_1, \ldots, q_s)$ is the generating vector of polynomials [13].

Let $m, s, d \in \mathbb{N}$ be defined with $d > 1$. Let $p \in Z_b$ be defined as $\deg(p) = m$ and let $q = (q_1, \ldots, q_{ds}) \in (Z_b)^{ds}$. An interlaced polynomial lattice rule of order d is actually a point set composed by b_m points $x_0, \ldots, x_b^{m-1}$ which are defined as [4]

$$x_n := D_d(z_n),$$

the n^{th} point z_n of a PLPS $P_{b^m,ds}(q, p)$ is presented by

$$z_n := \left(v_m\left(\frac{n(x)q_1(x)}{p(x)} \right), v_m\left(\frac{n(x)q_2(x)}{p(x)} \right), \ldots, v_m\left(\frac{n(x)q_{ds}(x)}{p(x)} \right) \right) \in [0,1)^{ds}$$

for $0 \leq n < b^m$. For explanation of $n(x)$ see [13]. A QMC rule, that uses this point set is called an interlaced polynomial lattice rule (of order d). Generating matrices used for the construction of interlaced polynomial lattice point sets were based on the implementation of the Sobol sequence with interlacing factor $d = 2$, as described in [15].

III. Case study - UNI-DEM model

The efficient MC algorithms for multidimensional numerical integration described above have been applied to sensitivity studies of concentration variations of air pollutants with respect to emission levels and some chemical reactions rates. More information can be found in [8], [9], [10].

A. Sensitivity Studies with Respect to Emission Levels

TABLE I
COMPARISON OF RELATIVE ERRORS FOR SENSITIVITY INDICES ESTIMATION USING VARIOUS STOCHASTIC APPROACHES ($n \approx 65536$).

EQ	RV	MSS-2S	IPLR2	CBCCM	FIBO
f_0	0.048	**2e-08**	7e-06	8e-06	9e-06
$\mathbf{D}$	0.0002	**2e-06**	8e-05	6e-04	3e-04
S_1	9e-01	5e-04	**2e-05**	**2e-05**	4e-04
S_2	2e-04	7e-02	**3e-03**	4e-01	2e-01
S_3	1e-01	1e-02	**9e-05**	3e-03	3e-03
S_4	4e-05	6e-01	**1e-02**	1e+00	5e-01
S_1^{tot}	9e-01	1e-03	**1e-05**	3e-04	5e-04
S_2^{tot}	2e-04	3e-03	**2e-03**	1e-01	3e-01
S_3^{tot}	1e-01	4e-03	**8e-05**	9e-04	2e-03
S_4^{tot}	5e-05	1e-01	**2e-02**	4e-01	5e-01

Firstly we will study the sensitivity of the model output (in terms of mean monthly concentrations of several important pollutants) with respect to variation of input emissions of the anthropogenic pollutants. The anthropogenic emissions input consist of 4 different components $\mathbf{E} = (\mathbf{E^A}, \mathbf{E^N}, \mathbf{E^S}, \mathbf{E^C})$ as follows:

$\mathbf{E^A}$ – ammonia (NH_3);
$\mathbf{E^S}$ – sulphur dioxide (SO_2);
$\mathbf{E^N}$ – nitrogen oxides $(NO + NO_2)$;
$\mathbf{E^C}$ – anthropogenic hydrocarbons.

The output of the model is mean monthly concentration of the following 3 pollutants:

- s_1 – ozone (O_3);
- s_2 – ammonia (NH_3);
- s_3 – ammonium sulphate and ammonium nitrate ($NH_4SO_4 + NH_4NO_3$).

In our particular case we are interested in sensitivity studies of the mean monthly concentrations of ammonia in Milan. The domain under consideration is the 4-dimensional hypercubic domain $[0.5, 1]^4$).

This table I compares the relative errors (RE) for estimating sensitivity indices of various input parameters using four stochastic approaches, with a sample size of approximately 65536. The Reference Values (RV) are included for context. for the model function MSS-2S stands out with the smallest relative error (2e-08), indicating exceptional accuracy in estimating the mean output. IPLR2, CBCCM, and FIBO are all orders of magnitude less precise here (errors in the range of 7e-06 to 9e-06). This suggests MSS-2S is highly efficient for estimating basic function averages. For the total variance ($\mathbf{D}$) MSS-2S again shows the lowest RE (2e-06), confirming its strength in capturing overall variance. IPLR2 (8e-05) and FIBO (3e-04) are better than CBCCM (6e-04) but lag behind MSS-2S.

For S_1 (most influential parameter), both IPLR2 and CBCCM achieve the best RE (2e-05), outperforming MSS-2S and FIBO. This suggests these two methods are more precise in estimating dominant first-order effects. For S_2, IPLR2 performs best (3e-03), significantly better than MSS-2S (7e-02), CBCCM (4e-01), and FIBO (2e-01). For S_3 and S_4, IPLR2 consistently has the smallest errors (9e-05 and 1e-02 respectively), with MSS-2S and FIBO trailing and CBCCM performing worst. This indicates IPLR2 is very efficient at estimating sensitivity for less dominant variables. For S_1^{tot} IPLR2 outperforms others with 1e-05 error, better than CBCCM (3e-04), MSS-2S (1e-03), and FIBO (5e-04). For S_2^{tot} IPLR2 again leads with 2e-03, while MSS-2S (3e-03) performs well but CBCCM and FIBO lag behind (0.1 and 0.3). For S_3^{tot} and S_4^{tot}, IPLR2 maintains the best accuracy with relative errors of 8e-05 and 2e-02, respectively, indicating strong performance on total effect indices across less influential parameters.

To summarize, in this case, if the goal is to accurately estimate the mean and total variance, MSS-2S is the method of choice. For detailed sensitivity analysis—precisely estimating both first-order and total-order indices, especially those with smaller magnitude—IPLR2 offers superior accuracy. CBCCM and FIBO show relatively larger errors, suggesting they may be less suitable for high-precision sensitivity studies with this sample size.

B. Sensitivity Studies with Respect to Chemical Reactions Rates

In this part we will study the sensitivity of the ozone concentration in Genova according to the rate variation of some chemical reactions: ## $1, 3, 7, 22$ (time-dependent) and $27, 28$ (time independent) reactions of the condensed CBM-IV scheme ([23]). The simplified chemical equations of these reactions are as follows:

$$
\begin{aligned}
&[\#\mathbf{1}] \quad NO_2 + h\nu \Longrightarrow NO + O;\\
&[\#\mathbf{3}] \quad O_3 + NO \Longrightarrow NO_2;\\
&[\#\mathbf{7}] \quad NO_2 + O_3 \Longrightarrow NO_3;\\
&[\#\mathbf{22}] \quad HO_2 + NO \Longrightarrow OH + NO_2;\\
&[\#\mathbf{27}] \quad HO_2 + HO_2 \Longrightarrow H_2O_2;\\
&[\#\mathbf{28}] \quad OH + CO \Longrightarrow HO_2.
\end{aligned}
$$

The domain under consideration is the 6-dimensional hypercubic domain $[0.6, 1.4]^6$).

TABLE II
COMPARISON OF RELATIVE ERRORS (RE) FOR AE OF SIS ACROSS SELECTED METHODS ($n \approx 2^{16}$).

EQ	RV	MSS-2S	D-α2-Cs-1	CBCCM	FIBO
f_0	0.27	**5e-07**	8.4e-06	1e-05	3e-04
$\mathbf{D}$	0.0025	1e-04	**2.8e-05**	1e-04	2e-03
S_1	4e-01	2e-02	**2.4e-04**	3e-03	4e-02
S_2	3e-01	6e-02	**1.2e-04**	5e-03	1e-02
S_3	5e-02	8e-02	**2.3e-04**	4e-03	5e-01
S_4	3e-01	4e-03	**1.2e-05**	5e-04	1e-02
S_5	4e-07	2e+02	**5.9e-01**	2e+02	3e+03
S_6	2e-02	4e-02	**1.0e-03**	3e-02	1e+00
S_1^{tot}	4e-01	5e-02	**2.2e-04**	7e-03	8e-02
S_2^{tot}	3e-01	3e-02	**2.8e-04**	2e-03	3e-02
S_3^{tot}	5e-02	4e-02	**5.9e-04**	2e-02	1e+00
S_4^{tot}	3e-01	4e-02	**5.8e-05**	1e-03	4e-01
S_5^{tot}	2e-04	1e+00	**1.5e-02**	2e-01	9e+01
S_6^{tot}	2e-02	4e-02	**1.1e-03**	3e-03	2e+00
S_{12}	6e-03	7e-01	**2.8e-03**	3e-01	3e+00
S_{14}	5e-03	1e+00	**6.5e-03**	3e-02	8e+00
S_{24}	3e-03	1e+00	**4.5e-03**	5e-02	1e+01
S_{45}	1e-05	4e+00	**3.3e-02**	1e+00	4e+01

This table II compares the relative errors (RE) for estimating sensitivity indices of various input parameters using four stochastic approaches, with a sample size of approximately 65536.

FIBO performs poorly overall, especially in detecting higher-order and small-effect indices. It delivers acceptable errors only for dominant first-order indices, such as S_1 and S_2, but is far less accurate for total-effect and interaction indices. Also, it is characterized with substantial divergence in weak-effect indices like S_5, S_5^{tot}, and interaction terms—REs are several orders of magnitude larger than other methods. FIBO is inadequate for high-precision sensitivity analysis, particularly in higher-dimensional or low-sensitivity contexts.

CBCCM offers moderately good performance, better than FIBO but generally not better than MSS-2S or IPLR2 It excels in select indices like S_1^{tot} and S_2^{tot} but is not the best in any single row of the table. It tends to perform worse than IPLR2 in detecting fine-grained or subtle effects, although it outperforms MSS-2S and FIBO in some higher-order interactions. CBCCM is reliable but not exceptional. It represents a solid compromise between computational cost and accuracy.

MSS-2S delivers excellent accuracy for f_0 and $\mathbf{D}$, both with errors as low as 10^{-7} to 10^{-4}. It performs well for low-to-moderate variance indices (e.g., S_1, S_2, S_3), but poorly on

higher-order or subtle interactions like S_{24} and S_{45}. It notably fails in estimating some higher-order interactions and weak effects (e.g., very high RE in the smallest in value S_5 and S_5^{tot} which are very important for the reliability of the results). MSS-2S is accurate and efficient for dominant effects, but may be unstable for weak sensitivities or higher-order interactions.

Among all methods, IPLR2 provides the lowest RE for the majority of first-order and total-effect indices, including: S_1, S_2, S_3, S_4, S_1^{tot} through S_6^{tot} and many second-order terms like S_{12}, S_{14}, S_{24}, and S_{45}. It is exceptionally robust across both low and high sensitivity parameters, and particularly effective in capturing small effect sizes (e.g., S_5). It is very competitive even for interaction terms, where other methods tend to diverge. IPLR2 stands out as the most consistently accurate and robust method, particularly effective for precise SI estimation across the full spectrum of effects.

IV. Conclusion

The computational experiments reveal that the optimization method IPLR2 developed are amongst the most effective stochastic strategies currently available for determining sensitivity indices, particularly for the most challenging task – assessing the least value sensitivity indices, which are crucial for the dependability of the model's outcomes. These findings are of considerable significance for environmental conservation and the credibility of future predictions.

The results underscore the trade-off between method complexity and accuracy: MSS-2S excels in some global measures, while IPLR2 is more balanced for all sensitivity indices. Among the evaluated methods for estimating sensitivity indices, our optimization method IPLR2 consistently delivers the most accurate and robust results, especially across both dominant and subtle effects, including higher-order interactions. MSS-2S performs well for main effects but lacks stability in weak or complex interactions. CBCCM offers balanced performance but is never the top performer. In contrast, FIBO and RV show significant limitations in accuracy and should be avoided in precision-critical sensitivity analyses. Overall, the proposed optimization method is the recommended method for comprehensive and reliable global sensitivity estimation.

Acknowledgment

Venelin Todorov was partially supported by the Centre of Excellence in Informatics and ICT under the Grant No BG16RFPR002-1.014-0018, financed by the Research, Innovation and Digitalization for Smart Transformation Programme 2021-2027 and co-financed by the European Union and by BNSF under Project KP-06-N62/6 "Machine learning through physics-informed neural networks".

References

[1] N. Bahvalov. On the Approximate Computation of Multiple Integrals. *Vestn. Mosc. State Univ.* **1959**, *4*, 3–18.

[2] J. Baldeaux, J. Dick, G. Leobacher, D. Nuyens, F. Pillichshammer, Efficient calculation of the worst-case error and (fast) component-by-component construction of higher order polynomial lattice rules, *Numer. Algorithms* **2012**, *59*, 403–431.

[3] R. Cools, F. Kuo, D. Nuyens, Constructing Embedded Lattice Rules for Multivariate Integration, *SIAM J. Sci. Comput.* **2006** *28*, 2162–2188.

[4] J. Dick, F. Pillichshammer, 2010. Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge.

[5] I. Dimov, Monte Carlo Methods for Applied Scientists, New Jersey, London, Singapore, World Scientific, (2008).

[6] I.T. Dimov, R. Georgieva, Monte Carlo Method for Numerical Integration based on Sobol' Sequences, Numerical Methods and Applications (I. Dimov, S. Dimova, N. Kolkovska - Eds.), LNCS 6046, Springer, 2011, 50–59.

[7] I. T. Dimov, R. Georgieva, Multidimensional Sensitivity Analysis of Large-scale Mathematical Models. O.P. Iliev et al. (eds.), Numerical Solution of Partial Differential Equations: Theory, Algorithms, and Their Applications, Springer Proceedings in Mathematics & Statistics 45, Springer Science+Business Media, New York, 2013, 137–156.

[8] I. T. Dimov, R. Georgieva, Tz. Ostromsky, Z. Zlatev, Variance-based Sensitivity Analysis of the Unified Danish Eulerian Model According to Variations of Chemical Rates, – In: I. Dimov, I. Faragó, and L. Vulkov (Eds.) Proc. NAA 2012, LNCS 8236, Springer, 2013, 247–254.

[9] I. T. Dimov, R. Georgieva, Tz. Ostromsky, Z. Zlatev, Sensitivity Studies of Pollutant Concentrations Calculated by UNI-DEM with Respect to the Input Emissions. Central European Journal of Mathematics, Numerical Methods for Large Scale Scientific Computing Vol. 11, No 8, 2013, 1531–1545.

[10] I.T. Dimov, R. Georgieva, Tz. Ostromsky, Z. Zlatev, Advanced Algorithms for Multidimensional Sensitivity Studies of Large-scale Air Pollution Models based on Sobol Sequences, Special issue of Computers and Mathematics with Applications 65 (3), Efficient Numerical Methods for Scientific Applications, Elsevier, 2013, 338 - 351.

[11] I. Dimov, Z. Zlatev, Testing the sensitivity of air pollution levels to variations of some chemical rate constants, *Notes Numer. Fluid Mech.* **1997**, *62*, 167–175.

[12] F. Ferretti, A. Saltelli, S. Tarantola, Trends in sensitivity analysis practice in the last decade, Journal of Science of the Total Environment 568, (2016), 666–670.

[13] T. Goda, 2015. Good interlaced polynomial lattice rules for numerical integration in weighted Walsh spaces. Journal of Computational and Applied Mathematics 285, 279–294.

[14] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, Reliability Engineering and System Safety 52, (1996), 1–17.

[15] S. Joe, F. Kuo, 2008. Constructing Sobol' sequences with better two-dimensional projections. SIAM J. Sci. Comput. 30, 2635–2654.

[16] T. Ostromsky, (2022). Performance and Scalability Experiments with a Large-scale Air Pollution Model on the EuroHPC Petascale Supercomputer DISCOVERER. In FedCSIS (Communication Papers) (pp. 81-84).

[17] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, Halsted Press, New York, (2004).

[18] I.H. Sloan A.V. Reztsov, Component-by-component construction of good lattice rules, *Math. Comp.* **2002**, *71*, 263–273.

[19] I. Sobol, Numerical methods Monte Carlo, Nauka, Moscow, (1973).

[20] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models, Mathematical Modeling and Computational Experiment 1(4), (1993), 407–414.

[21] I.M. Sobol, S. Tarantola, D. Gatelli, S. Kucherenko, W. Mauntz, Estimating the approximation error when fixing unessential factors in global sensitivity analysis, Reliability Engineering & System Safety 92, (2007), 957–960.

[22] Y. Wang, F.J. Hickernell, An historical overview of lattice point sets, In *Monte Carlo and Quasi-Monte Carlo Methods 2000*; Fang, KT., Niederreiter, H., Hickernell, F.J., Eds.; Springer: Berlin/Heidelberg, Germany, 2002, pp. 158–167.

[23] Z. Zlatev, Computer Treatment of Large Air Pollution Models, KLUWER Academic Publishers, Dorsrecht-Boston-London, (1995).

[24] Z. Zlatev, I.T. Dimov, K. Georgiev, Three-dimensional version of the Danish Eulerian model, Z. Angew. Math. Mech. 76(S4), (1996), 473–476.

[25] Z. Zlatev, I.T. Dimov, Computational and Numerical Challenges in Environmental Modelling, Elsevier, Amsterdam, (2006).

[26] The Danish Eulerian Model, Available online: http://www2.dmu.dk/AtmosphericEnvironment/DEM/

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 783–787
DOI: 10.15439/2025F9395
ISSN 2300-5963 ACSIS, Vol. 43

Illumination-targeting Data Augmentation for Monochrome Images

Andrzej Śluzek, Piotr Stachura
0000-0003-4148-2600, 0000-0002-7133-2729
Warsaw University of Life Sciences - SGGW,
Institute of Information Technology,
Nowoursynowska 159, bld.34,
02-776 Warsaw, Poland
Email: {andrzej_sluzek, piotr_stachura1}@sggw.edu.pl

***Abstract*—This paper addresses the challenge of data augmentation for monochrome images, which are still highly relevant in machine vision. Current augmentation methods designed for natural scenes focus on geometric transformation and often fall short in simulating real-world illumination variations. We propose a novel approach that leverages a two-step monochrome image processing technique consisting of image pseudo-colorization followed be a specific decolorization scheme. Our method generates diverse and unpredictable intensity variations, effectively emulating illumination changes and, if only a specific category of color maps are used, preserving the naturalness of the resulting images. To exclude the results which closely replicate images already in the dataset, we rank these maps, identifying a subset that significantly alters illumination characteristics of the originals. A popular SSIM measure is used for that purpose. The proposed technique enhances the illumination diversity of datasets, offering a valuable tool for improving the robustness of AI/ML models in monochrome imagery.**

NOTE: ***All figures are best viewed in high resolution*.**

I. Motivation and Introduction

DATA augmentation is a crucial AI/ML technique that artificially expands training datasets by generating modified samples, [1]. This process enhances robustness of machine vision models to unseen data, reduces overfitting, and improves performance, particularly in scenarios with data scarcity or imbalance, [2]. By leveraging existing data, the need for expensive and time-consuming data collection efforts is minimized.

The data augmentation problem for a specific category of visual data, i.e., monochrome images, is the subject of this paper. These images continue to be a popular and informative source, comparable in relevance to color images.

Commonly, the augmentation of monochrome visual data (e.g., [3], [4]) relies on geometric transformations, encompassing rotation, scaling, cropping, shearing, homography, and random occlusions. Illumination and intensity transformations are less frequently employed, largely due to their limited options, which primarily include adjustments to contrast and brightness, alongside standard image processing techniques like blurring or sharpening. Significantly, current augmentation methods often fall short in their ability to simulate the unpredictable intensity variations caused by real-world illumination changes, such as those arising from diverse weather or visibility conditions.

Consequently, this paper focuses on augmenting visual data by addressing this specific limitation. Directly visualizing the effects of real-world illumination or visibility changes is generally infeasible due to the extremely complex computational demands of 3D scene reconstruction and illumination modeling based on identified or artificially embedded light sources. Instead, we propose utilizing an image processing technique capable of producing varying and unpredictable outcomes that can effectively emulate diverse illumination changes, even though the underlying physical characteristics of these changes remain unspecified and undetermined.

The proposed approach leverages a recently developed image decolorization algorithm [5], which has been subsequently adapted to enhance the quality of grayscale images in [6], [7]. The algorithm is a specific grayscaling scheme which effectively enhances the visual prominence of original color images by incrementally determining pixel intensities based on the colors and intensities of their already processed neighbors (details in [5]). Example results are shown in Fig. 1.



.

Fig. 1. Example images from COLOR250 dataset [8] decolorized using the method described in [5].

Subsequently, any monochrome image can be transformed into a color variant using pseudo-colorization, a common technique for visualizing grayscale data from diverse, primarily non-visual domains, e.g., [9], [10], and then decolorized using the aforementioned approach. We have observed that this two-step process can significantly enhance image perceptibility compared to the original grayscale image, given the application of a suitable *color map* in the initial step. Optimal color maps for this purpose are identified in [6], [7], both for general monochrome images (targeting primarily those from non-visual domains) [7]), and for images depicting natural

scenes, [6]. However, we also observed that for images of natural scenes, many color maps produce monochrome results that, regardless of improved perceptibility, can resemble the same scenes under different illumination conditions, as shown in Fig. 2.



Fig. 2. The original image (left) and its variants, generated using diversified color maps in the first step of our method.

This finding forms the core principle of this paper, as such images are incorporated into the augmented dataset.

First, Section II categorizes color maps employed in pseudo-colorization and identifies the sub-category suitable for the proposed data augmentation scheme. Within this context, the key features of the adopted decolorization scheme are also briefly highlighted. In Section III, those color maps are ranked so that only maps which statistically generate the most diverse (and different from the dataset originals) images are retained. In this way, we exclude the augmented images that basically replicate visual information already provided in the training dataset.

Finally, Section IV presents examples of this augmentation and discusses further aspects of the method, including its limitations.

II. Principles of the Method

A. Remarks on Image Decolorization (more in [5])

The decolorization scheme presented in [5] assumes, but does not directly apply, linear *rgb-to-gray* mappings defined by:

$$I = k_R R + k_G G + k_B B, \tag{1}$$

where the coefficients $[k_R, k_G, k_B]$ are optimized either globally (e.g., [11], [12]) or locally (e.g., [13], [14]) to enhance the visual expressiveness of the resulting monochrome images.

Instead, we observe that a pixel with $[R, G, B]$ values can only be assigned an intensity I within the constrained range:

$$min(R, G, B) \leq I \leq max(R, G, B), \tag{2}$$

regardless of the specific values of $[k_R, k_G, k_B]$ in Eq.1.

Eq. 2 allows us to identify pixels with the narrowest range of feasible grayscale intensities. For these pixels, their grayscale values are either randomly selected from this restricted range or assigned deterministically when when $min(R, G, B) = max(R, G, B)$.

These pixels serve as the starting points (the *initial list*) for the decolorization process, which employs a randomized variant of a *flood-fill* algorithm. Subsequently, pixels with already decolorized neighbors are assigned grayscale intensities (constrained by Eq. 2) in proportion to their color differences from those neighbors. This iterative process continues until all pixels are processed, completing the decolorization step.

Furthermore, as discussed below, Eq. 2 indirectly determines the color maps that can be effectively used in the first step (image pseudo-colorization) of the proposed image augmentation.

B. Color Maps

Various color maps are designed for diverse purposes (e.g., [15], [16]); however, we specifically focus on *sequential color maps*, where the luminance Y of assigned colors changes monotonically with intensity. That is, *if* $I_1 < I_2$ *then* $Y(RGB(I_1)) < Y(RGB(I_2))$. This constraint ensures that the pseudo-colorized image preserves the perceptual order of luminance, which is important for maintaining the natural appearance of the augmented images.

To further preserve perceptual naturalness, we impose another constraint on the color maps, derived from Eq. 2. Specifically, we only use those sequential color maps that satisfy the condition:

$$max[R(0), G(0), B(0)] < 0.5 < min[R(255), G(255), B(255)]. \tag{3}$$

This condition implies that after pseudo-coloring and subsequent decolorization, the darkest pixels in the original image always remain darker than the brightest pixels. Combined with the fact that the average intensity in Eq. 2 generally increases for sequential color maps, the original hierarchy of intensities is statistically maintained. We will jointly refer to sequential maps satisfying Eg. 3 as SEQ3 maps.

At least 42 SEQ3 maps exist, as identified from the publicly available map catalog [17] and other sources.

Figure 3 provides examples of those categories of color maps.

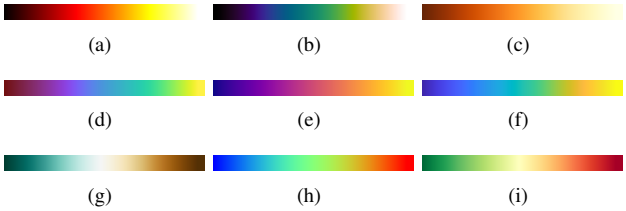


Fig. 3. Examples of: (a,b,c) sequential color maps satisfying Eq. 3 (SEQ3 maps); (d,e,f) non-SEQ3 sequential color maps; and (g,h,i) non-sequential color maps.

To further illustrate the significance of using only SEQ3 color maps, examples are provided in Fig. 4. The results obtained with SEQ3 maps effectively emulate alternative illumination conditions, whereas the image generated using a non-sequential map appears too unrealistic (despite its high visual clarity) to be considered an augmented variant of the original.

III. Restricting the Number of Color Maps

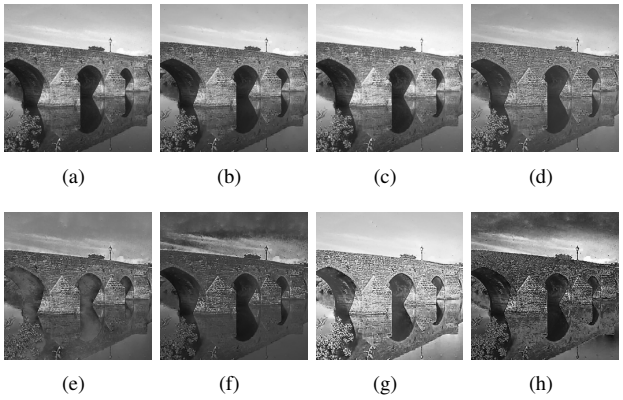
Although our proposed method allows any SEQ3 map to generate realistically-looking alternatives of monochrome images, some maps yield outputs nearly indistinguishable from the originals. Such near-identical images are unsuitable for data augmentation as they essentially replicate existing



(a) (b) (c) (d)

Fig. 4. Original image (a) compared with two variants using SEQ3 maps SEQ3 maps (b,c) and a variant using a non-sequential color map (d), which shows an unacceptable loss of naturalness.

dataset items in terms of overall visual impression. Instead, we should prioritize maps that introduce substantial variations in illumination characteristics. Illustrative examples of both scenarios are provided in Fig. 5.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. (a) Original image; (b-d) variants closely resembling the original (*bone*, *greys*, *ice* SEQ3 maps); and (e-h) variants with significantly different illumination characteristics (*haline*, *inferno*, *OrRd*, *rainforest* SEQ3 maps).

To assess the statistical suitability of various SEQ3 maps for illumination-targeting data augmentation in monochrome imagery, we utilized a large number (N) of grayscale images randomly extracted from the widely used SUN dataset [18]. This dataset was selected to facilitate future research comparing the classification performance of diverse DL architectures trained on either color images or their grayscale versions.

Each test image was processed by our method using all (SEQ3(1) through SEQ3(42)) maps. Subsequently, we evaluated the illumination disparity between each generated image and its original. The most dissimilar results would be potential candidates for inclusion in the augmented dataset.

While visual inspection could assess this dissimilarity, its burden on human evaluators quickly becomes impractical even with a small number of test images. Therefore, after considering several common image similarity metrics, we opted for the structural similarity index measure (SSIM) for automatic evaluation.

The SSIM is defined by a weighted combination of three measures that broadly represent statistical similarities between the intensity, contrast, and structure of two images, X and Y:

$$SSIM(X,Y) = i(X,Y)^{\alpha} \times c(X,Y)^{\beta} \times s(X,Y)^{\gamma} \quad (4)$$

where $i(X,Y) = \frac{2\mu_X\mu_Y+c_1}{\mu_X^2+\mu_Y^2+c_1}$, $c(X,Y) = \frac{2\sigma_X\sigma_Y+c_2}{\sigma_X^2+\sigma_Y^2+c_2}$ and $s(X,Y) = \frac{\sigma_{XY}+c_3}{\sigma_X\sigma_Y+c_3}$, see [19].

The structural similarity index measure (SSIM) ranges from -1 to 1, with a maximum value of 1 indicating identical images.

While lower SSIM values suggest various distortions between an original image and its processed version, in our specific scenario, the image structure is perfectly preserved. Therefore, intensity variations are the primary factor contributing to the SSIM-measured dissimilarity, making it a suitable metric for our focus on illumination changes.

Consequently, for each test images $X(n)$ (where n = 1,...,N), we calculate 42 SSIM values, denoted as their 42 $SSIM(X(n), X_i(n))$ (where i=1,...,42). These values represent the similarity between the original image and its variants $X_i(n)$ generated using each of the 42 SEQ3 maps.

Subsequently, we estimate the norms of the SSIM value vectors across all test images.

$$norm\left\{[SSIM(X(1), X_i(1)), ..., SSIM(X(N), X_i(N))]\right\} \quad (5)$$

The SEQ3 maps are ranked based on the increasing values of the metric defined in Eq. 5. Consequently, maps exhibiting the lowest vector norms are considered strong candidates for data augmentation, as they statistically produce images with significant dissimilarity from the originals. Conversely, maps with the highest vector norms tend to generate near-identical images, rendering them unsuitable for this purpose.

We explored various vector norms, primarily L^k norms with different values of k, both with and without outlier removal. Remarkably, the resulting rankings of SEQ3 maps remained nearly identical across these different norms, especially within the top (recommended maps) and bottom (least recommended maps) tiers. Specifically, the **top** 5 and **bottom** 5 map sets consistently contained the same members, with only minor variations in their internal order, as detailed in Table I, irrespective of the chosen norm.

TABLE I
RANKING OF SEQ3 MAPS, INDICATING THE TOP AND BOTTOM TIERS.

rank	1	2	3	4	5	...	38	39	40	41	42
map	*rainforest*	*hotM*	*hot*	*YlOrBr(OrRd)*	*OrRd(YlOrBr)*	.....	*ocean(ice)*	*ice(ocean)*	*copper*	*greys*	*bone*

Additionally, we have found that the maximum subset of color maps which are always on top of this ranking (although sometimes in a slightly varying order) irrespective of the norm applied in Eq. 5, contains 14 SEQ3 maps.

We further assessed, using SSIM, the cross-dissimilarity of the outputs generated by these color maps. Generally, the results exhibited substantial differences, with the notable exception of *hotM* and *hot*. These two maps, being nearly identical, understandably produced very similar images, suggesting that one should be excluded.

Consequently, for illumination-targeting data augmentation in monochrome natural scene imagery, we recommend employing the following color maps in the first step of our proposed monochrome-to-monochrome (via pseudo-coloring) image alternations:
(1) *rainforest*, (2) *hot*, (3) *YlOrBr*, (4) *OrRd*, (5) *YlOrRd*, (6) *oranges*, (7) *nuclear*, (8) *RdPu*, (9) *reds*, (10) *PuRd,* (11) *speed*, (12) *warm* and (13) *inferno*.

Detailed information on these maps can be found in [17] and other resources.

Specifically, Figures 3(a) 3(b) and 3(c) illustrate the *hot*, *rainforest* and *YlOrBr*, respectively.

Our limited-scale visual validation strongly supports the suitability of these automatically identified color maps.

IV. Summary

A. Examples

Figures 6 to 9 illustrate the augmentation results achieved with the recommended color maps. These figures display original monochrome images of natural scenes and the corresponding outputs produced by a random selection of seven out of the thirteen recommended maps. While some image fragments may show resemblance, their illumination characteristics diverge significantly in other parts.

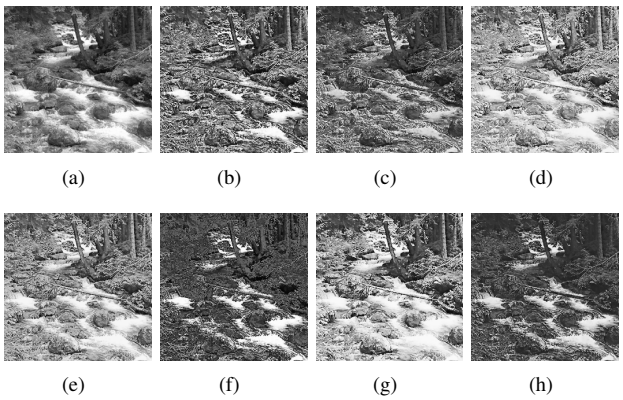


Fig. 6. (a) Original image; augmented variants obtained by using: (b) *hot*, (c) *inferno*, (d) *OrRd*, (e) *PuRd*, (f) *rainforest*, (g) *RdPu* and (h) *speed* SEQ3 maps.

B. Concluding Remarks

Overall, the SEQ3 color maps generate randomized and often unpredictable intensity changes while preserving the natural appearance of the original images. This means our proposed augmentation method effectively produces scene variations that could occur under various unspecified illumination conditions, an effect not achievable with traditional image processing techniques.

Nevertheless, it's important to highlight that our approach addresses only a limited range of potential illumination variations. Specifically, within the proposed model, we cannot simulate the effects of relocating light sources or changing their numbers. Despite this limitation, we believe our method offers a valuable tool for augmenting training data in monochrome imagery, introducing items with unseen and diversified illumination characteristics. Notably, when combined with geometry-based augmentation, this method can yield sufficiently large and diverse training datasets, even when the availability of images from a particular domain is limited.

Finally, it should be noted that restricting the augmenta-

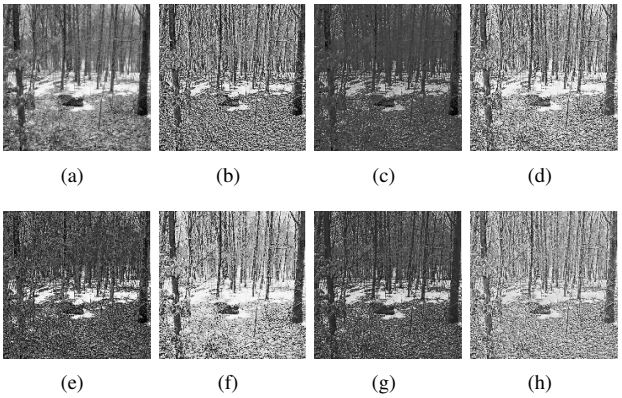


Fig. 7. (a) Original image; augmented variants obtained by using: (b) *hot*, (c) *nuclear*, (d) *oranges*, (e) *rainforest*, (f) *RdPu*, (g) *speed* and (h) *YlOrRd* SEQ3 maps.

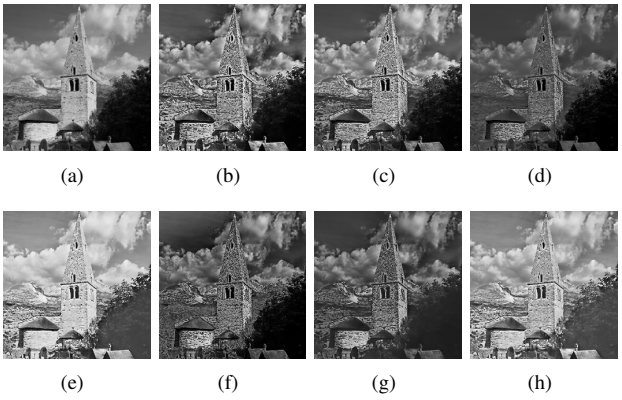


Fig. 8. (a) Original image; augmented variants obtained by using: (b) *hot*, (c) *inferno*, (d) *nuclear*, (e) *OrRd*, (f) *rainforest*, (g) *speed* and (h) *YlOrRd* SEQ3 maps.

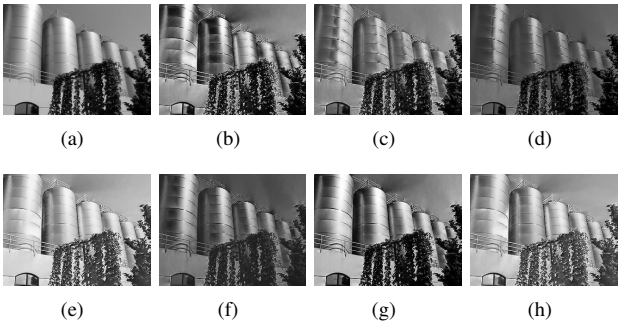


Fig. 9. (a) Original image; augmented variants obtained by using: (b) *hot*, (c) *inferno*, (d) *nuclear*, (e) *OrRd*, (f) *speed*, (g) *warm* and (h) *YlOrRd* SEQ3 maps.

tion to only SEQ3 color maps is not a strict requirement. Generally, non-SEQ3 sequential color maps or non-sequential maps produce unacceptably unnatural outputs, as illustrated in Fig. 3(d). However, if extremely unusual and unpredictable visibility conditions are anticipated in a specific scenario, one might consider incorporating non-SEQ3 or even non-sequential maps into the data augmentation process. Fig. 10 provides an illustrative example of such a case.

Additionally, we have experimentally verified that our approach offers a superior level of intensity diversification compared to standard image processing techniques. We evaluated over ten common, *non-task-specific* image enhancement algorithms. Of these, only two - *image sharpening* and the *LoG* filter - produced statistically comparable SSIM dissimilarity results. However, even these methods were noticeably inferior to ours. It is important to note that image sharpening and the LoG filter primarily enhance edges. This means the visual differences observed with these techniques largely stem from structural modifications, not from the genuine intensity diversification our proposed method achieves.

In the future, we plan to extend our proposed data augmentation scheme to natural color images. Preliminary experimental results for this extension are shown in Figure 11.



Fig. 10. Original image (left) and three augmented variants using non-SEQ3 color maps.



(a) (b) (c) (d)

Fig. 11. (a) Original color image; augmented results obtained via SEQ3 color maps: (b) *inferno*, (c) *RdPu*, and (d) *rainforest*.

REFERENCES

[1] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022. doi: https://doi.org/10.1016/j.array.2022.100258. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000911

[2] Y. Zhang, A. Yilmaz, M. Popa, and C. Brewster, "Analysis of the impact of data augmentation on the performance of deep learning models in multispectral food authenticity identification," in *Proc. 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, vol. 35. IEEE, 2023. doi: 10.15439/2023F3643 p. 823–832.

[3] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018. doi: 10.1109/IIPHDW.2018.8388338

[4] L. Nanni, M. Paci, S. Brahnam, and A. Lumini, "Comparison of different image data augmentation approaches," *Journal of Imaging*, vol. 7, no. 12, p. 222–239, 2014. doi: 10.3390/jimaging7120254

[5] A. Śluzek, "Incremental image decolorization with randomizing factors," in *32nd European Signal Processing Conference (EUSIPCO)*, 2024. doi: 10.23919/EUSIPCO63174.2024.10715444 pp. 591–595.

[6] ——, "An unorthodox technique for enhancing monochrome images of natural scenes," in *Proc. 6th Polish Conference on Artificial Intelligence, Katowice, Poland*. Springer, 2025, in print.

[7] ——, "A new technique for enhanced monochrome visualization of non-visual data," in *Computational Science – ICCS 2025 Workshops*, M. Paszynski, A. S. Barnard, and Y. J. Zhang, Eds. Springer Nature Switzerland, 2025, pp. 51–65.

[8] C. Lu, L. Xu, and J. Jia, "Contrast preserving decolorization with perception-based quality metrics," *International Journal of Computer Vision*, vol. 110, p. 222–239, 2014. doi: 10.1007/s11263-014-0732-6

[9] N. S. Aghdam, M. M. Amin, M. E. Tavakol, and E. Ng, "Designing and comparing different color map algorithms for pseudo-coloring breast thermograms," *Journal of Medical Imaging and Health Informatics*, vol. 3, no. 4, pp. 487–493, 2013. doi: 10.1166/jmihi.2013.1191

[10] A. Rahimian, M. Etehadtavakol, M. Moslehi, and E. Ng, "Comparing different algorithms for the pseudo-coloring of myocardial perfusion single-photon emission computed tomography images," *Journal of Imaging*, vol. 8, no. 12, p. 331, 12 2022. doi: 10.3390/jimaging8120331

[11] Q. Liu, J. Xiong, L. Zhu, M. Zhang, and Y. Wang, "Extended RGB2Gray conversion model for efficient contrast preserving decolorization," *Multimedia Tools and Applications*, vol. 76, p. 14055–14074, 2017. doi: 10.1007/s11042-016-3748-9

[12] T. Wu, C. Eising, M. Glavin, and E. Jones, "An efficient and effective image decolorization algorithm based on cumulative distribution function," *Journal of Imaging*, vol. 10, no. 3, 2024. doi: 10.3390/jimaging10030051

[13] T. Wu and A. Toet, "Color-to-grayscale conversion through weighted multiresolution channel fusion," *Journal of Electronic Imaging*, vol. 23, no. 4, p. 043004, 2014. doi: 10.1117/1.JEI.23.4.043004

[14] J. Yu, F. Li, and X. Lv, "Contrast preserving decolorization based on the weighted normalized l1 norm," *Multimedia Tools and Applications*, vol. 80, p. 31753–31782, 2021. doi: 10.1007/s11042-021-11172-9

[15] M. Khan, Y. Gotoh, and N. Nida, "Medical image colorization for better visualization and segmentation," in *Medical Image Understanding and Analysis*, M. Valdés Hernández and V. González-Castro, Eds. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-60964-5_50 pp. 571–580.

[16] K. Moreland, "Why we use bad color maps and what you can do about it," in *Proc. IS&T Int'l. Symp. on Electronic Imaging: Human Vision and Electronic Imaging*, vol. 28, 2016. doi: 10.2352/ISSN.2470-1173.2016.16.HVEI-133

[17] "Colormap catalog," accessed on Jan. 5, 2025. [Online]. Available: https://cmap-docs.readthedocs.io/en/latest/catalog/

[18] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Conference CVPR*, 2010. doi: 10.1109/CVPR.2010.5539970 pp. 3485–3492.

[19] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. doi: 10.1109/TIP.2003.819861

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 789–794
DOI: 10.15439/2025F4752
ISSN 2300-5963 ACSIS, Vol. 43

Improved upper bounds on the shortest watchman route in simple polygons: Dependence on reflex vertices and triangulation strategies

Lubomír Štěpánek[1, 2, 3]
ORCiD: 0000-0002-8308-4304
[1]Department of Statistics and Probability, [2]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
[3]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, 12000 Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

***Abstract*—This paper investigates upper bounds on the length of the shortest watchman route in simple polygons using minimal structural information – specifically, the number and placement of reflex vertices – rather than full geometric detail. We show that the set of reflex vertices alone suffices to guarantee full visibility of the polygon: every interior or boundary point is visible from at least one reflex vertex. Moreover, these vertices can always be connected by a polyline within the polygon, enabling a constructive approach to route design. While it is known that the shortest watchman route is bounded above by the perimeter of the polygon, we prove that this bound is tight and cannot be improved in general. However, we identify several important classes of polygons where significantly better bounds are achievable. In particular, for polygons with exactly two reflex vertices, we establish that the shortest watchman route is strictly shorter than half the perimeter. We also introduce a novel triangulation method – *successive dual-edge triangulation* – in which each triangle is constructed by attaching to existing polygon edges. When such a triangulation is possible, the resulting watchman route has length strictly less than half the perimeter, regardless of the number of sides or reflex vertices. These results are not only theoretical but also constructive, offering efficient route designs when geometric data is available. The approach is particularly well suited to real-world environments such as industrial halls or private courtyards, where structural simplicity often permits route planning without requiring full geometric reconstruction. In such cases, our methods provide a significant and elegant reduction in route length compared to perimeter-based patrolling, making them applicable to scenarios such as autonomous drone navigation and indoor surveillance.**

I. Introduction

THE watchman route problem is a fundamental question in computational geometry and visibility theory. It seeks the shortest path within a polygonal domain from which every point is visible, with applications in robotic surveillance, patrolling, and security optimization.

For simple polygons (without holes), several polynomial-time algorithms exist. The first general solution, by Carlsson et al. [1], had worst-case complexity $\mathcal{O}(n^6)$, later improved to $\mathcal{O}(n^5)$ by Tan [2], and to $\mathcal{O}(n^4 \log\log n)$ in special cases by Chin and Ntafos [3], where n is the number of vertices. These methods assume full geometric detail, making them less practical when only structural features are known.

In polygons with holes, Mitchell showed the problem is NP-hard and proposed the first polynomial-time approximation algorithm with an $\mathcal{O}(\log^2 n)$ factor and showed strong inapproximability results [4]. Dumitrescu and Tóth [5] provided asymptotic upper bounds on the watchman route length in polygons with holes, expressed as $\mathcal{O}\left(\mathrm{perim}(\mathcal{P}) + \sqrt{k} \cdot \mathrm{diam}(\mathcal{P})\right)$ for a polygon $\mathcal{P}$ with k holes, perimeter $\mathrm{perim}(\mathcal{P})$, and diameter $\mathrm{diam}(\mathcal{P})$. However, they did not derive explicit route lengths or provide exact constants in the upper bounds. In contrast, this paper provides exact and constructive bounds for specific polygon classes.

Extensions of the problem include limited-visibility models [6], [7], multi-watchman strategies [8], and static coverage ratios from art gallery results [9], [10]. Similar geometric reasoning appears in other domains such as bioinformatics. For example, Lavor et al. [11] solve the discretizable molecular distance geometry problem using partial distance constraints and structural sparsity to infer configurations.

Most existing approaches to the watchman route problem focus on minimizing algorithmic complexity or improving approximation ratios, typically assuming full geometric data.

In practice, however, environments such as industrial halls or courtyards often have simple layouts, where detailed mapping is unnecessary and only coarse information – such as the number of vertices or reflex angles – is available.

This paper studies the watchman route problem in simple polygons under such limited assumptions. We show that the set of reflex vertices alone suffices to ensure full visibility, and that these vertices can always be connected by a polyline. The shortest watchman route is thus bounded above by the polygon's perimeter, and we prove that this bound is tight in general.

We then identify special structural cases that allow for strictly better bounds. For example, if the polygon has exactly two reflex vertices, the length of the shortest route is strictly less than half the perimeter. We also introduce a geometric method – *successive dual-edge triangulation* – in which the polygon is incrementally triangulated by attaching new triangles to existing edges. When such a triangulation is possible, the route length is guaranteed to be strictly less than half the perimeter, regardless of the polygon's size or complexity.

While our results are derived from general structural principles, they also yield constructive algorithms. These are well suited to practical scenarios where detailed geometry is unavailable, and even when such data is present, our method offers an elegant and analytically grounded alternative to exhaustive geometric search.

II. Preliminaries

This section introduces the geometric foundations, mostly elementary but essential, that underlie our results. We define point visibility and show that every point in a simple polygon is visible from some boundary point. In convex polygons, all pairs of points are mutually visible. For polygons with reflex vertices, each point is visible from at least one reflex vertex. Finally, we show that all reflex vertices can be connected by a polyline within the polygon, a key step in the proposed constructing watchman routes.

Definition 1 (Visibility of a point by another)**.** *Let $\mathcal{P}$ be a simple polygon and let $A \in \mathcal{P}$, $B \in \mathcal{P}$ be two points (either in the interior or on the boundary of $\mathcal{P}$). We say that* point A is visible by point B *(or equivalently,* B is visible from A*) if the closed line segment $\overline{AB}$ lies entirely within $\mathcal{P}$, that is, $\overline{AB} \subset \mathcal{P}$. Equivalently, we may say that* point A sees point B*, or vice versa.*

Lemma 1 (Visibility from the boundary of polygon)**.** *Let $\mathcal{P}$ be a simple polygon (of a finite size). Then every point $A \in P$ (including points in the interior and on the boundary) is visible by at least one point $B \in \partial\mathcal{P}$, where $\partial\mathcal{P}$ denotes the boundary of the polygon.*

Proof. Obviously $\partial\mathcal{P} \subset \mathcal{P}$ (†). Let $A \in P$ be an arbitrary point.

If A lies on the boundary $\partial\mathcal{P}$, i.e. $A \in \partial\mathcal{P}$, then it is trivially visible by itself, so $B = A$, because

$$\overline{AB} = B = A \in \partial\mathcal{P} \overset{(\dagger)}{\subset} \mathcal{P}.$$

Suppose A lies in the interior of $\mathcal{P}$. Since $\mathcal{P}$ is a simple polygon of a finite size, the line segment from A to any direction will eventually intersect the boundary of $\mathcal{P}$. Let us select an arbitrary direction and consider the ray starting at A in that direction. Let C be the first point where this ray intersects the boundary $\partial\mathcal{P}$, so $C \in \partial P$. Since $\mathcal{P}$ is simple and of a finite size (i.e., it has no self-intersections and encloses a connected finite region), the point C exists, segment $\overline{AC}$ has $|\overline{AC}| < \infty$ and lies entirely within $\mathcal{P}$, i.e., $\overline{AC} \subset \mathcal{P}$ (‡). Hence, $B = C$ and A is visible by B because

$$\overline{AB} = \overline{AC} \overset{(\ddagger)}{\subset} \mathcal{P}.$$

Thus, in all cases, there exists a boundary point $B \in \partial\mathcal{P}$ such that A is visible by B. □

Lemma 2 (Mutual visibility in convex polygons)**.** *Let $\mathcal{P}$ be a simple and convex polygon. Then every pair of distinct points $A \in \mathcal{P}$, $B \in \mathcal{P}$ (including interior and boundary points) are mutually visible; that is, the segment $\overline{AB}$ lies entirely within $\mathcal{P}$.*

Proof. Let $A, B \in \mathcal{P}$ be two distinct points, each either in the interior or on the boundary of the convex polygon $\mathcal{P}$. Consider the line $\overleftrightarrow{AB}$ passing through points A and B. Since $\mathcal{P}$ is convex and of finite size, the intersection of this line with the polygon $\mathcal{P}$ is a single line segment $\overline{CD}$, where $C, D \in \partial\mathcal{P}$ are the boundary points where the line enters and exits the polygon. Because $\mathcal{P}$ is convex, the entire segment $\overline{CD}$ lies within $\mathcal{P}$. Since both A and B lie on the segment $\overline{CD}$, so $\overline{AB} \subseteq \overline{CD}$, it follows that $\overline{AB} \subseteq \mathcal{P}$. Therefore, $\overline{AB} \subseteq \mathcal{P}$, which implies that A is visible by B and vice versa. □

Theorem 1 (Visibility from reflex vertices)**.** *Let $\mathcal{P}$ be a simple polygon with boundary $\partial\mathcal{P}$, and suppose that $\mathcal{P}$ contains at least one vertex $V \in \partial\mathcal{P}$ with a reflex angle. Then every point $A \in \mathcal{P}$ (including interior and boundary points) is visible by at least one reflex vertex of $\mathcal{P}$.*

Proof. Let $A \in \mathcal{P}$ be an arbitrary point. Let $\mathcal{W}_A$ denote the set of all boundary points visible from A, that is,

$$\mathcal{W}_A = \left\{\forall B \in \partial\mathcal{P} : \overline{AB} \subset \mathcal{P}\right\}.$$

Case 1: Suppose A sees the entire boundary, i.e., $\mathcal{W}_A = \partial\mathcal{P}$. Since $\mathcal{P}$ contains at least one reflex vertex V, it follows that $V \in \mathcal{W}_A$, and therefore $\overline{AV} \subset \mathcal{P}$. Hence, A is visible by a reflex vertex V.

Case 2: Suppose A does not see the entire boundary, i.e., $\mathcal{W}_A \subset \partial\mathcal{P}$. Let $C \in \partial\mathcal{P} \setminus \mathcal{W}_A$ be a boundary point not visible from A.

Subcase 2a: If $\mathcal{W}_A$ contains at least one reflex vertex V, then A is visible by V, and the statement holds.

Subcase 2b: Suppose $\mathcal{W}_A$ contains no reflex vertices; that is, every boundary point visible from A is either a convex vertex or a point on an edge between convex vertices. Since C is not visible from A and $\mathcal{P}$ is a simple polygon (i.e., it has no self-intersections or holes), the obstruction to visibility must be caused by some part of the boundary of $\mathcal{P}$. Consider the

collection of points $D \in \partial\mathcal{P}$ that obstruct the visibility from A to C; thus, $\overline{AC} \not\subset \mathcal{P}$. Among these, let $D^* \in \partial\mathcal{P}$ be a point such that the angle $\angle AD^*C$ is maximized. Since $\overline{AC} \not\subset \mathcal{P}$, this angle must be reflex, i.e., $|\angle AD^*C| > 180°$. This implies that D^* is a vertex of $\mathcal{P}$ with a reflex angle. Let $\overleftrightarrow{AD^*}$ denote the line passing through points A and D^*. Furthermore, since $C^* \in \overleftrightarrow{AD^*} \cap \partial\mathcal{P}$ is visible from A (if not, angle $\angle AD^*C$ cannot be maximized), so $\overline{AC^*} \subset \mathcal{P}$ and because $\overline{AD^*} \subset \overline{CD^*}$, the segment $\overline{AD^*}$ lies within $\mathcal{P}$, confirming that A is visible by D^*.

In both cases, we have shown that point A is visible by at least one reflex vertex of $\mathcal{P}$. Since $A \in \mathcal{P}$ was arbitrary, the result holds for all interior and boundary points of $\mathcal{P}$. □

Theorem 2 (Connectivity of reflex vertices)**.** *Let $\mathcal{P}$ be a simple polygon with a finite number of vertices, and suppose that $\mathcal{P}$ contains at least one reflex vertex. Then all reflex vertices of $\mathcal{P}$ can be connected by a polyline lying entirely within $\mathcal{P}$.*

Proof. Let $\mathcal{R} = \{R_1, R_2, \ldots, R_k\}$ denote the set of all reflex vertices of $\mathcal{P}$, where $k \geq 1$ and each $R_i \in \partial\mathcal{P}$ is a vertex, i.e. a point of polygon boundary $\partial\mathcal{P}$, at which the internal angle exceeds $180°$.

Case 1: If $k = 1$, there is only one reflex vertex. It is trivially connected to itself, so the statement holds.

Case 2: Let $k \geq 2$. We aim to construct a polyline connecting all reflex vertices using segments that lie entirely within $\mathcal{P}$. Select any two distinct reflex vertices, say A and B, where $A \in \mathcal{R}$, $B \in \mathcal{R}$.

Subcase 2a: If the segment $\overline{AB} \subset \mathcal{P}$, then A and B are directly connected.

Subcase 2b: If $\overline{AB} \not\subset \mathcal{P}$, the segment is obstructed by the boundary of $\mathcal{P}$. Since $\mathcal{P}$ is a simple polygon, the obstruction must be due to other parts of the boundary, not self-intersections or holes. Following a similar argument as used in reflex vertex visibility, we consider points along the boundary that block the visibility from A to B. Among those, there exists a vertex $D \in \partial\mathcal{P}$ such that the angle $\angle ADB$ is maximized. Because the segment $\overline{AB}$ is not entirely in $\mathcal{P}$, this angle must be greater than $180°$, implying that D is a reflex vertex. Clearly, the segment $\overline{AD} \subset \mathcal{P}$, so A and D are connected. If $\overline{DB} \subset \mathcal{P}$, then D and B are connected as well, and hence A, D, and B are connected via a polyline. If $\overline{DB} \not\subset \mathcal{P}$, we repeat the same argument: select a reflex vertex $D' \in \partial\mathcal{P}$ such that $\angle DD'B$ is maximized. Again, D' is a reflex vertex, and we connect D to D' via a segment in $\mathcal{P}$. If needed, this process is repeated. Since $\mathcal{P}$ has a finite number of vertices, and each iteration introduces a new reflex vertex not previously used, this process must terminate after a finite number of steps. Eventually, a polyline connecting A to B through intermediate reflex vertices is constructed.

By repeating this process for all pairs of reflex vertices, a single polyline connecting the entire set $\mathcal{R}$ within $\mathcal{P}$ can be constructed.

Hence, all reflex vertices in $\mathcal{P}$ can be connected by a polyline lying entirely within the polygon. □

III. Upper bounds on the shortest watchman route length based on the number of reflex vertices

In this section, we study upper bounds on the length of the shortest watchman route within a simple polygon, based on the number of reflex vertices it contains. We begin by formally defining a watchman route as a path within the polygon from which every point is visible from at least one position along the path.

We then analyze several cases: polygons with no reflex vertices, one reflex vertex, and two reflex vertices, showing how these configurations lead to increasingly tighter bounds on the route length. For polygons with three or more reflex vertices, we demonstrate that the general upper bound – equal to the perimeter of the polygon – cannot be improved in all cases. This is supported by a specific construction in which the shortest watchman route is (almost) necessarily equal to the perimeter. However, we also introduce a geometric technique we call *successive dual-edge triangulation*, and show that if such a triangulation is possible for the given polygon, the upper bound on the route length can be improved significantly – specifically, to below half the perimeter.

Definition 2 (Watchman route)**.** *Let $\mathcal{P}$ be a simple polygon. A* watchman route *is a continuous path $\Gamma \subseteq \mathcal{P}$ such that for every point $A \in \mathcal{P}$, there exists a point $B \in \Gamma$ for which $\overline{AB} \subseteq \mathcal{P}$. In other words, every point of the polygon is visible from at least one point along the route.*

For a polygon with no reflex vertices or with only a single reflex vertex, the shortest watchman route is straightforward to determine, and its length admits a tight and easily computable upper bound.

Theorem 3 (Upper bound on watchman route length in polygon with no reflex vertex)**.** *Let $\mathcal{P}$ be a simple polygon with no reflex vertices. Then the shortest watchman route in $\mathcal{P}$ has length zero, and any single point within $\mathcal{P}$ (including boundary points) serves as a valid watchman route.*

Proof. Since $\mathcal{P}$ has no reflex vertices, it is convex by definition. By Lemma 2, every point in $\mathcal{P}$ is visible from every other point. Therefore, any single point $A \in \mathcal{P}$ satisfies the condition that every point in the polygon is visible from A. Hence, the constant path $\Gamma = \{A\}$ is a valid watchman route. Because this path consists of a single stationary point, its length is zero, $|\Gamma| = |\{A\}| = 0$. Thus, the upper bound on the shortest watchman route length is exactly zero, and any point in $\mathcal{P}$ realizes this bound. □

Theorem 4 (Upper bound on watchman route length in polygon with one reflex vertex)**.** *Let $\mathcal{P}$ be a simple polygon with exactly one reflex vertex R. Then the shortest watchman route in $\mathcal{P}$ has length zero, and the reflex vertex R itself serves as a valid watchman route, so $\Gamma = \{R\}$.*

Proof. By Theorem 1, every point in $\mathcal{P}$ (interior or boundary) is visible from at least one reflex vertex. Since $\mathcal{P}$ contains exactly one reflex vertex R, it follows that every point in $\mathcal{P}$

is visible from R. Thus, the constant path $\Gamma = \{R\}$ is a valid watchman route. Since it consists of a single stationary point, its length is zero, $|\Gamma| = |\{R\}| = 0$. Therefore, the upper bound on the shortest watchman route length is zero, and R realizes this bound. □

Remark 1. *While the reflex vertex R is guaranteed to serve as a valid watchman route, other points in $\mathcal{P}$ may also provide full visibility. However, R is a point that ensures this property without requiring further geometric analysis.*

We now consider the case of a simple polygon containing exactly two reflex vertices. In this setting, we show that a valid watchman route not only exists but can also be constructed with a length strictly less than half the perimeter of the polygon.

Theorem 5 (Upper bound on watchman route length in polygon with two reflex vertices)**.** *Let $\mathcal{P}$ be a simple polygon with exactly two reflex vertices, denoted R_1 and R_2. Then the shortest watchman route in $\mathcal{P}$ has length strictly less than half the perimeter of $\mathcal{P}$.*

Proof. By Theorem 1, every point in $\mathcal{P}$ is visible from at least one of the two reflex vertices R_1 or R_2. By Theorem 2, the two reflex vertices can be connected by a polyline lying entirely within $\mathcal{P}$. Since there are only two reflex vertices, the polyline reduces to the segment $\overline{R_1R_2}$. Therefore, any point visible from R_1 or R_2 is also visible from any point on the segment $\overline{R_1R_2}$, as visibility is preserved along the connecting segment. We now prove that the length of this segment is strictly less than half the perimeter of $\mathcal{P}$. Let the perimeter of $\mathcal{P}$ be denoted by $p = \text{perim}(\mathcal{P})$. Traversing from R_1 to R_2 along the boundary of $\mathcal{P}$, there are two distinct polygonal paths: one in the clockwise direction, denoted Π_1, and one in the counter-clockwise direction, denoted Π_2. Let ℓ_1 and ℓ_2 be the respective lengths of these two paths. Then,

$$|\Pi_1| + |\Pi_2| = \ell_1 + \ell_2 = p = \text{perim}(\mathcal{P}). \tag{1}$$

Since $\mathcal{P}$ is simple and R_1, R_2 lie on its boundary, the segment $\overline{R_1R_2}$ is a chord that spans across the interior of the polygon and lies entirely within $\mathcal{P}$. At least one of the polygonal (perimeter) paths Π_1 or Π_2 must contain at least one additional vertex, as a polygon $\mathcal{P}$ cannot consist of only two vertices. Without loss of generality, assume that the path Π_1 contains additional (distinct) vertices $A_1, A_2, \ldots, A_m$ for $m \geq 1$. By the triangle inequality, we have

$$\begin{array}{rcccl} |\overline{R_1R_2}| & < & |\overline{R_1A_1}| & + & |\overline{R_2A_1}|, \\ |\overline{R_2A_1}| & < & |\overline{A_1A_2}| & + & |\overline{R_2A_2}|, \\ \cdots & < & \cdots & + & \cdots, \\ |\overline{R_2A_{m-1}}| & < & |\overline{A_{m-1}A_m}| & + & |\overline{R_2A_m}|. \end{array}$$

Repeatedly substituting $|\overline{R_2A_i}|$ for $i \in \{1, 2, \ldots, m-1\}$ in each inequality with the corresponding bound from the next inequality yields

$$|\overline{R_1R_2}| < |\overline{R_1A_1}| + \sum_{i=1}^{m-1} |\overline{A_iA_{i+1}}| + |\overline{R_2A_m}| = |\Pi_1| = \ell_1.$$

The path Π_2 may contain no additional vertices, so it follows directly that $|\overline{R_1R_2}| \leq |\Pi_2| = \ell_2$. Hence, we have

$$|\overline{R_1R_2}| < \ell_1 \qquad \text{and} \qquad |\overline{R_1R_2}| \leq \ell_2,$$

and thus, $|\overline{R_1R_2}| + |\overline{R_1R_2}| = 2 \cdot |\overline{R_1R_2}| < \ell_1 + \ell_2$, so

$$|\overline{R_1R_2}| < \frac{\ell_1 + \ell_2}{2} \overset{(1)}{=} \frac{p}{2}.$$

So, the segment $\overline{R_1R_2}$ has length strictly less than half the perimeter, and it serves as a valid watchman route since every point in $\mathcal{P}$ is visible from R_1 or R_2 and thus from a point on $\overline{R_1R_2}$.

This establishes that the shortest watchman route in a polygon with two reflex vertices has length strictly less than half the polygon perimeter. □

Remark 2. *The following examples illustrate the variability in shortest watchman routes for polygons with two reflex vertices:*

(a) In a butterfly-shaped polygon, the reflex vertices form symmetrical "wings", and a centrally placed point can see the entire region. Here, the shortest watchman route Γ has length zero, consistent with Theorem 5. See Fig. 1, part (a).

(b) In contrast, a polygon with a long rectangular base and two small inward-curved reflex "horns" may require a segment connecting the reflex vertices for full visibility. In this case, Γ approaches half the perimeter, showing that the bound in Theorem 5 is tight. See Fig. 1, part (b).

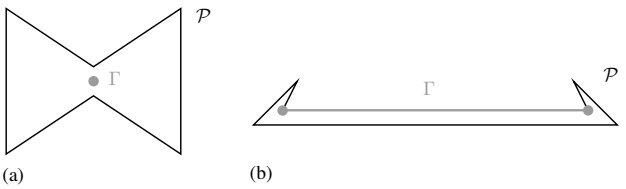


Fig. 1. Examples of polygons with two reflex vertices: (a) a butterfly-shaped polygon $\mathcal{P}$, where a single-point watchman route Γ (in grey) suffices; (b) a skinny polygon $\mathcal{P}$ with inward reflex "horns", where the shortest route Γ (in grey) approaches half the perimeter.

We now turn to the case of polygons with three or more reflex vertices. In general, the shortest watchman route in such polygons can approach the full perimeter, and we show that this upper bound cannot be improved in all cases. However, for certain polygonal structures, the application of a *successive dual-edge triangulation* – when possible – enables a significant improvement, reducing the upper bound on the route length to below half the perimeter.

Theorem 6 (Upper bound on watchman route length in polygon with three or more reflex vertices)**.** *Let $\mathcal{P}$ be a simple polygon with at least three reflex vertices. Then the length of the shortest watchman route in $\mathcal{P}$ is at most equal to*

the perimeter of $\mathcal{P}$. Moreover, this upper bound cannot be generally improved.

Proof. By Lemma 1, every point in $\mathcal{P}$ is visible from at least one point on the boundary $\partial\mathcal{P}$, so a route along the boundary guarantees full visibility. Thus, the perimeter is a valid upper bound for the shortest watchman route. To potentially improve this bound, one might consider a polyline connecting all reflex vertices, since Theorem 1 ensures that every point in $\mathcal{P}$ is visible from at least one of them. However, this polyline may still approach the perimeter length. To show this, we construct a family of polygons, see Fig. 2, with alternating convex and reflex vertices arranged as repeated triples (A_i, B_i, R_i), where R_i is reflex. Each "tooth" – a subregion formed near vertex B_i, which lies behind a reflex corner – necessitates that the watchman visit the corresponding reflex vertex R_i to maintain visibility, particularly of the B_i vertices. Connecting the reflex vertices yields a route Γ whose length closely matches the perimeter, as the segments $\overline{R_iR_{i+1}}$ are only slightly shorter than the corresponding polygonal arcs $\overline{R_iA_{i+1}B_{i+1}R_{i+1}}$. Therefore, while the perimeter is always a valid upper bound, it is asymptotically tight in general and cannot be improved without additional structural assumptions. □

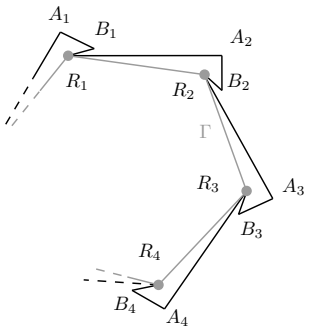


Fig. 2. A polygon constructed from repeated vertex triples (A_i, B_i, R_i), where A_i and B_i are convex and R_i is reflex. The visibility constraints force the watchman to visit each reflex vertex, resulting in a route Γ (in grey) whose length approaches the full perimeter. This demonstrates that the general upper bound cannot be improved for polygons with three or more reflex vertices.

Definition 3 (Successive dual-edge triangulation)**.** *Let $\mathcal{P}$ be a simple n-sided polygon with vertices $A_1, A_2, \ldots, A_n$. A triangulation $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ is called a* successive dual-edge triangulation *of polygon $\mathcal{P}$ if the following conditions are satisfied,*

- *each triangle $\tau_j \in \mathcal{T}$ is formed by three vertices of $\mathcal{P}$ and lies entirely within $\mathcal{P}$;*
- *the union of all triangles in $\mathcal{T}$ exactly covers the interior of $\mathcal{P}$ without overlaps or gaps;*
- *the first triangle $\tau_1 = \triangle A_{i_1}A_{i_2}A_{i_3}$ shares at least two of its edges with the boundary $\partial\mathcal{P}$ of polygon $\mathcal{P}$;*
- *Every subsequent triangle τ_j for $j > 1$ shares one edge with the previously constructed triangle τ_{j-1}, and one or two of its edges must coincide with edges of $\partial\mathcal{P}$. Consequently, triangle τ_j has two edges not shared with $\partial\mathcal{P}$ for $j < m$, and one such edge if $j = m$.*

Fig. 3 illustrates an example of a polygon with a successive dual-edge triangulation. The triangulation proceeds from triangle $\tau_1 = \triangle A_1A_2A_3$ and continues through adjacent triangles such as $\tau_2 = \triangle A_1A_3A_{24}$ and $\tau_3 = \triangle A_3A_{23}A_{24}$, ultimately reaching $\tau_{21} = \triangle A_{12}A_{13}A_{14}$. The dotted lines indicate the internal diagonals added during the successive construction.

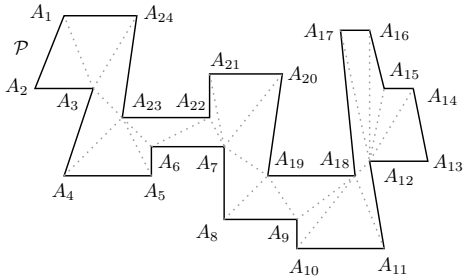


Fig. 3. An example of a polygon $\mathcal{P}$ with 24 vertices, demonstrating a successive dual-edge triangulation. Dotted lines represent internal diagonals added in a step-by-step manner. The triangulation includes, among others, the triangles $\tau_1 = \triangle A_1A_2A_3$, $\tau_2 = \triangle A_1A_3A_{24}$, $\tau_3 = \triangle A_3A_{23}A_{24}$, and continues up to $\tau_{21} = \triangle A_{12}A_{13}A_{14}$.

Theorem 7 (Upper bound improvement via successive dual-edge triangulation)**.** *Let $\mathcal{P}$ be a simple polygon with perimeter p, and suppose that $\mathcal{P}$ admits a successive dual-edge triangulation as defined in Definition 3. Then the length of the shortest watchman route in $\mathcal{P}$ is strictly less than $\frac{p}{2}$.*

Proof. Let $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_m\}$ be a successive dual-edge triangulation of $\mathcal{P}$, as defined in Definition 3. We construct a watchman route Γ by connecting the midpoints of edges not shared with the polygon boundary $\partial\mathcal{P}$ that appear in the triangles of $\mathcal{T}$.

For the initial triangle τ_1, we add a segment connecting the midpoint of one boundary edge to the midpoint of the edge not shared with $\partial\mathcal{P}$. For each subsequent triangle τ_j with $1 < j < m$, we add a segment connecting the midpoints of its two edges that are not shared with the boundary. For the final triangle τ_m, we again add a segment connecting the midpoint of the edge not shared with $\partial\mathcal{P}$ to the midpoint of one boundary edge. By the construction of the triangulation, these segments are connected consecutively and lie entirely within the interiors of the corresponding triangles. Since the union of all triangles in $\mathcal{T}$ covers the entire polygon $\mathcal{P}$, the resulting path Γ guarantees visibility of the entire region.

Each segment in the route corresponds to half the length of a polygon boundary edge (or two such edges in the case of τ_1 and τ_m). Since no boundary edge is used in more than one triangle, the total contribution to the route is at most the sum of half-lengths of all boundary edges. Moreover, the first and last triangles each involve two boundary edges but contribute only one midline segment, meaning that two boundary edges are not halved and do not contribute to the route.

Therefore, the total length of the constructed watchman route Γ is strictly less than half the perimeter, $|\Gamma| < \frac{p}{2}$. □

Figure 4 illustrates the construction of a watchman route

based on a successive dual-edge triangulation of a 24-vertex polygon, using midlines of interior triangles to ensure visibility.

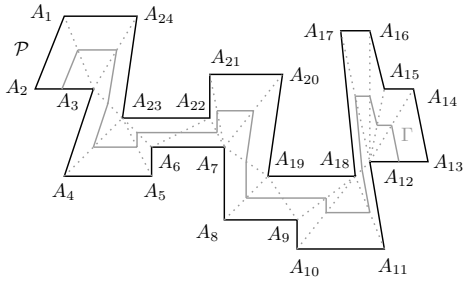


Fig. 4. A simple polygon $\mathcal{P}$ with 24 vertices, along with its successive dual-edge triangulation (indicated by dotted lines) and a watchman route Γ (shown as a bold grey polyline) constructed by connecting midpoints of selected triangle edges. The route lies entirely within the polygon and ensures full visibility while having total length strictly less than half the perimeter.

The example in Fig. 4 also illustrates the practical applicability of our approach. Although the polygon has a complex structure with 24 vertices and numerous reflex angles – resembling environments such as industrial halls or an irregular courtyard of a large private property – it admits a successive dual-edge triangulation. By connecting midpoints of interior triangle edges, we construct a watchman route that ensures full visibility and is provably shorter than half the perimeter. This method serves not only as a theoretical upper bound but also as a constructive algorithm for efficient route planning in scenarios where full geometric detail is available. Compared to strategies that rely on exhaustive perimeter traversal, our approach can substantially reduce route length. However, as visible in the figure, further local improvements may be possible in subregions without reflex vertices, suggesting opportunities for route refinement beyond the triangulation-based construction.

IV. Conclusion

We have analyzed upper bounds on the length of the shortest watchman route in simple polygons under varying structural conditions. For polygons with no reflex vertex, the upper bound is trivial: any point within the polygon suffices, and the shortest watchman route has length zero. Similarly, in polygons with a single reflex vertex, the upper bound remains zero, with the reflex vertex itself serving as a valid watchman position. In polygons with exactly two reflex vertices, we showed that the shortest watchman route can be constructed to have a length strictly less than half the polygon's perimeter, and we further demonstrated that this bound cannot be improved in general. For polygons with three or more reflex vertices, the general upper bound equals the full perimeter, and we likewise showed that this bound is tight in the worst case. However, if the polygon admits a successive dual-edge triangulation, the upper bound can be reduced to less than half the perimeter.

This triangulation not only offers a sharper theoretical bound but also enables a constructive algorithm when the full geometric structure of the polygon is known. It provides a systematic way to plan efficient watchman routes in practical scenarios. Nonetheless, we note that the triangulation-based route may not always be optimal, and local refinements – particularly in convex or reflex-free subregions – could lead to even shorter paths.

V. Acknowledgment

This paper is supported by the grant IG410035 with no. F4/51/2025, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] S. Carlsson, H. Jonsson, and B. J. Nilsson, "Finding the shortest watchman route in a simple polygon," *Discrete & Computational Geometry*, vol. 22, no. 3, p. 377–402, Oct. 1999. doi: 10.1007/pl00009467

[2] X. Tan, "Fast computation of shortest watchman routes in simple polygons," *Information Processing Letters*, vol. 77, no. 1, p. 27–33, Jan. 2001. doi: 10.1016/s0020-0190(00)00146-0

[3] W.-P. Chin and S. Ntafos, "Shortest watchman routes in simple polygons," *Discrete & Computational Geometry*, vol. 6, no. 1, p. 9–31, Mar. 1991. doi: 10.1007/bf02574671

[4] J. S. B. Mitchell, "Approximating watchman routes," in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2013. doi: 10.1137/1.9781611973105.60 p. 844–855.

[5] A. Dumitrescu and C. D. Tóth, "Watchman tours for polygons with holes," *Computational Geometry*, vol. 45, no. 7, p. 326–333, Aug. 2012. doi: 10.1016/j.comgeo.2012.02.001

[6] J. Faigl, "Approximate solution of the multiple watchman routes problem with restricted visibility range," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, p. 1668–1679, Oct. 2010. doi: 10.1109/tnn.2010.2070518

[7] J. Mikula and M. Kulich, "Towards a continuous solution of the d-visibility watchman route problem in a polygon with holes," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, p. 5934–5941, Jul. 2022. doi: 10.1109/lra.2022.3159824

[8] B. J. Nilsson and E. Packer, "Approximation algorithms for the two-watchman route in a simple polygon," *Algorithmica*, vol. 86, no. 9, p. 2845–2884, Jun. 2024. doi: 10.1007/s00453-024-01245-0

[9] L. Štěpánek, F. Habarta, I. Malá, and L. Marek, "A lower bound for proportion of visibility polygon's surface to entire polygon's surface: Estimated by art gallery problem and proven that cannot be greatly improved," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. FedCSIS 2023, vol. 35. IEEE, Sep. 2023. doi: 10.15439/2023f4335. ISSN 2300-5963 p. 1229–1233.

[10] ——, *A Proportion of Visibility Polygon's Surface to the Entire Polygon's Surface: A Lower Bound of the Proportion Derived for General Polygons of Any Shape and Orthogonal Polygons*. Springer Nature Switzerland, 2025, p. 73–98. ISBN 9783031747588

[11] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, "The discretizable molecular distance geometry problem," *Computational Optimization and Applications*, vol. 52, no. 1, p. 115–146, Mar. 2011. doi: 10.1007/s10589-011-9402-6

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 795–800
DOI: 10.15439/2025F0998
ISSN 2300-5963 ACSIS, Vol. 43

A statistical hypothesis test for primality based on random divisor sampling: Principles, properties, adaptive design, and algorithmic analysis

Lubomír Štěpánek[1, 2, 3]
ORCiD: 0000-0002-8308-4304
[1]Department of Statistics and Probability
[2]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 13067 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
[3]Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, 12000 Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

***Abstract*—We propose a statistical hypothesis test for determining whether a given integer $n \geq 2$ is prime. Under the null hypothesis H_0, we assume that n is not a prime (i.e., composite). The test operates by randomly sampling integers from the candidate divisor set $\mathcal{D} = \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ and checking whether any of them divide n. If a proper divisor is found, H_0 is not rejected and n is declared composite. If no divisor is found among an initial set of k_0 samples, additional k samples are drawn, and a p-value is computed based on the probability of missing all actual divisors under H_0. This probability is calculated exactly via a hypergeometric distribution and approximated using an exponential bound. We derive a closed-form upper bound for the minimal number of trials k required to reject H_0 at a given significance level α under the conservative assumption of only one true divisor ($m = 1$). The algorithm has worst-case time complexity $\mathcal{O}(\sqrt{n})$, matching that of classical trial division, but its expected runtime is substantially lower when n has multiple divisors. The proposed test is simple, statistically interpretable, and well-suited both as an educational tool and as a lightweight probabilistic pre-check in layered primality testing pipelines.**

I. Introduction

THE problem of determining whether a given integer n is a prime number is a fundamental question in number theory and theoretical computer science. Primality testing has widespread applications in modern cryptography (such as RSA encryption [1]), secure key generation, error-correcting codes, hashing schemes, and symbolic computation. In many such contexts, fast and reliable primality tests are crucial for both correctness and efficiency [2].

A straightforward approach to primality testing is to check whether any integer in the range $\{2, 3, \ldots, n-1\}$ divides n, or, more effectively, the range of possible divisors is firstly reduced to $\{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ and each one is tested whether divides n. If no such divisor is found, then n is declared prime. While conceptually simple, this method becomes computationally expensive for large n, as it requires up to $\mathcal{O}(\sqrt{n})$ divisibility tests in the worst case. This naive method does not scale well for applications involving large integers or repeated primality checks.

Several more advanced primality tests have been developed, many of which rely on algebraic or number-theoretic properties. Notable examples include the following.

- The *Fermat primality test*, based on Fermat's little theorem, assumes $H_0 : n$ is prime, and checks whether $a^{n-1} \equiv 1 \pmod{n}$ for a random base a [3]. Failure to satisfy this congruence implies n is composite, but passing it does not guarantee primality due to pseudoprimes and Carmichael numbers [4].
- The *Miller–Rabin test*, a widely used probabilistic test, strengthens the Fermat test by testing additional congruences derived from factorization of $n - 1$. It is efficient and has a controllable error probability [5].
- The *AKS primality test*, a deterministic polynomial-time algorithm, decides primality unconditionally. However, it is relatively complex and not competitive in practice for large inputs [6].

While these methods are powerful, most of them assume the null hypothesis that n is prime and look for algebraic contradictions [7]. In this paper, we propose a novel *statistical hypothesis test* for primality that takes a fundamentally different approach: we assume the null hypothesis H_0 that n is

not prime, and assess how "surprising" the observed outcomes of random divisor tests are under this assumption.

Specifically, we randomly sample integers from the set $\{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ and test whether they divide n. If a proper divisor is found, we do not reject H_0 and declare that n is composite. If no divisor is found, we compute a p-value based on the probability of missing all actual divisors under H_0, using both exact combinatorial formulas and exponential approximations. We also derive closed-form bounds for the minimum number of integers that must be tested as potential divisors to achieve a given significance level α, and we analyze the algorithmic complexity of this test in comparison to the classical exhaustive approach.

The proposed test is simple to implement, statistically interpretable, and often requires significantly fewer checks than exhaustive search – especially when n has many proper divisors. It can serve as a practical and educational tool, or as a component in probabilistic or layered primality testing frameworks.

II. Preliminaries

This section establishes basic mathematical facts used in the formulation and justification of the proposed test.

We begin by showing that it is sufficient to restrict our attention to potential divisors of n that are less than or equal to $\lfloor\sqrt{n}\rfloor$.

Lemma 1 (Sufficiency of checking divisors up to $\sqrt{n}$)**.** *Let $n \geq 2$ be an integer. Then n is composite if and only if there exists a proper divisor d such that $2 \leq d \leq \lfloor\sqrt{n}\rfloor$.*

Proof. Let $n \geq 2$ be arbitrary. Suppose first that n is composite. Then there exist integers d, q such that $n = d \cdot q$ with $2 \leq d < n$ and $2 \leq q < n$. Without loss of generality, assume $d \leq q$. Then it is $d \cdot d \leq d \cdot q$, but $d \cdot q = n$, so $d^2 \leq d \cdot q = n$, or simply $d^2 \leq n$. Consequently, it is $d \leq \sqrt{n}$, and since $d \in \mathbb{N}$, it is also $d \leq \lfloor\sqrt{n}\rfloor$. So, d is a proper divisor of n and satisfies $2 \leq d \leq \lfloor\sqrt{n}\rfloor$, the claim follows.

Conversely, suppose there exists a proper divisor d of n such that $2 \leq d \leq \lfloor\sqrt{n}\rfloor$. Then $d \mid n$, so n is not a prime. Hence, n must be composite. □

The following lemma formalizes the intuitive fact that, when sampling from a finite population with few successes, drawing without replacement decreases the chance of obtaining no successes at all compared to drawing with replacement.

Lemma 2 (Probability of no success is lower or equal without replacement)**.** *Let $N \geq 1$ and $1 \leq M < N$ be integers representing a population of N items, M of which are marked as successes. Consider drawing k items from this population with $1 \leq k \leq N - M$. Then the probability of drawing zero successes without replacement is less than or equal to the probability of drawing zero successes with replacement.*

Proof. The probability of no success in k draws *with replacement* is given by

$$P_{\text{with}}(0) = \left(1 - \frac{M}{N}\right)^k. \tag{1}$$

The probability of no success in k draws *without replacement* is

$$P_{\text{without}}(0) = \frac{\binom{N-M}{k}}{\binom{N}{k}} = \prod_{i=0}^{k-1} \left(1 - \frac{M}{N-i}\right). \tag{2}$$

Since for each i in $0 \leq i \leq k-1$ we have $N - i \leq N$, it follows that $\frac{M}{N-i} \geq \frac{M}{N}$, and thus

$$\left(1 - \frac{M}{N-i}\right) \leq \left(1 - \frac{M}{N}\right).$$

Consequently, the product of k such terms satisfies

$$\prod_{i=0}^{k-1} \left(1 - \frac{M}{N-i}\right) \leq \prod_{i=0}^{k-1} \left(1 - \frac{M}{N}\right)$$

$$\prod_{i=0}^{k-1} \left(1 - \frac{M}{N-i}\right) \leq \left(1 - \frac{M}{N}\right)^k$$

$$P_{\text{without}}(0) \overset{(1,2)}{\leq} P_{\text{with}}(0).$$

Hence, the probability of no success is lower than or equal to when drawing without replacement. □

III. Proposed methodology

We now introduce a statistical hypothesis testing procedure for primality, based on the idea of randomly sampling potential divisors of a given integer n. The method is built upon a probabilistic interpretation of missing actual divisors under the assumption that n is composite. In addition to defining the logic, hypotheses, and operational steps of the proposed test, we derive exact and approximate expressions for the resulting p-value and analyze the algorithm's time complexity in comparison to classical exhaustive methods.

A. Test logic and procedure

Let $\mathcal{D} = \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ be the set of candidate proper divisors of $n \in \mathbb{N}$ with $n \geq 2$ large. Using Lemma 1, we know that investigating potential divisors not lower than or equal to $\lfloor\sqrt{n}\rfloor$ is sufficient to determine primality. Denote its cardinality by $|\mathcal{D}|$, where obviously

$$|\mathcal{D}| = |\{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}| = \lfloor\sqrt{n}\rfloor - 1. \tag{3}$$

If n is composite, then it must have at least one divisor in $\mathcal{D}$. This observation motivates a statistical hypothesis test based on sampling elements from $\mathcal{D}$ and checking whether they divide n.

We define the null and alternative hypotheses as follows.

H_0 : n is not a prime, i.e., there exists at least one $d \in \mathcal{D}$ such that $d \mid n$.

H_1 : n is a prime, i.e., no element of $\mathcal{D}$ divides n, or also $\forall d \in \mathcal{D} : d \nmid n$.

The test consists of the following steps.

(i) We randomly select $k_0 \geq 0$ integers from the set $\mathcal{D}$ without replacement, and test each to see whether it divides n.
- If any sampled value divides n, we do not reject H_0 and declare that n is composite.
- If no divisor is found, we proceed to the next step.

(ii) We randomly select an additional $k \geq 1$ integers from the remaining elements of $\mathcal{D}$, again without replacement, and test them for divisibility.
- If any sampled value divides n, we do not reject H_0 and declare that n is composite.
- If no divisor is found, we compute the p-value under H_0.

(iii) We define the p-value as the probability that none of the k additional randomly selected integers divide n, under the assumption that n is composite and has $m \geq 1$ actual divisors in $\mathcal{D}$. Although the exact value of m is unknown in practice – if it were known, we could straightforwardly decide whether n is prime – we can still compute valid upper bounds on this probability. The smaller the p-value, the more extreme the observed outcome (i.e., absence of divisors) appears under H_0, and the stronger the evidence against the assumption that n is composite. If the p-value is less than or equal to a given significance level α, we reject H_0 and declare that n is likely a prime.

This method enables a sequential, interpretable, and probabilistically grounded test for primality. The exact and approximate computation of the p-value, as well as guidance for selecting k given a desired confidence level, are discussed in the following subsections.

B. Exact derivation of the p-value

We now derive the exact p-value used in the proposed hypothesis test. This value represents the probability of observing no divisors among a randomly selected subset of potential divisors, under the assumption that n is composite.

Let $\mathcal{D} = \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ be the set of candidate proper divisors of n. Suppose – under H_0 – that n is composite and has exactly $m \geq 1$ proper divisors in $\mathcal{D}$. We assume m is fixed but unknown.

After testing an initial set of $k_0 \ll |\mathcal{D}|$ integers from $\mathcal{D}$ with no observed divisor, we draw an additional k values uniformly at random without replacement from the remaining $|\mathcal{D}| - k_0$ elements. We are interested in the probability that none of these k values divide n, conditioned on the assumption that n is composite and has m proper divisors in $\mathcal{D}$.

Let X denote the number of divisors observed in the additional k draws. Under H_0 and fixed m, the distribution of X follows the hypergeometric distribution,

$$X \sim \text{Hypergeometric}(|\mathcal{D}| - k_0, m, k).$$

Then, the p-value is defined as the probability of observing $X = 0$ under H_0,

$$p\text{-value} = \mathbb{P}(X = 0 \mid H_0) = \frac{\binom{|\mathcal{D}|-m-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}}, \tag{4}$$

or also

$$p\text{-value} = \frac{\binom{|\mathcal{D}|-m-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}} = \prod_{i=0}^{m-1} \frac{|\mathcal{D}| - k_0 - k - i}{|\mathcal{D}| - k_0 - i}. \tag{5}$$

This expression quantifies the likelihood of missing all true divisors in the k additional samples, under the null hypothesis that n is not a prime. Inspecting formulas (4) and particularly (5), we can see that as k increases or m increases, the p-value decreases, providing stronger evidence against H_0. Thus, the exact p-value decreases as the number of actual divisors m increases. That is, the more true divisors exist in the candidate set $\mathcal{D}$, the less likely it is that a random sample of size k misses all of them. Formally,

$$p\text{-value} = \frac{\binom{|\mathcal{D}|-m-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}} \leq \frac{\binom{|\mathcal{D}|-1-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}},$$

or also

$$p\text{-value} \leq \frac{\binom{|\mathcal{D}|-1-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}} \overset{(5)}{=} 1 - \frac{k}{|\mathcal{D}| - k_0} \overset{(3)}{=} 1 - \frac{k}{\lfloor\sqrt{n}\rfloor - 1 - k_0}, \tag{6}$$

for all $m \geq 1$, where $m \leq \lfloor\sqrt{n}\rfloor - 1 - k_0 - k$. Thus, the case $m = 1$ corresponds to the worst-case scenario (i.e., highest possible p-value under H_0), and can be used as a conservative upper bound. This allows the test to remain valid even when the number of actual divisors is unknown. Also, controlling the p-value under $m = 1$ yields a conservative test.

C. Exponential approximation and upper bound of the p-value

While the exact p-value expression in (4) is combinatorially precise, it can be computationally demanding to evaluate, especially when one wishes to determine the minimal value of k for which the p-value falls below a given significance level α, particularly under the worst-case assumption $m = 1$.

To enable more practical reasoning and simplify the estimation of the required sample size k, we now derive an upper bound on the p-value using an exponential approximation.

Let's assume $k \ll |\mathcal{D}| - m - k_0$, which is natural since k is typically chosen much smaller than $|\mathcal{D}| = |\{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}| = \lfloor\sqrt{n}\rfloor - 1$. Under this assumption, sampling without replacement is numerically close to sampling with replacement. In particular, using Lemma 2, the probability of missing all m true divisors when drawing without replacement is slightly lower than when sampling with replacement, and hence,

$$p\text{-value} = \frac{\binom{|\mathcal{D}|-m-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}} \leq \left(1 - \frac{m}{|\mathcal{D}| - k_0}\right)^k. \tag{7}$$

This bound follows from the fact that each draw without replacement strictly reduces the remaining pool size, increasing the probability of hitting a true divisor. On the other hand, when drawing with replacement, the chance of selecting a true divisor remains constant across draws. As a result, sampling with replacement underestimates the chance of detection, making the bound conservative.

D. Minimum number k of potential divisors to satisfy a given significance level

Having derived an exponential upper bound on the p-value in (7), we now invert this inequality to estimate the minimum number k of integers that need to be checked as potential divisors of n to ensure that the p-value falls below a desired significance level α, where $0 < \alpha < 1$.

Combining formulas (4) and (7) and comparing the result to the significance level α, we obtain

$$p\text{-value} = \frac{\binom{|\mathcal{D}|-m-k_0}{k}}{\binom{|\mathcal{D}|-k_0}{k}} \leq \left(1 - \frac{m}{|\mathcal{D}| - k_0}\right)^k \leq \alpha,$$

and focus on the rightmost inequality to solve for k. Taking logarithms on both sides yields

$$k \cdot \log\left(1 - \frac{m}{|\mathcal{D}| - k_0}\right) \leq \log(\alpha). \tag{8}$$

Since $\log(1-x) < 0$ for $x \in (0,1)$, inequality (8) leads to the following lower bound

$$k \geq \frac{\log(\alpha)}{\log\left(1 - \frac{m}{|\mathcal{D}|-k_0}\right)},$$

or also

$$k \overset{(3)}{\geq} \frac{\log(\alpha)}{\log\left(1 - \frac{m}{\lfloor\sqrt{n}\rfloor-1-k_0}\right)}. \tag{9}$$

This expression provides a closed-form estimate of the number of random checks needed after the initial k_0 samples in order to reject H_0 at significance level α. The most conservative bound arises under the worst-case scenario $m = 1$ (i.e., n has exactly one proper divisor in $\mathcal{D}$). Substituting $m = 1$ into (9) gives

$$k \geq \frac{\log(\alpha)}{\log\left(1 - \frac{1}{|\mathcal{D}|-k_0}\right)},$$

or also

$$k \overset{(3)}{\geq} \frac{\log(\alpha)}{\log\left(1 - \frac{1}{\lfloor\sqrt{n}\rfloor-1-k_0}\right)}. \tag{10}$$

This value of k guarantees that the p-value under H_0 does not exceed α for any possible number of actual divisors $m \geq 1$, thus ensuring a valid and conservative decision threshold even when m is unknown.

IV. Practical usage of the proposed test

The proposed statistical hypothesis test for primality can be applied in two main modes: a classical version with a precomputed sample size k, and an adaptive version that incrementally samples potential divisors until a statistically justified conclusion can be reached. We describe both approaches in this section.

A. Fixed-sample usage with conservative decision threshold

A straightforward way to use the test is to choose a fixed sample size k of candidate divisors in advance, based on a desired significance level α. This guarantees that the computed p-value will fall below α unless a divisor is found, as long as k satisfies the inequality derived in formula (10) under the worst-case assumption that n has only one proper divisor in $\mathcal{D}$ (i.e., when $m = 1$). The steps for such usage are as follows.

(i) The user first selects $k_0 \geq 0$ initial values to test.
(ii) Then, they choose k based on the exponential bound from formula (10) to ensure that the p-value will be sufficiently low if no proper divisor of n is found.
(iii) If any of the tested values divide n, the number is declared composite, since H_0 cannot be rejected.
(iv) If no divisor is found and the computed p-value is below α, then H_0 is rejected, and n is declared probably prime (at the given statistical significance level α).
(v) Otherwise (i.e., if k was chosen too small), the test may be inconclusive: no divisors of n are found in $\mathcal{D}$ (so n may be a prime), but p-value $> \alpha$, and thus H_0 (claiming that n is composite) cannot be rejected.

This fixed-sample approach is simple and deterministic, but may be inefficient if k is much larger than necessary, or inconclusive if k is too small.

B. Adaptive usage with incremental k sampling

An alternative approach is to use the test adaptively: incrementally draw one candidate divisor at a time, and terminate as soon as either a divisor is found (in which case H_0 cannot be rejected), or the p-value's upper bound, calculated using formula (6), drops below α (in which case H_0 is rejected). This version avoids the need to specify k in advance and can potentially stop earlier, depending on the observed evidence.

We present this version formally in Algorithm 1, which assumes a conservative worst-case bound with $m = 1$ and computes the p-value after each additional draw.

This variant can be seen as a probabilistic filter with automatic stopping conditions and is particularly useful when computational efficiency is critical or when n is large and sparsely divisible.

V. Algorithmic Complexity Analysis

We now analyze the computational complexity of the proposed statistical primality test. The analysis is based on the number of basic divisibility checks required to reach a statistically justified conclusion, assuming that a single test of the form $d \mid n$ takes constant time.

Algorithm 1: Adaptive statistical primality test via incremental k sampling

Input : Integer $n \geq 2$; significance level α; initial batch size $k_0 \geq 0$
Output: Decision: "n is `composite`" or "n is `probably prime` (at statistical significance α)"

Let $\mathcal{D} \leftarrow \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$
Randomly sample k_0 integers from $\mathcal{D}$ without replacement
foreach d *in the* k_0*-sample* **do**
  if $d \mid n$ **then**
    return do not reject $H_0 \implies n$ is `composite`
Let $\mathcal{D}' \leftarrow \mathcal{D} \setminus \{\text{already tested values}\}$
Set $k \leftarrow 0$ `// counter of post-initial checks`
while $\mathcal{D}'$ *is not empty* **do**
  Randomly select one $d \in \mathcal{D}'$ and remove it from $\mathcal{D}'$
  if $d \mid n$ **then**
    return do not reject $H_0 \implies n$ is `composite`
  $k \leftarrow k + 1$
  Compute p-value's upper bound:
  $$p\text{-value}_{\text{upper}} \overset{(6)}{\leftarrow} 1 - \frac{k}{\lfloor\sqrt{n}\rfloor - 1 - k_0}$$
  if $p\text{-value}_{upper} \leq \alpha$ **then**
    return reject $H_0 \implies n$ is `probably prime` (at statistical significance α)

A. Search space size

Let $\mathcal{D} = \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ be the set of candidate proper divisors. Its cardinality satisfies

$$|\mathcal{D}| = \lfloor\sqrt{n}\rfloor - 1 = \Theta(\sqrt{n}). \tag{11}$$

A classical exhaustive trial division test checks all values in $\mathcal{D}$, resulting in a worst-case complexity of $\mathcal{O}(\sqrt{n})$.

B. Worst-case bound for the proposed test

Unlike exhaustive search, the proposed statistical test samples values randomly from $\mathcal{D}$ and stops either when a divisor is found or when the computed p-value drops below the significance threshold α.

From formula (9), the minimum number of values k to sample (after initial k_0 values) that guarantees p-value $\leq \alpha$ is

$$k_{\min} = \left\lceil \frac{\log(\alpha)}{\log\left(1 - \frac{m}{|\mathcal{D}|-k_0}\right)} \right\rceil, \tag{12}$$

where m is the number of proper divisors in $\mathcal{D}$. In the worst-case, $m = 1$, yielding

$$k_{\min} = \left\lceil \frac{\log(\alpha)}{\log\left(1 - \frac{1}{|\mathcal{D}|-k_0}\right)} \right\rceil. \tag{13}$$

Using the standard approximation $\log(1-x) \approx -x$ for small $x > 0$ coming from Taylor series, we derive the asymptotic behavior as

$$\begin{aligned}
\Theta(k_{\min}) &= \Theta\left(\left\lceil \frac{\log(\alpha)}{\log\left(1 - \frac{1}{|\mathcal{D}|-k_0}\right)} \right\rceil\right) \overset{(\dagger)}{\approx} \\
&\overset{(\dagger)}{\approx} \Theta\left(\left\lceil \frac{\log(\alpha)}{-\frac{1}{|\mathcal{D}|-k_0}} \right\rceil\right) \approx \\
&\approx \Theta\left(\lceil -\log(\alpha)(|\mathcal{D}| - k_0)\rceil\right) \approx \\
&\approx \Theta\left(\left\lceil \log\left(\frac{1}{\alpha}\right)(|\mathcal{D}| - k_0) \right\rceil\right) \overset{(11)}{\approx} \\
&\overset{(11)}{\approx} \Theta\left(\left\lceil \log\left(\frac{1}{\alpha}\right)\sqrt{n} \right\rceil\right) \approx \\
&\approx \Theta\left(\log\left(\frac{1}{\alpha}\right)\sqrt{n}\right) \approx \\
&\approx \Theta\left(\sqrt{n}\right).
\end{aligned} \tag{14}$$

Thus, the number of divisor checks required to make a valid decision at level α is bounded by $\mathcal{O}\left(\sqrt{n}\right)$.

C. Expected number of checks

In practice, the number of checks is usually smaller due to the following two stopping mechanisms,

(i) a divisor is found early, particularly when m is large;
(ii) no divisor is found, but the p-value falls below α.

Assuming uniform sampling without replacement from $\mathcal{D}$ and a uniform distribution of the m divisors within $\mathcal{D}$, there is, on average, one divisor in every $\frac{1}{m}$-fraction of the set. Therefore, we expect to check approximately $\frac{|\mathcal{D}|}{m} = \frac{\sqrt{n}-1}{m}$ potential divisors. Thus, the expected number of trials until the first divisor is observed is approximately

$$\mathbb{E}[T_{\text{hit}}] = \frac{|\mathcal{D}|}{m} = \frac{\sqrt{n}-1}{m} \approx \frac{\sqrt{n}}{m}. \tag{15}$$

Alternatively, if no divisor is found, the test terminates when $k = k_{\min}$ as given in (12). Therefore, the number of checks is bounded by

$$T(n, m) \overset{(11,12,15)}{\approx} \mathcal{O}\left(\min\left(\frac{\sqrt{n}}{m}, \frac{\log(\alpha)}{\log\left(1 - \frac{m}{\sqrt{n}}\right)}\right)\right).$$

D. Summary

The test adapts to the actual number of divisors: it runs faster when n has many proper divisors. In the extreme case $m = \Theta(\sqrt{n})$, the expected number of checks becomes constant. The worst-case complexity of $\mathcal{O}(\log(1/\alpha) \cdot \sqrt{n})$ arises only when $m = 1$. This adaptivity makes the proposed method significantly more efficient in practice compared to exhaustive search, which always requires $\Theta(\sqrt{n})$ steps.

VI. Illustration of p-value behavior for varying number of divisors

To better understand the behavior of the p-value in the proposed test, we analyze how it changes as a function of the number of random trials k for different values of m, the number of actual proper divisors in the candidate set $\mathcal{D} = 2, 3, \ldots, \lfloor\sqrt{n}\rfloor$. Fig. 1 illustrates this relationship for $n = 10007$, where $|\mathcal{D}| = \lfloor\sqrt{10007}\rfloor - 1 = 99$.

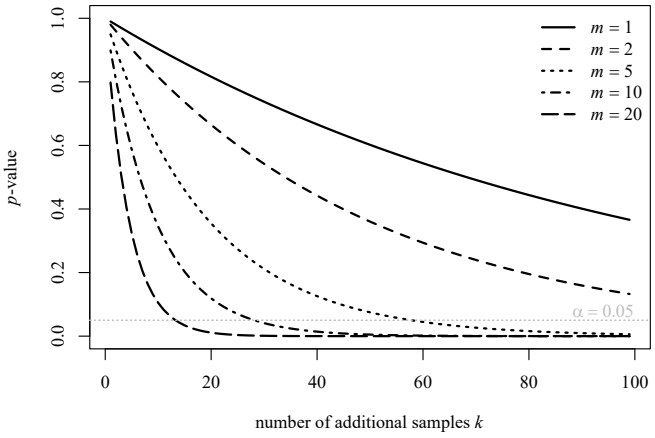


Fig. 1. Decay of the p-value with increasing number k of additionally sampled candidate divisors from $\mathcal{D}$ for various numbers m of true divisors under H_0. The horizontal dotted line marks the significance level $\alpha = 0.05$. Each curve shows the upper bound $\left(1 - \frac{m}{|\mathcal{D}|-k_0}\right)^k$ on the p-value, highlighting how the test becomes more decisive as either m or k increases.

Table I shows corresponding values of the minimum number $k_{\min}$ of samples required to reduce the p-value below the significance threshold $\alpha = 0.05$, based on the exponential bound $\left(1 - \frac{m}{|\mathcal{D}|-k_0}\right)^k$ from formula (7). This bound is conservative, especially for small values of m, and in some cases leads to estimates $k_{\min}$ exceeding the size of the candidate set, i.e., $k_{\min} > \lfloor\sqrt{n}\rfloor - 1$. This inefficiency illustrates that the bound is useful for providing safe, but not necessarily tight, thresholds.

TABLE I
Minimum number of random divisor checks $k_{\min}$ to reject H_0 at significance level $\alpha = 0.05$, for various numbers of actual divisors m, using the exponential bound with $n = 10007$, where $|\mathcal{D}| = \lfloor\sqrt{10007}\rfloor - 1 = 99$.

m	$k_{\min}$ such that p-value ≤ 0.05	is $k_{\min} \leq \lfloor\sqrt{n}\rfloor - 1$?
1	297	no
2	148	no
5	59	yes
10	30	yes
20	15	yes
40	8	yes
75	4	yes

Obviously, the p-value decreases exponentially with k, and this decay is faster when the number of divisors m increases. This demonstrates the adaptivity of the proposed method: when a composite number has more divisors, fewer random checks are typically needed to either discover one or to obtain a statistically significant p-value.

VII. Conclusion

We have proposed and rigorously analyzed a statistical hypothesis test for assessing the primality of a given integer n. The test is based on randomly sampling potential divisors from the set $\mathcal{D} = \{2, 3, \ldots, \lfloor\sqrt{n}\rfloor\}$ and computing a p-value under the null hypothesis H_0 that n is composite. If no divisor is found and the p-value falls below a given significance level α, the test rejects H_0, suggesting that n is likely a prime.

We derived both exact and exponential approximations of the p-value and provided a closed-form expression for the minimum number of additional checks k needed to guarantee a statistically valid conclusion. The exponential upper bound for the p-value decreases as either the number of actual divisors m or the number of random trials k increases. The test allows for an optional initial sample of $k_0 \geq 0$ values, which can be used for early termination or warming up the process, and is fully accounted for in the p-value computations and bounds. Moreover, we analyzed the worst-case and expected algorithmic complexity of the test, showing that the method is adaptive: it performs significantly better when the tested number has more divisors.

Compared to classical exhaustive methods, the proposed approach offers some practical advantages: it is statistically interpretable, simple to implement, and can terminate early in favorable cases. While the worst-case complexity remains $\mathcal{O}(\sqrt{n})$, the average-case performance is often much better, especially for composite numbers.

This paper provides a theoretical introduction to the proposed statistical primality test. Practical evaluations, methodological improvements, simulations, and comparisons with existing methods are part of ongoing work and will be addressed in future research.

VIII. Acknowledgment

This paper is supported by the grant IG410035 with no. F4/51/2025, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978. doi: 10.1145/359340.359342

[2] R. Crandall and C. Pomerance, *Prime Numbers: A Computational Perspective*, 2nd ed. Springer, 2005. ISBN 9780387252820

[3] G. L. Miller, "Riemann's hypothesis and tests for primality," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 300–317, 1976. doi: 10.1016/S0022-0000(76)80043-8

[4] K. H. Rosen, *Elementary Number Theory and Its Applications*, 6th ed. Addison-Wesley, 2011. ISBN 9780321500311

[5] M. O. Rabin, "Probabilistic algorithm for testing primality," *Journal of Number Theory*, vol. 12, no. 1, pp. 128–138, 1980. doi: 10.1016/0022-314X(80)90084-0

[6] M. Agrawal, N. Kayal, and N. Saxena, "Primes is in p," *Annals of Mathematics*, vol. 160, no. 2, pp. 781–793, 2004. doi: 10.4007/annals.2004.160.781

[7] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Duxbury Press, 2002. ISBN 9780534243128

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 801–806
DOI: 10.15439/2025F6497
ISSN 2300-5963 ACSIS, Vol. 43

Estimating the Difficulty of Chess Puzzles by Combining Fine-Tuned Maia-2 with Hand-Crafted and Engine Features

Sebastian Björkqvist
0009-0006-9039-8623
IPRally Technologies Oy
Helsinki, Finland
Email: sebastian@iprally.com

***Abstract*—A common way for chess players to practice tactical awareness is to solve chess puzzles, consisting of an initial position and a sequence of moves to achieve a winning position. This practice is more effective when puzzles are matched to the player's skill level. In this work, we present an approach for estimating the difficulty of a chess puzzle using only the initial position and the sequence of correct moves. Our approach uses a fine-tuned modification of the Maia-2 model combined with a set of hand-crafted features and features extracted from chess engines such as Leela Chess Zero and Stockfish. All of these features are then used as input to a gradient boosted decision tree model that predicts the final rating of the puzzle. We applied our approach to the *FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty Part 2*, where it achieved first place.**

***Index Terms*—chess, chess puzzle, machine learning, transformer, neural network, gradient boosted decision tree**

I. Introduction

SOLVING chess puzzles, where a player is presented a specific chess position and needs to find a sequence of moves leading to a winning advantage, is an important part of any chess player's training routine. Solving puzzles allows the player to learn important tactical patterns much more quickly compared to just playing chess games, since a puzzle always contains a critical position, and similar patterns can thus be practiced far more frequently than they would appear in actual games.

Chess puzzles are widely available on online platforms [1], [2], [3] and may be automatically created by analyzing online games and finding critical positions where a specific sequence of moves leads to a winning position for one side [4]. While the creation of puzzles can be automated, the process of determining the difficulty remains mostly dependent on human input. The puzzle difficulty is usually estimated by having a number of different people attempt the puzzle and measuring the success rate. This information can then be used to calculate the puzzle rating, using for instance the Elo [5] or Glicko-2 [6] algorithm. The drawback of this process is that determining the puzzle rating accurately is quite time-consuming, requiring the input of multiple human solvers. Automating the process of determining the difficulty would allow newly included puzzles to be immediately shown to players of suitable strength, which would improve the puzzle solving experience.



Fig. 1. An example of a chess puzzle. The opponent (black pieces), just made the move `Qd8xd4`, capturing a pawn (red circle), leaving the black queen undefended. Thus the puzzle solver (white pieces) can make the move `Bd3xb5`, capturing the black pawn (green arrow) and checking the black king (green circle). After the opponent captures the bishop (red arrow) to escape check, the white queen captures the black queen with the move `Qd1xd4` (blue arrow). In the final position black is without a queen while white has only lost a bishop, leaving white with a decisive material advantage.

The FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty Part 2 [7] is centered around automating the task of predicting the difficulty of chess puzzles using only the initial position and the sequence of moves, without the need for repeated human solving. The competition is a follow-up to the IEEE Big Data Cup 2024 Challenge on Predicting Chess Puzzle Difficulty [8]. The second edition of the competition introduces more chess puzzles in the training dataset. Additionally, the training data features also include success probabilities computed using the Maia-2 [9] model.

In this work, we introduce an approach for estimating the

difficulty of chess puzzles which adapts and fine-tunes Maia-2, originally trained to predict the probability of humans making certain moves in a specific position. We also utilize ideas from the approaches of the first [10] and second place [11] teams from the first edition of the competition, as well as extend our own previous method [12] which placed third. Our new approach was applied to the second edition of the challenge, achieving first place.

II. Related work

While the estimation of chess puzzle difficulty had been analyzed on a small scale previously [13], [14], the main body of work related to the problem of predicting the difficulty of chess puzzles is due to the IEEE Big Data Cup 2024 Challenge on Predicting Chess Puzzle Difficulty [8], which we later refer to as the *first edition* of the competition. The top performing teams of the first edition used the following methods:

1) The winning team [10] used multiple Maia-1 [15] and Leela Chess Zero [16] models to encode the chess positions in a puzzle into vectors and then used an RNN to combine these position embeddings into one puzzle embedding, which was used to predict the puzzle rating.
2) The second place team [11] used a CNN architecture similar to Leela Chess Zero to encode each position in the puzzle and trained the model from scratch to predict the rating. The model also predicted the next move in the position and used the loss of that prediction as an additional feature when predicting the rating.
3) The third place team [12] (our submission to the first edition) used hand-crafted features and features extracted from chess engines, as well as a rating predicted by a residual neural network as input to a LightGBM [17] model to predict the puzzle rating.

Other approaches used in the first edition of the competition include a RankNet-based deep neural network model [18], a Transformer-based approach using a modified version of ChessFormer [19] and other CNN-based solutions [20], [21].

III. Chess Puzzle Challenge Dataset

The dataset of chess puzzles available for training provided by the FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty Part 2 [7] consists of approximately 4.55 million chess puzzles obtained from the Lichess open database [22]. An example puzzle can be seen in Figure 1. Each puzzle in the training dataset contains an initial position given in FEN notation [23] and the sequence of the correct moves of the puzzle in long algebraic notation [24]. Additionally, the training dataset contains metadata such as rating deviation, attempt count, popularity, tactical themes, and opening info. A new addition to the second edition of the competition is the inclusion of 22 success probability predictions, describing how likely humans of varying playing strength are to correctly solve the puzzle. The target variable is the rating of the puzzle, determined by repeated solving of the puzzle in the Lichess trainer [1]. The distribution of ratings is shown in Figure 2.

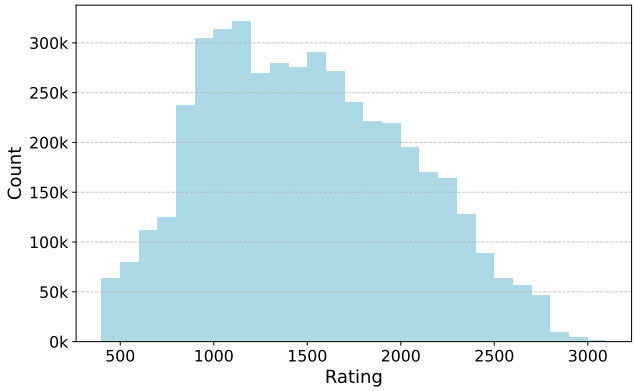


Fig. 2. The distribution of the puzzle ratings in the training set. A large number of puzzles have a rating near the initial puzzle rating of 1500, with another visible concentration of puzzles around 1000.

The competition also provides a public test set containing 2,235 puzzles. The public test set does not contain any of the metadata that is available for the training set — it only has the initial position, the solution moves as well as the predicted human success probabilities. The goal of the competition is to predict accurate ratings for the public test set. The public test set is further split into a holdout set used to compute the preliminary score, and a final test set, used for the final competition ranking.

IV. Approach

Our approach is based mainly on the following observation: The first place solution [10] in the first edition used pre-trained chess engine models to extract position embeddings but did not fine-tune the engine model weights. The second place solution [11], on the other hand, used the model architecture of a chess engine model and trained it end-to-end, but did not initialize the model weights with any pre-trained model. Thus, we wanted to analyze whether these two methods could be combined by both training a model end-to-end and initializing the model with pre-trained weights. The reason why no team did this in the first edition of the competition is likely related to the difficulty of extracting the trained weights from Maia-1 and Leela Chess Zero models for use in a commonly used deep learning library like PyTorch.

Shortly after the first edition of the competition finished, the Maia-2 [9] model was released. This model turned out to be a great candidate to both initialize a model with and to fine-tune, since the model code and weights are freely available, and the model is written in PyTorch, making it easy to modify and extend. Thus, we decided to use the Maia-2 model as the backbone of our approach.

We also wanted to verify whether the use of hand-crafted features and features extracted from other chess engines is still useful even when fine-tuning a deep learning model to predict the ratings, so we decided also to include and extend the features used in our third place solution for the first edition [12].

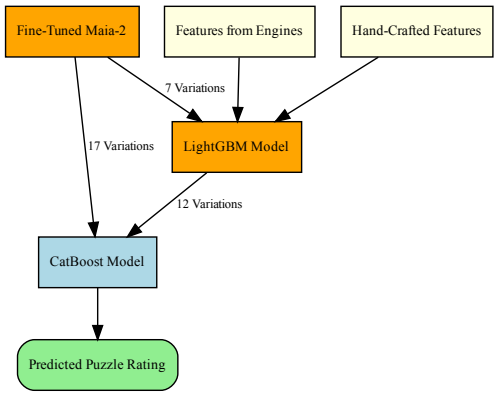


Fig. 3. Flowchart showing the basic structure of the method to predict the Glicko-2 rating for chess puzzles. Ratings are first predicted by a fine-tuned Maia-2 model as well as a LightGBM model. Multiple variations of both models are trained, and the predicted ratings are given as input to a final CatBoost model that ensembles the ratings into a final prediction. The different models and features are described in detail in Section IV.

An overview of our approach is shown in Figure 3. We first train a fine-tuned Maia-2 model to predict the rating of puzzles, and use this rating as well as hand-crafted and chess engine features as input to a LightGBM model to predict the puzzle rating. Finally, we use ratings predicted by multiple variants of the fine-tuned Maia-2 and the LightGBM model as input to a final CatBoost ensemble, which then predicts the final puzzle rating. For the predictions submitted to the competition, we additionally perform a linear scaling to account for the differences in the rating distribution between the training set and the public test set.

A. Fine-Tuned Maia-2

1) Architecture: Maia-2 [9] is a model trained to predict the probability of a human making a specific move in a certain chess position. Maia-2 consists of a ResNet backbone that embeds a position to a set of patch embeddings, and of a Transformer model that takes the patch embeddings as input and outputs a *position embedding* encoding relevant information about the board state. The Transformer layers incorporate skill-aware attention which allows the model to see information about the skill level of the players when predicting the next move. The position embedding is used as input to three different heads: The *move head* predicting the probability of each possible next move, the *auxiliary head* predicting, among others, the legal moves and move-specific information such as whether the move is a check or capture, and the *value head* predicting the result of the game.

An overview of how we use the Maia-2 model is shown in Figure 4. We utilize the model for predicting the puzzle rating as follows: First, for each position in the puzzle (up to and including the fifth move), we compute the position embedding using the ResNet backbone and the Transformer model. We inject a constant skill value to the Transformer model since puzzle solvers are unknown.

Additionally, for each position in the puzzle, we compute the move and auxiliary predictions by passing the position embedding to the heads. Then, we compute the cross-entropy loss of these using the ground truth (which is known because all the puzzle moves are available), following how move predictions were used in [11]. Finally, we embed the correct move in the position using a two-layer MLP. The final embedding of the position is a concatenation of the Maia-2 position embedding and the move and auxiliary prediction losses, as well as the correct move embedding. We also experimented with using the Maia-2 value head here, but it did not improve the rating predictions, so we ultimately left it out.

To obtain a single embedding for an entire puzzle, we pass the final position embedding sequence into a bi-directional GRU model (with two layers and a hidden size of 512) to obtain a single 1024-dimensional embedding for the entire puzzle, similarly as in [10]. The main difference from their solution is that we use the Maia-2 model instead of Maia-1 and Leela, and we also fine-tune the entire model instead of keeping the position embedder frozen.

The final predicted rating is obtained by passing the puzzle embedding given by the GRU to a final linear layer. This layer predicts the rating of the puzzle, normalized to zero mean and unit variance.

As an additional variant, we also include a set of other features as input to the final linear layer. These features are a subset of the ones used as input to the LightGBM model (described in section IV-B). These features are embedded using a residual neural network, and the output embedding is concatenated with the puzzle embedding and given as input to the final linear layer.

2) Training: The Maia-2 model is implemented using *PyTorch* [25], so we also implement our additional layers using PyTorch and use it to train the model. We initialize the Maia-2 model with the pre-trained rapid or blitz weights. The entire model is trained end-to-end, meaning that also the pre-trained Maia-2 layers change during training. We use mean squared error (MSE) loss for the predicted rating, and we additionally also compute the cross-entropy loss for the move and auxiliary predictions from Maia-2 to keep them from diverging, with the cross-entropy losses having significantly lower weight than the rating MSE loss. We also weigh the loss of each puzzle by the inverse of its rating deviation to prevent the model from paying too much attention to puzzles that have a very uncertain rating.

We use the Schedule-Free AdamW optimizer [26] to train our model, which allows us to use a fixed learning rate for the entire training. The model is trained using approximately 4.1 million puzzles. We use a batch size of 128 puzzles and evaluate the model every 5,000 iterations. The training is stopped when the validation loss hasn't improved for three consecutive evaluations.

A single NVIDIA RTX 4070 GPU is used to train the model, and we run the training using bfloat16 automatic mixed

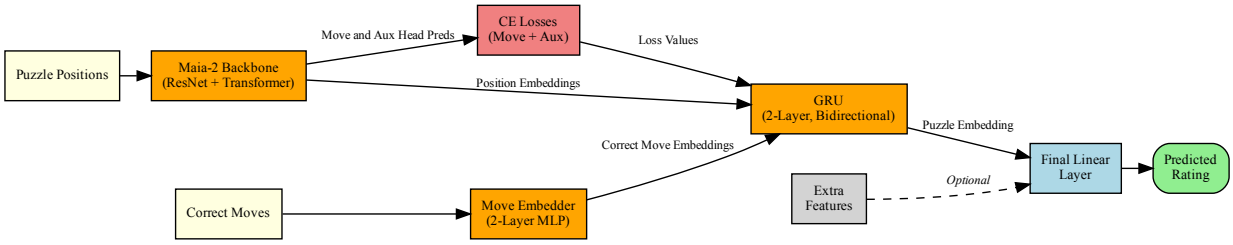


Fig. 4. Diagram describing how the fine-tuned and customized Maia-2 model for predicting puzzle ratings. First, each position in the puzzle is encoded using the Maia-2 backbone. Then, the predictions of the move and auxiliary heads are used to compute a loss using the ground-truth values. Additionally the correct move in the position is embedded using an MLP. The position embeddings, loss values and move embeddings of a puzzle are then combined into a single puzzle embedding using a GRU model. The predicted rating is then computed using a final linear layer, which also may obtain as input additional hand-crafted or engine features. The model is trained end-to-end with all layers being trained or fine-tuned simultaneously. For more details, see Section IV-A.

precision to speed up the training and reduce memory usage. A single training run takes about two hours. Multiple variants of the model are trained with different hyperparameters (dropout, weight decay, learning rate, loss weights) and slightly different model configurations and weights (Maia-2 rapid or blitz).

B. LightGBM model

The next step in our rating prediction pipeline is training a gradient boosted decision tree model using *LightGBM* [17] that receives as inputs three different types of features:

1) Ratings predicted by the fine-tuned Maia-2 model, as described above in section IV-A
2) Features extracted from various chess engines such as Leela, Maia-1 and Stockfish
3) Hand-crafted features created from the initial position and the solution moves

The LightGBM model is trained with approximately 4.4 million puzzles using MSE loss. The training is stopped if the validation loss doesn't improve in 50 iterations. Multiple different versions of the model are trained using slightly different feature sets and hyperparameters (like different dropout and bagging factor). We chose to use LightGBM due to its fast training speed when using a large number of features and puzzles. In the final ensemble we instead used CatBoost [27], which is slower but results in a slightly more accurate model.

1) Ratings predicted by fine-tuned Maia-2: The LightGBM model uses as input the rating predicted by seven different versions of the fine-tuned Maia-2 model. For six of these versions we trained the Maia-2 model multiple times using different folds, and then predicted the rating for the training set with the model where the fold was not used for training, following our previous approach [12]. We also included the rating of one fine-tuned Maia-2 model that was trained only once due to time constraints, meaning the LightGBM model retrieved ratings predicted on the train set of the Maia-2 model. We added Gaussian noise to these ratings to make sure the model didn't overfit to this specific feature.

2) Features extracted from chess engines: The main feature extracted from chess engines is the probability that the correct puzzle move is made in different positions during the puzzle, as in [12]. The probability is computed for the first five player and opponent moves, with a default value of 1.0 being used if the puzzle is shorter than five moves. We compute the probability using the engine evaluation function only, without any search being performed. This feature is obtained using the following models:

- All Maia-1 models (1100, 1200, ..., 1900) [15]
- Leela "best nets" T1-256x10 and T1-512x15x8h [28]
- BadGyal-8, GoodGyal-5, GoodGyal-7 and TinyGyal-8 [29]
- Additional Leela networks LD2, J104.1-30, T73_RL.2-10000 and J64-180 [30]

We also use Stockfish 17 [31] to extract the material, positional, and total evaluations according to its evaluation function. Stockfish is also used to compute the final centipawn value after the last move of the puzzle, as well as to compute whether the correct move is found on a certain depth. For this purpose we used depths 1 through 100,000 in powers of ten.

Additionally we use the success probabilities provided by the competition organizers in the original dataset.

3) Hand-crafted features: We extract three feature families using *python-chess* [32] and *pandas* [33]: puzzle-level basic features, move/position features, and tactical features.

a) Puzzle-level basic features: These features are computed once per puzzle: solution length; is the player white?; Lichess themes from the puzzle tagger [4]; initial piece placement (board array) before the first player move.

b) Move/position features: These features are computed for the first five moves by each side, and are listed in Table I.

c) Tactical features: Features indicating tactical motifs, described in Table II. Unless noted, computed for the first five moves by each side.

C. Final CatBoost ensemble

To obtain the final predicted rating, we train a gradient boosted tree model using *CatBoost* [27] that receives as input

TABLE I
MOVE- AND POSITION-SPECIFIC FEATURES

Category	Features
Position state	# legal moves; in-check flag
Forcing moves	# checking moves / pieces; # capturing moves / pieces; captures winning material; materially safe captures
Correct move	Check flag; moving piece mobility; captures undefended piece; material gain; piece type; from/to column and row
Material and structure	Side material and material diff.; mobility by piece type; undefended pieces (# and value); attackers/defenders near each king
Pawn structure	Pawn islands; doubled / isolated / passed pawns
Other	Over-/under-defended pieces; castling status (both sides); castling/en passant flags

TABLE II
TACTICAL FEATURES

Feature	Notes
Accepted sacrifice	Correct move accepts a sacrifice
Interposition defence	Correct move blocks a check
Mate threat	First five *player* moves only; correct move poses a mate threat, detected using Stockfish assuming opponent skips a turn
Recapture	First five *player* moves only; correct move recaptures previously lost piece

the predictions of 17 different versions of the fine-tuned Maia-2 model, as well as 12 different versions of the LightGBM model. Additionally, the model uses as input the number of moves in the puzzle solution and the probability of making the correct puzzle move according to the following Leela nets: T3-512x15x16h [30], T82-768x15x24h and BT3-768x15x24h [28]. We include these probabilities in this late stage instead of in the larger LightGBM model due to the computation of the probabilities with these models being quite slow.

The CatBoost model is trained using 100,000 puzzles, with MSE loss. We note that using a gradient boosted tree model for ensembling gives slightly better results compared to using just a simple average or learned weighted average.

D. Scaling of predicted ratings

Since the public test set puzzles have on average fewer attempts than the ones in the training set [8], they are also likely to have fewer extremely high and low ratings, due to puzzles with such extreme true ratings requiring many solving attempts to converge. To account for this difference, we follow the method used in [11] and apply linear scaling to align the predicted ratings more closely with the estimated mean $c = 1600$ of the test set distribution using the formula

$$\hat{y}_{scaled} = c + g * (\hat{y} - c),$$

where g is the scaling factor. The mean value was found by weighing our own test set (extracted from the training data) to better match the distribution of the public test set (similarly to how we weighed our own validation set in the first edition [12]). The raw predicted ratings we used for the final submissions in the FedCSIS 2025 competition all came from the same final CatBoost ensemble. We then applied the scaling using factors g of 0.7, 0.8 and 0.9 to obtain three different variations of the ratings.

V. RESULTS

TABLE III
FINAL FEDCSIS 2025 COMPETITION RESULTS (TOP 5 TEAMS OUT OF 31)

Rank	Prelim. MSE	*Final MSE*	Rel. diff.
1	55,382	*52,311*	-
2	58,186	*54,377*	+3.9%
3	58,892	*55,938*	+6.9%
4	61,524	*57,492*	+9.9%
5	67,062	*61,045*	+17%

In the FedCSIS 2025 Challenge, the predicted puzzle ratings are evaluated on the public test set by comparing the predicted puzzle ratings to the ground truth ratings and computing the MSE loss. Our method achieved a preliminary score of 55,382, and a final score of 52,311, resulting in first place in the final results (as well as on the preliminary leaderboard). The scores of the top five teams (out of a total of 31 teams with a final score) are shown in Table III. The final score of the second place team is 3.9% higher than our score, and all of the top four teams are within 10% of our score. This indicates that the second edition of the competition was much closer than the first edition, where the difference between the first and second place teams was around 15% [8].

TABLE IV
RESULTS OF DIFFERENT VARIATIONS OF OUR MODEL

Model	Test set MSE	Rel. diff.
Final CatBoost ensemble	35,400	-
LightGBM (all feats)	36,400	+2.9%
LightGBM w/o hand-crafted feats	36,800	+4.1%
LightGBM w/o engine feats	37,300	+5.4%
Fine-tuned Maia-2 (with extra feats)	38,900	+10%
Fine-tuned Maia-2 (w/o extra feats)	40,800	+15%
LightGBM w/o fine-tuned Maia-2	52,600	+49%

We also performed an ablation study to verify the impact of different parts of our solution, using a test set of 50,000 puzzles extracted randomly from competition training set. The results of the ablation study are shown in Table IV. From the results we can see that the fine-tuned Maia-2 model is critical to our solution, since the LightGBM model without these ratings has an MSE that is 49% higher than our final solution. We also see that the final CatBoost ensemble is useful, since without it the loss is 2.9% higher.

The ablation study also confirms that both the chess engine features and hand-crafted features are still useful. Additionally we see that even though the fine-tuned Maia-2 model is critical for our model, by itself it performs around 15% worse than

our final ensemble, which would place it close to the fifth place finisher, assuming the results transfer to the competition public test set.

VI. Conclusion

In this work, we described an approach for predicting the difficulty of a chess puzzle based on fine-tuning the Maia-2 model and combining the predicted rating from this model with hand-crafted features as well as features extracted from chess engines. We showed that fine-tuning the pre-trained Maia-2 model achieved strong performance for the puzzle rating prediction by itself, and that hand-crafted and chess engine features are still a useful addition to this deep learning-based method. We demonstrated the effectiveness of the method, achieving first place in the FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty.

References

[1] Lichess.org, "Lichess training," https://lichess.org/training, 2025, accessed: 2025-08-16.

[2] Chess.com, "Chess.com puzzles," https://www.chess.com/puzzles, 2025, accessed: 2025-08-16.

[3] ChessTempo, "Chesstempo," https://chesstempo.com/, 2025, accessed: 2025-08-16.

[4] T. Duplessis, "lichess puzzler," https://github.com/ornicar/lichess-puzzler. [Online]. Available: https://github.com/ornicar/lichess-puzzler

[5] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Pub., 1978. ISBN 0668047216 9780668047210. [Online]. Available: http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216

[6] M. E. Glickman, "Example of the glicko-2 system," http://www.glicko.net/glicko/glicko2.pdf, 2022, published: 2022-03-22, accessed: 2025-08-16.

[7] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. doi: 10.15439/2025F5937. [Online]. Available: http://dx.doi.org/10.15439/2025F5937

[8] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "Ieee big data cup 2024 report: Predicting chess puzzle difficulty at knowledgepit.ai," in *IEEE International Conference on Big Data, Big Data 2024, Washington DC, USA, December 15-18, 2024*. IEEE, 2024.

[9] Z. Tang, D. Jiao, R. McIlroy-Young, J. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A unified model for human-ai alignment in chess," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 919–20 944, 2024.

[10] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji team's submission to the ieee bigdata 2024 cup: Predicting chess puzzle difficulty challenge," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8415–8422.

[11] A. Schütt, T. Huber, and E. André, "Estimating chess puzzle difficulty without past game records using a human problem-solving inspired neural network architecture," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8396–8402.

[12] S. Björkqvist, "Estimating the puzzlingness of chess puzzles," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8370–8376.

[13] S. Stoiljkovikj, I. Bratko, M. Guid, and F. UNI, "A computational model for estimating the difficulty of chess problems," in *Proceedings of the third annual conference on advances in cognitive systems ACS*, 2015, p. 7.

[14] D. Hristova, M. Guid, and I. Bratko, "Assessing the difficulty of chess tactical problems," *International journal on advances in intelligent systems*, vol. 7, no. 3, pp. 728–738, 2014.

[15] R. McIlroy-Young, S. Sen, J. Kleinberg, and A. Anderson, "Aligning superhuman ai with human behavior: Chess as a model system," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020. doi: 10.1145/3394486.3403219. ISBN 9781450379984 p. 1677–1687. [Online]. Available: https://doi.org/10.1145/3394486.3403219

[16] The LCZero Authors, "Leela chess zero." [Online]. Available: https://lczero.org/

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017. ISBN 9781510860964 p. 3149–3157.

[18] A. Rafaralahy, "Pairwise learning to rank for chess puzzle difficulty prediction," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8385–8389.

[19] S. Miłosz and P. Kapusta, "Predicting chess puzzle difficulty with transformers," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8377–8384.

[20] D. Ruta, M. Liu, and L. Cen, "Moves based prediction of chess puzzle difficulty with convolutional neural networks," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 8390–8395.

[21] M. Omori and P. Tadepalli, "Chess rating estimation from moves and clock times using a cnn-lstm," in *International Conference on Computers and Games*. Springer, 2024, pp. 3–13.

[22] Lichess.org, "Lichess open database: Chess puzzles," https://database.lichess.org/#puzzles, 2025, accessed: 2025-08-16.

[23] Wikipedia contributors, "Forsyth–edwards notation — Wikipedia," https://en.wikipedia.org/wiki/Forsyth%E2%80%93Edwards_Notation, 2025, accessed: 2025-08-16.

[24] ——, "Algebraic Notation (Chess) — Wikipedia," https://en.wikipedia.org/wiki/Algebraic_notation_(chess), 2025, accessed: 2025-08-16.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: an imperative style, high-performance deep learning library*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[26] A. Defazio, X. Yang, A. Khaled, K. Mishchenko, H. Mehta, and A. Cutkosky, "The road less scheduled," *Advances in Neural Information Processing Systems*, vol. 37, pp. 9974–10 007, 2024.

[27] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018. [Online]. Available: https://arxiv.org/abs/1810.11363

[28] Leela Chess Zero contributors, "Best networks for lc0," https://lczero.org/dev/wiki/best-nets-for-lc0/, 2024, accessed: 2025-07-25. [Online]. Available: https://lczero.org/dev/wiki/best-nets-for-lc0/

[29] dkappe, "Leela chess weights: Bad gyal," https://github.com/dkappe/leela-chess-weights/wiki/Bad-Gyal, 2020, GitHub release, accessed 2025-07-25. [Online]. Available: https://github.com/dkappe/leela-chess-weights/wiki/Bad-Gyal

[30] Leela Chess Zero contributors, "Contributed networks for lc0," https://storage.lczero.org/files/networks-contrib/, 2024, accessed: 2025-07-26. [Online]. Available: https://storage.lczero.org/files/networks-contrib/

[31] The Stockfish developers, T. Romstad, M. Costalba, J. Kiiski, G. Linscott, Y. Nasu, M. Isozaki, and H. Noda, "Stockfish." [Online]. Available: https://stockfishchess.org/

[32] N. Fiekas, "python-chess: a chess library for python," https://github.com/niklasf/python-chess, 2025. [Online]. Available: https://github.com/niklasf/python-chess

[33] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51 – 56.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 807–812
DOI: 10.15439/2025F4532
ISSN 2300-5963 ACSIS, Vol. 43

A Multi-Stage Framework for Chess Puzzle Difficulty Prediction

Ling Cen
Neurospark Lab
Singapore
cen.ling@neurosparklab.tech

Jiahao Cen
University College Dublin
Ireland
cen.jiahao@ucdconnect.ie

Malin Song
Anhui University of Finance and Economics
China
songml@aufe.edu.cn

Zhuliang Yu
South China University of Technology
China
zlyu@scut.edu.cn

***Abstract*—Accurately estimating the difficulty of the chess puzzle is important for adaptive training systems, personalized recommendations, and large-scale content curation. Unlike engine evaluations optimized for perfect play, this task involves modeling human-perceived solving difficulty, typically expressed by Glicko-2 ratings. We present a multi-stage framework developed for the FedCSIS 2025 Challenge. The method trains four rating-banded neural regression models in different Elo ranges to capture localized difficulty patterns and reduce bias from unbalanced data. Their predictions are combined with statistical attributes, including success probabilities, failure distributions, and solution length, through a feature-based regression stage to improve cross-range generalization. A final calibration step adjusts the output to statistically plausible rating levels, mitigating systematic prediction biases without adding computational complexity. An additional mask selection procedure was explored as part of the competition extension to identify 10% of the puzzles that are most likely to benefit from the refined evaluation. The proposed solution ranked 5th on the public leaderboard and 6th in the final standings. These results demonstrate that a lightweight and interpretable regression pipeline can achieve competitive precision in modeling human-perceived chess puzzle difficulty.**

***Index Terms*—Chess puzzle difficulty prediction, Multi-stage framework, Regression, Calibration, Mask selection**

I. Introduction

PREDICTING the difficulty of chess puzzles is a challenging but important problem that integrates machine learning, human cognition, and game theory. The difficulty of the puzzle on platforms such as Lichess[1] is expressed as a Glicko-2 rating[2] that dynamically updates based on the success or failure of the players. Automating this rating prediction accelerates puzzle curation, enables personalized recommendations for players of varying skill levels, and provides insight into which tactical or strategic motifs challenge human solvers the most. This is also the goal of the FedCSIS 2025 Challenge [1] organized on the KnowledgePit platform[3].

A. Related Work

Research on puzzle difficulty prediction has advanced considerably in recent years. Early approaches used handcrafted features with classical machine learning: Björkqvist [2] modeled puzzle "puzzlingness" via positional and tactical indicators, and Rafaralahy [3] applied pairwise learning-to-rank to capture ordinal relations between puzzles. With the rise of deep learning, sequence-based and representation-learning methods became prominent. Ruta et al. [4] applied convolutional neural networks (CNNs) to predict puzzle ratings directly from move sequences, Miłosz and Kapusta [5] proposed Transformer-based models treating puzzles as sequences, and Omori and Tadepalli [6] developed a CNN-LSTM model incorporating moves and timing to jointly estimate puzzle and player ratings.

The IEEE BigData Cup 2024 [7] demonstrated the success of hybrid pipelines that integrate statistical features and learned representations. Woodruff et al. [8] trained neural models with rating-based features, while Schütt et al. [9] introduced a human-problem-solving-inspired architecture combining move distribution analysis and cognitive heuristics.

B. Motivation and Contributions

Despite these advances, two key challenges remain. First, models trained across broad rating ranges often exhibit systematic biases, overestimating simple puzzles and underestimating long tactical sequences due to global error optimization. Second, while models trained on different rating distributions provide complementary perspectives, effectively integrating them to improve generalization is non-trivial.

To address these challenges, we propose a multi-stage framework that first trains four independent neural models on separate rating ranges to capture localized difficulty patterns and reduce bias caused by imbalanced data. Their predictions are then combined with statistical attributes, including success probabilities, failure distributions, and solution length, in a feature-based regression stage to improve cross-range generalization. A final calibration step adjusts the outputs toward statistically plausible rating levels, reducing systematic prediction biases without increasing computational complexity. Additionally, the framework was extended with a mask selection procedure to identify the 10% of puzzles most likely to benefit from refined evaluation, as required in the competition extension.

Our method ranked 5th on the public leaderboard and 6th in the final standings of the FedCSIS 2025 Challenge, demonstrating that a lightweight, interpretable pipeline can achieve competitive accuracy in modeling human-perceived chess puzzle difficulty.

[1] https://lichess.org/
[2] https://en.wikipedia.org/wiki/Glicko_rating_system
[3] https://knowledgepit.ai/

The remainder of this paper is organized as follows. Section II briefly describes the challenge, dataset, and evaluation metric. Section III presents the methodology. Section IV discusses the mask selection task. Section V shows xperimental results. Finally, Section VI concludes the paper and outlines future directions.

II. FEDCSIS 2025 CHALLENGE

This competition continues the established task from IEEE Big Data 2024 [7], asking participants to predict the perceived difficulty of chess puzzles based on puzzle configurations and human solving statistics. The objective remains to estimate the Glicko-2 rating assigned to each puzzle, which reflects the likelihood that players of various skill levels can solve it, using only board position and solution moves.

A. Dataset

The official dataset consists of a large annotated training set and a separate unlabeled test set. The training set contains approximately 4.56 million puzzles, each labeled with a human-derived difficulty rating, while the test set includes 2,235 puzzles sharing the same feature structure but without ratings.

Each puzzle is described by multiple feature groups derived from both game records and engine analysis:

- Core identifiers: a unique puzzle ID, board state in Forsyth–Edwards Notation (FEN)[4], and the solution sequence in Portable Game Notation (PGN)[5].
- Human performance indicators (training only): Glicko-2 rating, rating deviation, popularity, and the number of attempts.
- Contextual metadata (training only): puzzle themes, associated game URLs, and opening tags.
- Engine-based success statistics: estimated success probabilities for players at different Elo levels, provided separately for rapid and blitz time controls (10 columns each), approximating human solving likelihood across skill tiers.

Puzzle ratings range from about 400 (trivial tactics) to over 3000 Elo (complex master-level combinations). The distribution is heavily imbalanced, with most puzzles clustered around intermediate difficulty. No official validation set is provided; participants typically constructed small internal validation splits for hyperparameter tuning and model selection.

A training sample (puzzle) of the initial state of the chessboard decoded by FEN and the rating is illustrated in Figure 1.

B. Evaluation Protocol

Predictions are evaluated using the mean squared error (MSE) between the predicted and actual ratings:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2,$$

[4]https://en.wikipedia.org/wiki/ForsythEdwards_Notation
[5]https://en.wikipedia.org/wiki/Portable_Game_Notation



Figure 1. Example of the chess puzzle Rating 1902: initial state of the chessboard decoded by FEN: "r6k/pp2r2p/4Rp1Q/3p4/8/1N1P2R1/PqP2bPP/7K b - - 0 24".

where N is the number of puzzles in the test set. The test set is partitioned internally, with 10% used for provisional public leaderboard updates and the remaining 90% for final ranking. Only the public subset is visible during the competition.

C. Optional Mask Extension

An additional post-evaluation extension allows participants to submit a binary mask marking 10% of test puzzles predicted to have the highest potential benefit from re-evaluation. For these selected puzzles, predictions are replaced by ground-truth ratings, and MSE is recalculated. While this mask-based scoring does not affect the main leaderboard, it serves as an auxiliary measure of a model's ability to recognize cases where its predictions are less reliable.

III. METHODS

Our approach employs a multi-stage pipeline that integrates localized learning of puzzle difficulty with statistical adjustment to correct systematic biases. The design philosophy stems from two key observations. First, the relationship between puzzle features—such as tactical motifs, material configurations, and success probabilities—and human-perceived difficulty is not uniform across the rating spectrum; a single global model tends to bias its predictions toward the overrepresented mid-range puzzles. Second, models trained purely to minimize global mean squared error often fail to respect structural regularities observed in human solving behavior, particularly for very easy or very hard puzzles. We therefore decompose the task into specialized components: localized rating-band models to better learn within-range patterns, a meta-learning step to integrate complementary predictions and additional structural attributes, a calibration step to align outputs with statistical difficulty indicators, and a final averaging procedure to stabilize predictions.

Figure 2 outlines this four-stage pipeline. The first stage trains multiple models specialized for different rating bands to disentangle the heterogeneous difficulty distribution. In the second stage, their predictions are combined with auxiliary

features using tree-based meta-regressors trained on a fixed validation subset. The third stage applies a statistical calibration that shifts the predictions toward historically plausible difficulty levels by exploiting failure distributions and their higher-order characteristics. Finally, the calibrated regressors are averaged to reduce variance and further improve generalization.

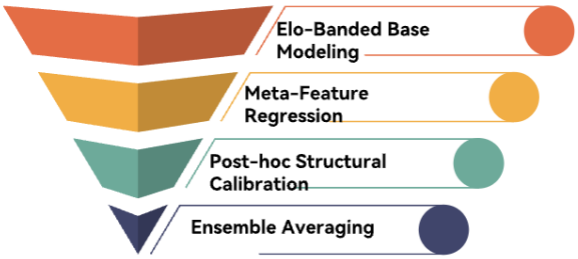


Figure 2. Overview of the proposed prediction pipeline.

A. Rating-Banded Base Models

The first stage employs a rating-banded learning strategy to address heterogeneous solving patterns across Elo levels. Lower-rated puzzles, often simple mate-in-one or basic forks, exhibit solving distributions heavily concentrated in low-skill buckets, whereas higher-rated puzzles show sparse but sharper failure transitions at advanced skill levels. A single global regressor tends to minimize global loss by fitting the over-represented mid-range, which leads to biased underestimation for high-rated puzzles. Inspired by the band-wise modeling idea explored in prior work [8], we explicitly partition the data into four difficulty bands and train separate regressors for each band to focus on localized difficulty dynamics.

Let the global training set be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the structured feature vector and y_i the true difficulty rating. We define four disjoint subsets:

$$\mathcal{D}_b = \{(x_i, y_i) \in \mathcal{D} \mid L_b \le y_i < U_b\}, \quad b \in \{1, 2, 3, 4\}, \tag{1}$$

with rating intervals:

$$[L_b, U_b) \in \{[0, 1000),\ [1000, 1400),\ [1400, 1700),\ [1700, \infty)\}. \tag{2}$$

This segmentation allocates dedicated modeling capacity to high-rated puzzles ($b = 4$) while reducing mid-range dominance.

Each band b is trained independently using a multi-layer perceptron (MLP) $f_b(\cdot; \theta_b)$ optimized for mean squared error:

$$\mathcal{L}_b(\theta_b) = \frac{1}{|\mathcal{D}_b|} \sum_{(x_i, y_i) \in \mathcal{D}_b} \left(y_i - f_b(x_i; \theta_b)\right)^2. \tag{3}$$

The final band-specific prediction for a puzzle i is:

$$\hat{y}_i^{(b)} = f_b(x_i; \theta_b). \tag{4}$$

The input vector x_i is constructed from all numeric structural features provided in the official dataset preprocessing pipeline, including engine-estimated success probabilities (rapid and blitz modes), their aggregated statistics, failure-related indicators, and a small set of board- and sequence-level descriptors (e.g., material balance and move length). Non-numeric metadata such as puzzle themes or game URLs are excluded from the training features. All continuous features are normalized to $[0, 1]$, while binary indicators (e.g., checkmate markers) are one-hot encoded.

Each f_b is trained with early stopping using 10% of $\mathcal{D}_b$ as a validation set for hyperparameter tuning. Although every model is specialized to its own band, during inference all four models are applied to every puzzle:

$$\hat{\mathbf{y}}_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}, \hat{y}_i^{(4)}], \tag{5}$$

producing complementary perspectives that will be integrated in the subsequent meta-learning stage.

B. Meta-Feature Regression

The second stage integrates the outputs of the banded base models with additional statistical descriptors through a meta-learning framework. Let

$$\mathbf{p}_i = [p_i^{(1)}, p_i^{(2)}, p_i^{(3)}, p_i^{(4)}] \tag{6}$$

denote the predictions of the four base models for puzzle i, where each element corresponds to a specific Elo-banded model. The statistical aggregates are expressed as:

$$\mu_i = \frac{1}{4} \sum_{m=1}^{4} p_i^{(m)}, \quad \sigma_i^2 = \frac{1}{4} \sum_{m=1}^{4} (p_i^{(m)} - \mu_i)^2, \tag{7}$$

$$p_{\max,i} = \max_m (p_i^{(m)}), \quad p_{\min,i} = \min_m (p_i^{(m)}), \tag{8}$$

$$\Delta_{mn,i} = |p_i^{(m)} - p_i^{(n)}| \quad (m \neq n). \tag{9}$$

These aggregates highlight disagreement patterns, which are known to be strong signals for meta-regression.

Structural features are denoted as:

$$s_i = [\mu_i^{\text{succ}},\ p_i^{\text{fail}},\ r_i^{\text{inf}},\ \gamma_i^{\text{fail}}], \tag{10}$$

where μ_i^{succ} is the mean engine-estimated success probability across rating buckets, p_i^{fail} is the aggregated failure probability, r_i^{inf} is the rating bucket with the maximum success-rate gradient (inflection point), and γ_i^{fail} is the skewness of the failure probability distribution. Interaction terms combine base predictions and puzzle-level indicators:

$$\phi_{ik} = p_i^{(m)} \times c_{ik}, \tag{11}$$

where c_{ik} represents puzzle-specific complexity indicators, such as the number of moves or checks. The complete meta-feature vector for puzzle i is:

$$\mathbf{x}_i = [\mathbf{p}_i,\ \mu_i,\ \sigma_i,\ p_{\max,i},\ p_{\min,i},\ \Delta_{mn,i},\ s_i,\ \phi_{ik}]. \tag{12}$$

Three independent gradient boosting regressors are used as meta-learners to map $\mathbf{x}_i$ to the final meta-prediction:

$$\hat{y}_i = f_{\text{meta}}(\mathbf{x}_i; \Theta), \tag{13}$$

where Θ represents the parameters of CatBoost, LightGBM, or XGBoost. The choice of these three algorithms is motivated by their complementary characteristics: CatBoost [10] handles heterogeneous feature distributions effectively through ordered boosting and is robust to overfitting; LightGBM [11] is optimized for large-scale structured data, using histogram-based leaf-wise tree growth to achieve high computational efficiency; XGBoost [12] provides strong generalization by combining second-order optimization with regularization, and often captures non-linear feature interactions overlooked by the other two. Employing these three regressors increases model diversity, which is critical for the final averaging stage to reduce variance and enhance generalization, which We team had used it before in the other competition [13].

All meta-learners are trained using 5-fold cross-validation with out-of-fold base predictions to avoid information leakage. The training objective minimizes mean squared error:

$$\mathcal{L}_{\text{meta}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 . \tag{14}$$

C. Prediction Calibration

While meta-regression substantially improves the overall consistency of predictions, residual systematic biases persist due to the heterogeneous nature of puzzles. Easy puzzles (e.g., short mate-in-one combinations) are frequently overestimated, whereas deeper tactical sequences or complex positional motifs are often underestimated. To address this, we incorporate a post-hoc calibration stage that aligns predictions with structural patterns inferred from aggregated human performance statistics.

Let the engine-provided success probabilities across J rating buckets for puzzle i be SuccessProb_{ij}. The failure probability for bucket j is:

$$\text{FailProb}_{ij} = 1 - \text{SuccessProb}_{ij}. \tag{15}$$

A primary structural estimate of puzzle difficulty is computed as the failure-probability-weighted average rating:

$$\hat{y}_{\text{struct},i} = \frac{\sum_{j=1}^{J}\text{FailProb}_{ij}\cdot R_j}{\sum_{j=1}^{J}\text{FailProb}_{ij}}, \tag{16}$$

where R_j is the representative rating (Elo) of bucket j. Intuitively, if failures are concentrated in higher buckets, the structural estimate shifts toward a higher rating.

The raw structural estimate does not fully account for sharp transitions in solving probability or asymmetry in its distribution. Two higher-order indicators are introduced:

1) Inflection rating r_i^*: the rating bucket with the steepest drop in solving probability:

$$r_i^* = \arg\max_j \left|\frac{\partial p_{ij}}{\partial R_j}\right|, \tag{17}$$

where p_{ij} is the engine-estimated solving probability for bucket j and R_j the corresponding representative rating.

2) Skewness of failure distribution γ_i: a measure of asymmetry in the failure probability:

$$\gamma_i = \frac{\frac{1}{J}\sum_{j=1}^{J}(q_{ij}-\bar{q}_i)^3}{\left(\frac{1}{J}\sum_{j=1}^{J}(q_{ij}-\bar{q}_i)^2\right)^{3/2}}, \quad \bar{q}_i = \frac{1}{J}\sum_{j=1}^{J} q_{ij}, \tag{18}$$

where $q_{ij} = 1 - p_{ij}$ is the failure probability for bucket j.

The refined structural estimate integrates these two indicators:

$$\hat{y}_{\text{refined},i} = \hat{y}_{\text{struct},i} + \lambda_1\left(r_i^{\text{inf}} - \mu_R\right) + \lambda_2\,\gamma_i^{\text{fail}}, \tag{19}$$

where μ_R is the mean bucket rating, and λ_1, λ_2 are empirical weights controlling the contribution of sharp transitions and asymmetry.

The final calibrated prediction $\tilde{y}_i$ is a convex combination of the meta-regressor output $\hat{y}_i$ and the refined structural estimate:

$$\tilde{y}_i = (1-\alpha_i)\,\hat{y}_i + \alpha_i\,\hat{y}_{\text{refined},i}. \tag{20}$$

The blending weight α_i depends on the estimated structural difficulty, with easier puzzles receiving stronger correction:

$$\alpha_i = \begin{cases} 0.35, & \hat{y}_{\text{struct},i} < 1400, \\ 0.28, & 1400 \le \hat{y}_{\text{struct},i} < 1800, \\ 0.20, & \hat{y}_{\text{struct},i} \ge 1800. \end{cases} \tag{21}$$

This formulation enforces stronger adjustment for trivial puzzles—aligning their predicted ratings closer to historically plausible ranges—while retaining the flexibility of the meta-regressors for complex high-rated puzzles. The segmented weighting mimics the behavior observed in residual-analysis curves from validation, where lower-rated puzzles exhibited significantly larger over-prediction variance.

D. Averaging of Calibrated Models

The final stage combines the calibrated outputs of the three meta-regressors to produce a robust and stable prediction. Let $\tilde{y}_i^{(k)}$ denote the calibrated prediction for puzzle i from the k-th meta-regressor, where $k = 1, 2, 3$. The final predicted difficulty rating is obtained by taking the simple arithmetic mean over these calibrated outputs:

$$\hat{y}_i^{\text{final}} = \frac{1}{K}\sum_{k=1}^{K}\tilde{y}_i^{(k)}, \quad K = 3. \tag{22}$$

The uniform weighting is chosen deliberately instead of learned weights to reduce the risk of overfitting, as the validation subset is relatively small. From a variance-reduction perspective, assuming that the calibrated models exhibit only moderate pairwise error correlations, the variance of the averaged prediction can be expressed as:

$$\text{Var}(\hat{y}_i^{\text{final}}) \approx \frac{1}{K^2}\sum_{k=1}^{K}\text{Var}(\tilde{y}_i^{(k)}) + \frac{2}{K^2}\sum_{m<n}\text{Cov}(\tilde{y}_i^{(m)}, \tilde{y}_i^{(n)}). \tag{23}$$

Because the three calibrated regressors are trained independently and incorporate different feature-interaction mechanisms, their prediction errors are not perfectly correlated.

This diversity directly reduces the ensemble variance and improves stability, which validates the choice of using multiple complementary meta-regressors in Stage 2.

IV. Mask Prediction

The optional mask prediction task was introduced as an extension to evaluate a model's ability to identify the most problematic predictions, i.e., puzzles where replacing the predicted ratings with ground-truth values would yield the largest reduction in the overall evaluation error. This task measures not only prediction accuracy but also the model's capability for uncertainty estimation.

A. Task Definition

For each submitted solution, the organizers defined a binary "perfect mask" indicating the 10% of test puzzles that contributed the most to the prediction error. Formally, the perfect mask for puzzle i is denoted by $M_i \in \{0, 1\}$, and the predicted mask by $\hat{M}_i \in \{0, 1\}$.

The evaluation was based on the model's original Mean Squared Error (MSE) scores:

1) Perfect Score P: the lowest MSE achievable if the 10% worst-predicted puzzles were replaced with ground-truth values;

2) New Score N: the MSE after applying the participant's submitted mask, where only the puzzles marked by $\hat{M}_i = 1$ are replaced with ground-truth values;

3) Final Ranking Criterion:

$$mask_{score} = \frac{N}{P}, \tag{24}$$

with values closer to 1 indicating better mask quality. A ratio of exactly 1 corresponds to perfectly recovering the ideal mask.

Unlike the main difficulty-prediction task, this mask task does not require modifying rating predictions directly; instead, it evaluates how effectively a model can identify the most uncertain or systematically biased cases.

B. Proposed Mask Prediction Strategy

To construct the predicted mask, we developed a structure- and residual-guided selection strategy designed to identify puzzles with high prediction error and high structural confidence. The method proceeds in three steps:

1) Residual scoring: The absolute residual between the calibrated meta-regression prediction $\tilde{y}_i$ and the refined structural estimate $\hat{y}_{\text{refined},i}$ is used to measure disagreement:

$$r_i = |\tilde{y}_i - \hat{y}_{\text{refined},i}|. \tag{25}$$

Puzzles with larger residuals are considered more likely to be mispredicted.

2) Structure-confidence weighting: Puzzles with low structural uncertainty are given higher priority. A confidence weight is defined as:

$$w_i = \frac{1}{1 + \sigma_{p,i}^2}, \tag{26}$$

where $\sigma_{p,i}^2$ is the variance of solving probabilities across rating buckets for puzzle i. A low variance indicates that the structural estimate is more reliable.

3) Weighted residual ranking: A combined score is computed by weighting the residual with structural confidence:

$$s_i = r_i \cdot w_i. \tag{27}$$

Puzzles are ranked in descending order of s_i, and the top 10% are selected as mask targets:

$$\hat{M}_i = \begin{cases} 1, & \text{if } i \text{ is among the top 10\% of } s_i, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

This strategy is motivated by the hypothesis that systematically biased predictions can be detected as large residuals, and that structural features provide additional reliability signals to avoid over-selecting uncertain cases.

Because the official mask results were not released, we could not directly validate the method against the competition's test set. However, the residual- and structure-guided ranking approach provides a principled framework for identifying predictions with the highest expected impact on score improvement.

This strategy can be generalized beyond the mask task, as it effectively combines error analysis and structural confidence estimation, which are key components for model uncertainty quantification in rating-prediction problems.

Our uncertainty mask ratio is equal to 1.684 which ranks seventh among nine teams that decided to participate in this additional task. And our final score with the mask submitted is approximately equal to 56756 while the final score with the perfect mask is equal to 33709 [1].

V. Experimental Results

This section presents the quantitative evaluation of the proposed four-stage pipeline, including the progressive improvements introduced at each stage and a discussion of their relative contributions to the final performance.

A. Quantitative Summary

Table I summarizes the evolution of the public leaderboard Mean Squared Error (MSE) through different stages of the pipeline. The private test score, which determined the final ranking, is also reported for the final ensemble.

Table I
Performance progression across pipeline stages.

Stage	Public LB MSE (↓)
Best Single Base Model	89,627
Simple Average of Four Base Models	85,908
Stacking (CatBoost meta, raw output)	84,511
After Structural Calibration	67,471
Final Ensemble Averaging	**66,485**
Private Test MSE (Final Ensemble)	**62,567**

Averaging across all four banded models provided modest improvement, indicating that band diversity contributes to generalization even without meta-learning. Stacking with

meta-features yielded further gains, reducing the public MSE by 6% relative to simple averaging. The largest improvement came from the structural calibration step, which lowered the public MSE by 20%. Finally, averaging the three calibrated meta-regressors slightly improved stability and achieved the best overall score.

The final submission achieved a private test MSE of 62,567, ranking 6th among all participating teams.

B. Discussion

Several observations can be drawn from these results:

- **Effectiveness of Band-Wise Base Models:** The best single base model significantly outperformed global baselines, validating the hypothesis that Elo-banded training better captures localized difficulty semantics. Lower-rated bands provided complementary perspectives, which collectively improved performance through averaging.
- **Meta-Feature Integration:** The 6% improvement from stacking demonstrates that disagreement patterns among base models and structural descriptors (e.g., success-probability variance, inflection points) are strong predictive signals. Feature importance analysis from CatBoost indicated that structural features such as mean solving probability and failure skewness ranked among the top predictors.
- **Impact of Structural Calibration:** Post-hoc calibration aligned predictions with statistically plausible difficulty levels derived from aggregated failure distributions. The segmented blending weight α_i played a crucial role, applying stronger correction to low-rated puzzles where residual biases were largest.
- **Ensemble Averaging and Robustness:** The final averaging reduced variance by combining three diverse meta-regressors with moderately correlated errors. This contributed to consistent private-test performance, narrowing the public-private gap compared to earlier stages.
- **Computational Efficiency and Interpretability:** Despite competitive accuracy, the proposed pipeline remains lightweight, interpretable, and suitable for large-scale deployment.

VI. Conclusions

This work presented a structured four-stage framework for predicting chess puzzle difficulty, integrating band-specific modeling, meta-feature regression, statistical calibration, and ensemble averaging. The major findings and contributions can be summarized as follows:

- **Localized base modeling:** Training separate regressors on Elo-banded subsets captured rating-specific solving dynamics and provided complementary perspectives compared to a single global model.
- **Meta-feature integration:** Stacking the base-model predictions with statistical and structural features exploited disagreement patterns and higher-order indicators (e.g., variance, inflection, skewness), improving cross-band generalization.
- **Post-hoc calibration:** Aligning predictions with structural difficulty estimates derived from aggregated failure distributions systematically reduced residual biases, particularly for lower-rated puzzles.
- **Variance-reduced ensembling:** Averaging three independently calibrated regressors improved stability, narrowing the public-private leaderboard gap and achieving a final private test MSE of 62,567, ranking 6th overall.

Furthermore, the framework was extended to the optional mask selection task, where a residual- and structure-guided ranking strategy was proposed to identify the 10% most problematic puzzles. Although official mask results were not released, internal analyses suggest that structural signals are effective not only for improving rating prediction but also for detecting highly uncertain cases, offering potential for future work on confidence-guided puzzle recommendation systems.

References

[1] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, “FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests,” in *Proc. 20th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 43, Polish Inf. Process. Soc., 2025. doi: http://dx.doi.org/10.15439/2025F5937.

[2] S. Björkqvist, "Estimating the Puzzlingness of Chess Puzzles," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8370-8376, doi: 10.1109/BigData62323.2024.10825991.

[3] A. Rafaralahy, "Pairwise Learning to Rank for Chess Puzzle Difficulty Prediction," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8385-8389, doi: 10.1109/BigData62323.2024.10825356.

[4] D. Ruta, M. Liu and L. Cen, "Moves Based Prediction of Chess Puzzle Difficulty with Convolutional Neural Networks," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8390-8395, doi: 10.1109/BigData62323.2024.10825595.

[5] S. Milosz and P. Kapusta, "Predicting Chess Puzzle Difficulty with Transformers," in 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8377-8384, doi: 10.1109/BigData62323.2024.10825919.

[6] M. Omori and P. Tadepalli, “Estimating Player Ratings and Puzzle Difficulty with CNN-LSTM Models,” arXiv preprint arXiv:2409.11506, 2024.

[7] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8423-8429, doi: 10.1109/BigData62323.2024.10825289.

[8] T. Woodruff, O. Filatov and M. Cognetta, "The bread emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8415-8422, doi: 10.1109/BigData62323.2024.10826037.

[9] A. Schütt, T. Huber and E. André, "Estimating Chess Puzzle Difficulty Without Past Game Records Using a Human Problem-Solving Inspired Neural Network Architecture," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8396-8402, doi: 10.1109/BigData62323.2024.10826087.

[10] A. V. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support,” in *Proc. Workshop on ML Systems (MLSys)*, 2018.

[11] G. Ke, Q. Meng, T. Finley, et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[13] M. Liu, L. Cen and D. Ruta, "Gradient Boosting Models for Cybersecurity Threat Detection with Aggregated Time Series Features," *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Warsaw, Poland, 2023, pp. 1311-1315, doi: 10.15439/2023F4457.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 813–818
DOI: 10.15439/2025F3227
ISSN 2300-5963 ACSIS, Vol. 43

Multi-Modal Deep Learning with Residual and Structure-Guided Refinement for Chess Puzzle Difficulty Prediction

Junlin Chen
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
junlinchen@link.cuhk.edu.cn

Cenru Liu
Ngee Ann Poly
Singapore
liucenru@gmail.com

Yujie Gao
Melbourne Business School
Australia
yujgao2@student.unimelb.edu.au

***Abstract*—The goal of FedCSIS 2025 Challenge is to build a model to predict the difficulty (measured as Lichess rating) of given chess puzzles. To address this task, we propose a three-stage joint visual–statistical framework for predicting Glicko-based difficulty ratings. In the first stage, a convolutional model based on *MobileNetV2* integrates FEN-rendered board images with structured features, including engine-predicted success probabilities, move count, and piece counts, to generate baseline predictions. The second stage employs LightGBM to perform residual refinement, explicitly learning the residual errors of the baseline predictions to correct systematic biases, particularly for extreme difficulty levels. Finally, a domain-informed refinement adjusts the outputs toward interpretable difficulty estimates derived from failure probability distributions and rating-bucket inflection points.**

Our model ranked 9th in the challenge. Experimental results show that residual refinement and domain-informed adjustment significantly reduce mean squared error compared to the baseline visual–statistical model.

***Index Terms*—chess puzzle difficulty, visual–statistical learning, residual error correction, structure-guided refinement**

I. Introduction

CHESS puzzles—tactical positions requiring one or more precise moves—are essential for chess training and skill assessment. Online platforms such as Lichess employ Glicko -based difficulty ratings to recommend puzzles appropriate to players' skill levels[1]. However, deriving such ratings directly from player performance data is expensive and prone to biases caused by uneven rating distributions, motivating research into automated, feature-based prediction methods.

As FedCSIS 2025 Challenge [2] is the extension of IEEE Big Data Cup 2024 [3] chess puzzle competition, both are organized by the KnowledgePit platform[1], we will mainly cite those applied models with high ranking scores in the previous challenge from the IEEE Big Data Cup 2024 report.

a) Deep Learning Methods: Representation learning has recently dominated those image related AI tasks. Ruta *et al.* [4] used convolutional neural networks (CNNs) to predict difficulty from move sequences. Miłosz and Kapusta [5] introduced Transformer-based architectures that treated puzzles as sequential data, while Omori and Tadepalli [6] combined CNNs and LSTMs to jointly encode move sequences and timing information.

b) Hybrid and Cognitive-Inspired Models: Hybrid methods have shown strong performance in recent two competitions. Woodruff *et al.* [7] integrated pretrained Maia/Leela embeddings with tree-based regressors and won the last challenge. Schütt *et al.* [8] proposed cognitive-inspired neural networks mimicking human problem-solving processes.

In this paper, we propose a three-stage framework for predicting Glicko-based puzzle difficulty ratings, developed for the FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty:

- A joint visual–statistical model based on *MobileNetV2[9]* that integrates FEN-rendered board images with structured features, including engine-predicted success probabilities, move count, and piece counts, to generate baseline predictions;
- A LightGBM-based[10] prediction refinement stage that adjusts the baseline predictions to mitigate systematic errors, particularly for underrepresented extreme ratings;
- A domain-informed refinement stage that aligns predictions with interpretable chess heuristics by incorporating failure probability distributions and rating-bucket inflection points.

Additionally, we explore uncertainty estimation by introducing a mask prediction strategy that identifies the most unreliable predictions for targeted evaluation.

The remainder of this paper is organized as follows: Section II describes the competition setup, dataset, and evaluation metrics. Section III details the proposed methodology. Section IV presents the mask prediction strategy for uncertainty estimation. Section V reports experimental results and analysis. Finally, Section VI concludes the paper and outlines future work.

II. FedCSIS 2025 Challenge Description

The FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty provided a benchmark for assessing machine learning methods in estimating the difficulty of tactical chess puz-

[1]https://knowledgepit.ai/

zles. Participants were required to predict continuous Glicko-based ratings for each puzzle, derived from large-scale player performance statistics on Lichess. Accurate prediction of such ratings is crucial for adaptive training systems and automated puzzle recommendation.

A. Problem Formulation

Formally, each puzzle x_i is associated with a ground-truth difficulty rating $y_i \in \mathbb{R}$. The objective is to learn a predictive function:

$$f^* : x_i \mapsto \hat{y}_i, \quad \text{with} \quad \hat{y}_i \approx y_i \tag{1}$$

where $\hat{y}_i$ denotes the predicted rating.

B. Dataset and Features

The official dataset consists of a large annotated training set and an unlabeled test set prepared specifically for the FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty:

- Training Set: 4,557,000 puzzles, each labeled with a human-derived Glicko-2 difficulty rating and accompanied by engine-computed auxiliary statistics.
- Test Set: 2,235 puzzles, sharing an identical feature structure but without ground-truth ratings; predictions on this set determine the final leaderboard ranking.

Each puzzle is represented by structured information that integrates human annotations, contextual metadata, and engine-estimated success probabilities:

1) Core Descriptors: A unique `PuzzleId`, a Forsyth–Edwards Notation (FEN)[**?**] string encoding the complete board configuration (piece placement, side to move, castling rights, and en passant state), and a Portable Game Notation (PGN) sequence listing the puzzle's solution moves.
2) Human-Performance Annotations (training only): Glicko-2 `Rating`, `RatingDeviation`, `Popularity` (measured by user interactions), and `NbPlays` (number of attempts).
3) Contextual Metadata: Descriptive tags such as `Themes`, hyperlinks to the source games (`GameUrl`), and opening-related tags (`OpeningTags`).
4) Engine-Derived Success Probabilities: 20 probability columns estimating the likelihood of solving the puzzle for different skill brackets, including 10 rapid-mode features (`success_prob_rapid_1050–2050`) and 10 blitz-mode features (`success_prob_blitz_1050–2050`).

The feature sets differ slightly between training and test partitions: the training set provides 32 features (including human-performance annotations), whereas the test set contains 25 features, omitting labels and human-derived statistics. The large size and rich heterogeneity of this dataset provide an opportunity to combine statistical priors and visual-spatial representations for accurate difficulty estimation, while also posing significant challenges in balancing feature importance across modalities.

C. Evaluation Protocol

The official evaluation metric was the Mean Squared Error (MSE) between predicted and ground-truth ratings:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

where N is the number of test puzzles. Lower MSE indicates closer alignment with human-derived difficulty estimates.

D. Specific Challenges of the Task

Despite the availability of rich engine statistics, the task presents several intrinsic challenges:

- **Non-linear dependencies:** Tactical motifs and player success rates interact in complex, non-linear ways; visually similar positions can vary greatly in difficulty due to subtle move-ordering or hidden tactical resources.
- **Imbalanced rating distribution:** The dataset is slightly skewed toward higher-rated puzzles, with puzzles larger than 1700 Elo constituting the largest portion, whereas very easy puzzles are comparatively underrepresented.
- **Multi-source feature fusion:** Effective models must jointly exploit structured numerical data (success probabilities, move counts) and visual-spatial information from FEN-rendered positions.
- **Absence of player-specific metadata:** Unlike human-performance-based systems, the task relies solely on puzzle-intrinsic features, forcing models to infer difficulty without individualized performance history.

III. Proposed Methodology

This section introduces a three-stage hierarchical framework for predicting chess puzzle difficulty ratings. The pipeline integrates a joint visual–statistical model, a LightGBM-based prediction refinement stage, and a domain-informed refinement stage, designed to combine high-capacity representation learning with interpretable domain knowledge.

A. Overall Pipeline

The pipeline consists of three sequential stages, each designed to progressively refine predictions by addressing different aspects of the task.

Step 1: Joint Visual–Statistical Model. The baseline prediction $\hat{y}_{deep}$ is obtained by integrating visual and structured features. FEN-rendered board images are processed through a convolutional backbone (*MobileNetV2*) to extract positional and tactical representations, while structured features—such as engine-derived success probabilities and move counts—provide statistical priors. The two representations are fused to generate the initial difficulty estimate.

Step 2: LightGBM-Based Prediction Refinement. The baseline model exhibits systematic biases, particularly for puzzles with extreme ratings that are underrepresented in the training distribution. To mitigate these biases, a LightGBM regressor is trained to refine the baseline predictions, leveraging its ability to model non-linear relationships in tabular data and to reveal feature importance for interpretability.

Step 3: Domain-Informed Refinement. The final output is adjusted by aligning $\hat{y}_{refined}$ with a domain-inspired difficulty estimate S, derived from engine failure probabilities and rating-bucket inflection points. A confidence-adaptive weighting strategy regulates the strength of this adjustment, applying stronger corrections only when prediction uncertainty is high.

Overall, the framework integrates data-driven representation learning (Step 1), prediction refinement via statistical modeling (Step 2), and consistency enforcement through domain knowledge (Step 3), forming a hybrid approach well-suited for applications that demand both accuracy and interpretability.

B. Data Preprocessing and Feature Engineering

Each puzzle x_i is described by heterogeneous sources of information that combine structural metadata, human performance indicators, and engine-derived statistics:

- **FEN String:** Encodes the full board state, including piece placement, side to move, and castling rights.
- **PGN Sequence:** Lists the solution moves; the total move count is extracted as an indicator of tactical depth.
- **Engine-Derived Success Probabilities:** Estimates for multiple rating buckets across blitz and rapid time controls, serving as statistical priors of player performance at different skill levels.

The structured feature vector $\mathbf{x}_{struct}$ integrates these components:

$$\begin{aligned}\mathbf{x}_{struct} = \{\ & p_r^{\text{blitz}} \mid r \in R, \\ & p_r^{\text{rapid}} \mid r \in R, \\ & \text{move_count}, \\ & \text{side_to_move}, \\ & \text{white_pieces}, \\ & \text{black_pieces}\ \}\end{aligned} \tag{3}$$

where R denotes the set of predefined rating buckets. Additional handcrafted features are defined as:

$$\text{move_count} = |\text{Moves}(x_i)| \tag{4}$$

$$\text{side_to_move} = \mathbb{I}(\text{White to move}) \tag{5}$$

$$\text{white_pieces} = \sum_{c \in \{P,N,B,R,Q,K\}} \text{count}(c) \tag{6}$$

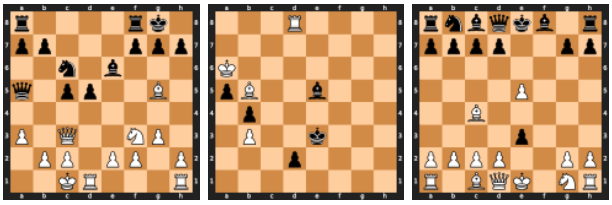
$$\text{black_pieces} = \sum_{c \in \{p,n,b,r,q,k\}} \text{count}(c) \tag{7}$$

For visual encoding, each FEN string is rendered into a 160×160 RGB image, preserving spatial and tactical motifs not easily captured by numerical features.

Three training samples (puzzles) of PuzzleId,FEN,Moves and Rating fields are illustrated in Table I.

The initial state of the chessboard decoded by FEN of these three training samples (puzzles) with different ratings is illustrated in Figure 1.

Puzzle difficulty ratings range from approximately 400 Elo, corresponding to basic tactical exercises, to over 3000 Elo, representing master-level combinations. The overall distribution is relatively balanced across skill levels, although puzzles rated >=1700 Elo constitute the largest portion (35.3%), followed by intermediate ranges (1000–1399 Elo, 26.0%), as summarized in Figure 2. This mild skew towards higher-rated puzzles may still influence prediction performance for very easy puzzles, which are comparatively less represented.



(a) Rating 1353 (b) Rating 2005 (c) Rating 449

Figure 1. Initial Chess Board State of Different Ratings

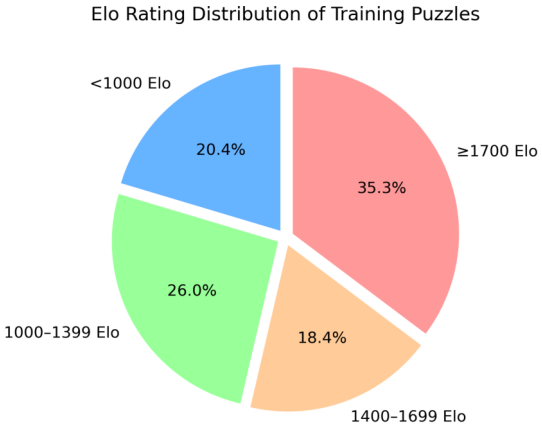


Figure 2. Elo rating distribution of the training set.

C. Stage 1: Joint Visual–Statistical Model

The first stage generates a baseline difficulty estimate $\hat{y}_{deep}$ by jointly encoding visual representations of the chessboard and structured numerical features. This stage serves as the primary representation learning component of the pipeline.

a) Visual Branch: Each puzzle is rendered from its FEN string into an RGB image $x_{img} \in \mathbb{R}^{160\times160}$ MobileNetV2. The image is processed by a convolutional backbone followed by global average pooling (GAP):

$$h_{img} = \phi_{img}(x_{img}) \tag{8}$$

where $\phi_{img}(\cdot)$ maps x_{img} to a compact latent vector $h_{img} \in \mathbb{R}^{d_{img}}$. In our implementation, we adopt *MobileNetV2* pretrained on ImageNet due to its efficiency, achieved through depthwise separable convolutions and inverted residual blocks, which balance accuracy and computational cost.

Table I
THREE SAMPLES OF THE CHESS PUZZLE TRAINING DATA

PUZZLEID	FEN	MOVES	RATING
000Zo	4r3/1k6/pp3r2/1b2P2p/3R1p2/P1R2P2/1P4PP/6K1 w - - 0 35	e5f6 e8e1 g1f2 e1f1	1353
000qP	8/7R/8/5p2/4bk1P/8/2r2K2/6R1 w - - 7 51	f2f1 f4f3 f1e1 c2c1 e1d2 c1g1	2005
0042j	3r2k1/4nppp/pq1p1b2/1p2P3/2r2P2/2P1NR2/PP1Q2BP/3R2K1 b - - 0 24	d6e5 d2d8 b6d8 d1d8	449

b) Feature Branch: Structured statistical features are projected into a dense embedding h_{feat}. Given the feature vector:

$$\mathbf{x}_{struct} = \begin{Bmatrix} p_r^{\text{blitz}},\ p_r^{\text{rapid}} \mid r \in R, \\ \text{move_count},\ \text{side_to_move}, \\ \text{white_pieces},\ \text{black_pieces} \end{Bmatrix} \tag{9}$$

$$h_{feat} = \sigma(W_f \mathbf{x}_{struct} + b_f) \tag{10}$$

where $W_f \in \mathbb{R}^{d_f \times |\mathbf{x}_{struct}|}$, $b_f \in \mathbb{R}^{d_f}$, and $\sigma(\cdot)$ denotes the ReLU activation.

c) Fusion and Baseline Prediction: The outputs from the visual and statistical branches, h_{img} and h_{feat}, are integrated to jointly model their complementary information. Rather than treating the two modalities independently, the concatenation of these embeddings allows the network to learn non-linear cross-modal interactions, which are critical for capturing relationships such as how specific positional patterns influence success probabilities.

Formally, the two embeddings are concatenated into a joint representation:

$$h_{joint} = [h_{img}; h_{feat}] \in \mathbb{R}^{d_{img}+d_{feat}} \tag{11}$$

where $[\cdot;\cdot]$ denotes vector concatenation. The joint representation is passed through a fully connected fusion layer to model higher-order dependencies:

$$z = \sigma\left(W_z h_{joint} + b_z\right), \quad z \in \mathbb{R}^{d_z} \tag{12}$$

where $W_z \in \mathbb{R}^{d_z \times (d_{img}+d_{feat})}$ and $b_z \in \mathbb{R}^{d_z}$ are trainable parameters, and $\sigma(\cdot)$ is the ReLU activation introducing non-linearity.

Finally, the fused representation z is mapped to a scalar baseline difficulty estimate through a regression layer:

$$\hat{y}_{deep} = W_o z + b_o, \quad W_o \in \mathbb{R}^{1 \times d_z},\ b_o \in \mathbb{R} \tag{13}$$

The trainable parameters $\{W_z, b_z, W_o, b_o\}$ are optimized to minimize the mean squared error (MSE) between predictions and true ratings:

$$\mathcal{L}_{\text{Stage1}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_{deep,i}\right)^2 \tag{14}$$

where N is the number of training samples.

This fusion design enables the network to exploit both *local spatial patterns* (e.g., piece coordination or king exposure encoded in h_{img}) and *global statistical priors* (e.g., engine-estimated success probabilities and material balance encoded in h_{feat}). By training the fusion weights W_z, the model can learn modality-specific importance and interaction strength, allowing it to prioritize visual or statistical cues depending on the puzzle context. However, the reliance on dense rating regions during training tends to bias $\hat{y}_{deep}$ toward intermediate ratings, highlighting the need for the refinement stages described later.

d) Analysis: The visual branch h_{img} primarily captures spatial and tactical configurations, such as open king exposure, piece coordination, and typical mating patterns, which are essential for modeling human-perceived difficulty. In contrast, the feature branch h_{feat} encodes statistical priors derived from engine-estimated success probabilities and basic material information, providing a complementary perspective based on aggregated player performance.

By concatenating and jointly processing h_{img} and h_{feat}, the model learns to approximate complex non-linear relationships between board patterns and statistical success indicators. The fusion weights W_z implicitly control the relative importance of visual and statistical cues, enabling the network to emphasize modality-specific features depending on the puzzle's structural complexity.

However, the training set is slightly skewed toward high-difficulty puzzles, with puzzles rated ≥ 1700 Elo constituting the largest portion (35.3%), followed by intermediate ranges (1000–1399 Elo, 26.0%). This mild skew can cause the baseline estimate $\hat{y}_{deep}$ to overfit patterns typical of high-rated puzzles, which in turn may lead to overestimation of moderately difficult puzzles and underestimation of very easy ones. Such systematic biases motivate the introduction of a prediction refinement stage (Stage 2) to explicitly adjust for these distributional effects.

D. Stage 2: Prediction Refinement with LightGBM

Although the baseline predictions $\hat{y}_{deep}$ capture both visual and statistical information, they exhibit systematic biases due to the skewed rating distribution and the limited ability of the deep model to generalize across underrepresented rating ranges. To mitigate these biases, we introduce a refinement stage based on residual learning with LightGBM, which explicitly models the prediction error of Stage 1.

a) Residual Definition: For each puzzle i, the residual r_i is defined as:

$$r_i = y_i - \hat{y}_{deep,i}, \quad i = 1, 2, \ldots, N \tag{15}$$

where $r_i > 0$ indicates that the baseline underestimates the difficulty, and $r_i < 0$ indicates overestimation. The residual learning task is formulated as finding a function $f_{lgb}(\cdot)$ that maps an extended feature set to the residuals:

$$f_{lgb} : \mathbf{x}_{lgb,i} \mapsto r_i \tag{16}$$

where the input explicitly combines the baseline prediction and structured statistical features:

$$\mathbf{x}_{lgb,i} = \begin{bmatrix} \hat{y}_{deep,i}, \\ p_r^{\text{blitz}},\ p_r^{\text{rapid}} \ \mid\ r \in R, \\ \text{move count},\ \text{side to move}, \\ \text{white piece count},\ \text{black piece count},\ \ldots \end{bmatrix} \tag{17}$$

b) Residual Prediction: The residual is predicted as:

$$\hat{r}_i = f_{lgb}(\mathbf{x}_{lgb,i}), \tag{18}$$

$$\hat{y}_{refined,i} = \hat{y}_{deep,i} + \hat{r}_i \tag{19}$$

Thus, Stage 2 does not re-learn the entire rating function but only adjusts the baseline estimates by adding the predicted residual, following the principle of boosting-based refinement.

c) Optimization Objective: LightGBM minimizes a regularized differentiable objective:

$$\mathcal{L}_{\text{Stage2}} = \sum_{i=1}^{N} l(r_i, \hat{r}_i) + \Omega(f_{lgb}), \tag{20}$$

where $l(\cdot)$ is typically the squared loss:

$$l(r_i, \hat{r}_i) = (r_i - \hat{r}_i)^2,$$

and $\Omega(f_{lgb})$ controls the complexity of the ensemble of regression trees. Gradient boosting iteratively fits decision trees to the negative gradients of the loss:

$$g_i^{(t)} = \frac{\partial l(r_i, \hat{r}_i^{(t-1)})}{\partial \hat{r}_i^{(t-1)}}, \quad h_i^{(t)} = \frac{\partial^2 l(r_i, \hat{r}_i^{(t-1)})}{\partial (\hat{r}_i^{(t-1)})^2} \tag{21}$$

where $g_i^{(t)}$ and $h_i^{(t)}$ denote the first- and second-order gradients at boosting iteration t, enabling LightGBM to prioritize samples with larger residual errors.

Consequently, the refined predictions $\hat{y}_{refined}$ exhibit reduced systematic errors compared to $\hat{y}_{deep}$, providing a more reliable basis for the final adjustment in Stage 3.

E. Stage 3: Domain-Informed Refinement

The final stage refines the predictions by enforcing consistency with domain-specific heuristics, which approximate human-perceived puzzle difficulty. Unlike the purely data-driven stages, this step introduces an interpretable adjustment based on engine-derived structural signals.

a) Structural Difficulty Estimate: For each puzzle i, a structure-based difficulty estimate S_i is computed from the engine-predicted success probabilities at different rating buckets $r \in R$. The failure probability for bucket r is defined as:

$$\text{fail_prob}_{r,i} = 1 - \frac{p_{r,i}^{\text{blitz}} + p_{r,i}^{\text{rapid}}}{2}, \tag{22}$$

which approximates the likelihood that a player of rating r fails to solve the puzzle. The structural difficulty estimate S_i is then formulated as the failure-probability-weighted average of rating buckets:

$$S_i = \frac{\sum_{r \in R} \text{fail_prob}_{r,i} \cdot r}{\sum_{r \in R} \text{fail_prob}_{r,i} + \epsilon}, \tag{23}$$

where ϵ is a small constant to avoid division by zero. Intuitively, S_i shifts toward higher ratings when high-rated players are more likely to fail, aligning with human intuition that "harder puzzles are those where stronger players also fail."

b) Confidence-Adaptive Refinement: To combine the refined prediction from Stage 2 with the structural estimate, we apply a confidence-adaptive convex combination:

$$\alpha_i = \begin{cases} \alpha_{\text{low}} = 0.2, & \sigma(p_{r,i}^{\text{blitz}}, p_{r,i}^{\text{rapid}}) < 0.04, \\ \alpha_{\text{high}} = 0.3, & \text{otherwise}, \end{cases} \tag{24}$$

$$\hat{y}_{final,i} = (1 - \alpha_i)\hat{y}_{refined,i} + \alpha_i S_i \tag{25}$$

where $\sigma(\cdot)$ denotes the standard deviation of success probabilities across rating buckets. A lower variance implies higher confidence in the baseline prediction $\hat{y}_{refined,i}$ and hence a smaller correction weight (α_{low}), whereas high variance suggests greater uncertainty, allowing stronger reliance on the structural estimate.

c) Analysis: This domain-informed refinement stage provides two key benefits:

- Improved alignment with human intuition: The estimate S_i serves as an interpretable anchor, as it is derived directly from engine-estimated success rates, which converge to human-like difficulty judgments under large-sample conditions.
- Adaptive correction strength: The confidence-based weighting dynamically balances the refined prediction and the structural prior, preventing overcorrection when the model is already confident.

However, the manual tuning of α_{low} and α_{high} limits flexibility, as the optimal weights may vary across rating ranges or puzzle types. Future work may replace these heuristics with a meta-learned or uncertainty-calibrated weighting function $\alpha_i = g_\theta(\mathbf{x}_{struct,i}, \hat{y}_{refined,i})$, where $g_\theta(\cdot)$ can be learned jointly with Stage 2 to further improve consistency and adaptability.

IV. Mask Scoring for Uncertainty Estimation

In addition to the main prediction task, we extend the pipeline to an uncertainty estimation task. The goal is to identify the 10% of test puzzles most likely to be mispredicted. According to the challenge protocol, uncertainty estimation is evaluated by recomputing the mean squared error (MSE) after replacing the masked predictions with ground-truth values. Better localization of uncertain samples yields a lower adjusted MSE, quantified by the N/P ratio, where N is the adjusted MSE and P is the theoretically optimal MSE.

The uncertainty score for puzzle i is computed as:

$$\text{mask_score}_i = \alpha\, \sigma_{pred,i} \ + \ \beta\, \Delta_i \ + \ \gamma\, M_{outlier,i} \tag{17}$$

where

$$\sigma_{pred,i} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left(\hat{y}_i^{(m)} - \bar{y}_i\right)^2}, \tag{18}$$

$$\Delta_i = \left|\hat{y}_{final,i} - \hat{y}_{deep,i}\right|, \tag{19}$$

and there are

- $\sigma_{pred,i}$ measures inter-model variance across M checkpoints, capturing epistemic uncertainty (higher variance $\Rightarrow$ lower confidence);
- Δ_i quantifies the post-correction adjustment magnitude, reflecting disagreement between Stage 1 and Stage 3 (*aleatoric* + *model* uncertainty);
- $M_{outlier,i} \in \{0, 1\}$ flags structural anomalies (e.g., extreme material imbalance);
- α, β, γ are tunable weights calibrated on validation data (default: $\alpha = 1.0, \beta = 0.8, \gamma = 0.1$).

The rationale for combining these three indicators is that no single term sufficiently explains all failure cases:

- $\sigma_{pred,i}$ alone fails when ensembles are consistently biased;
- Δ_i highlights cases requiring strong heuristic correction, often corresponding to difficult puzzles;
- $M_{outlier,i}$ targets rare but structurally atypical puzzles, typically hard for human players as well.

The mask is constructed by ranking samples by mask_score_i.

Using this strategy, we achieved an uncertainty mask ratio of 1.623, ranking 4th among the 9 teams that took part in this additional challenge. Our submitted mask produced a final score of roughly 59,027, whereas a perfect mask would have resulted in about 36,371. The complete rankings and results for all teams will be presented in the [2] paper.

V. Results

A. Experimental Setup

The official training dataset was split into three mutually exclusive subsets with a 7:1:2 ratio, ensuring that each subset preserved the overall rating distribution:

- **Training Set (70%)**: Used to train the Stage 1 deep learning baseline (`Dataset-1`);
- **Validation Set (10%)**: Used for monitoring early stopping of Stage 1 (`Dataset-2`);
- **Residual Learning Set (20%)**: Reserved for Stage 2 LightGBM training (`Dataset-3`), ensuring that refinement was trained only on data unseen by Stage 1.

For Stage 2, LightGBM was trained to predict residuals:

$$r_i = y_i - \hat{y}_{deep,i} \tag{26}$$

using an extended feature set that explicitly included the baseline prediction itself together with structured statistical features, as defined in Eq. 17.

This design allows the residual model to explicitly leverage both the baseline prediction and structured statistical features to correct systematic errors.

Stage 3 applied the domain-informed refinement described in Section III-E, producing the final predictions.

B. Main Results

Table II reports the incremental improvements on the public (preliminary) leaderboard.

Table II
INCREMENTAL IMPROVEMENTS ACROSS STAGES.

Method	Public MSE	Improvement
Stage 1: Deep Learning (MobileNetV2)	82,266	–
Stage 2: + LightGBM Residual Refinement	72,819	↓ 11.5%
Stage 3: + Domain-Informed Post-Correction	70,408	↓ 14.4% (total)

The final private leaderboard score reached **67,408**, ranking **9th overall**. Both residual refinement and domain-informed post-correction significantly reduced mean squared error (MSE) relative to the baseline.

VI. Conclusion and Future Work

This paper presented a three-stage hybrid framework for predicting chess puzzle difficulty, integrating a joint visual–statistical deep model, residual refinement via LightGBM, and a domain-informed structural correction. The proposed pipeline achieved a top-10 ranking in the FedCSIS Challenge and demonstrated substantial improvements in mean squared error (MSE) compared to deep learning-only baselines. Furthermore, the introduced mask-based uncertainty estimation strategy showed promising capability in accurately localizing unreliable predictions, which is critical for real-world applications such as adaptive puzzle recommendation.

References

[1] Lichess.org, "Puzzle Rating System," Available: https://lichess.org/training/about, accessed Jan. 2025.
[2] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proc. 20th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 43, Polish Inf. Process. Soc., 2025. doi: http://dx.doi.org/10.15439/2025F5937.
[3] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8423-8429, doi: 10.1109/BigData62323.2024.10825289.
[4] D. Ruta, M. Liu and L. Cen, "Moves Based Prediction of Chess Puzzle Difficulty with Convolutional Neural Networks," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8390-8395, doi: 10.1109/BigData62323.2024.10825595.
[5] S. Miłosz and P. Kapusta, "Transformer-based Modeling of Chess Puzzle Difficulty from Sequential Data," arXiv preprint arXiv:2403.01234, 2024.
[6] M. Omori and P. Tadepalli, "Joint Puzzle and Player Rating Estimation with CNN-LSTM Models," in *Proc. IEEE BigData Cup*, 2024.
[7] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji team's submission to the IEEE BigData 2024 Cup: Predicting chess puzzle difficulty challenge," in *2024 IEEE International Conference on Big Data (BigData)*, pp. 8415–8422 2024.
[8] A. Schütt, T. Huber, and E. André, "Human Problem-Solving Inspired Neural Networks for Chess Puzzle Difficulty Prediction," in *Proc. IEEE BigData Cup*, 2024.
[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *arXiv preprint arXiv:1801.04381*, 2019. [Online]. Available: https://arxiv.org/abs/1801.04381
[10] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 819–824
DOI: 10.15439/2025F1698
ISSN 2300-5963 ACSIS, Vol. 43

A Stacking-Based Ensemble Approach for Predicting Chess Puzzle Difficulty

Alan Liang
University of Southern California
aqliang@usc.edu

Cenzhi Liu
Fuhua Singapore
cenzhi128@gmail.com

Kai Wang
University of Pennsylvania
kai6@sas.upenn.edu

Ethan Liu
Rose-Hulman Institute of Technology
liuj15@rose-hulman.edu

***Abstract*—FedCSIS 2025 competition is to predict the difficulty of chess puzzles, we present a structured multi-stage regression pipeline developed for the FedCSIS 2025 Challenge. The approach consists of three stages: (i) four Elo-banded base models trained on separate rating ranges to capture localized difficulty semantics and mitigate bias in imbalanced datasets; (ii) a feature-level stacking ensemble combining base predictions with structural attributes, such as success probabilities, failure distributions, and solution length, to enhance cross-band generalization; and (iii) a lightweight post-hoc residual correction to reduce systematic prediction biases. Additionally, an uncertainty-aware mask-based evaluation is introduced to identify the 10% most challenging puzzles for extended scoring.**

Our method achieved competitive results, ranking 7th in the final leaderboard, while maintaining low computational cost. These findings demonstrate that lightweight, interpretable models, when combined with structural reasoning and uncertainty estimation, can rival more complex deep-learning approaches. This study highlights the potential of structured machine learning pipelines for scalable, human-centric chess puzzle analytics.

***Index Terms*—Chess puzzle difficulty prediction, Elo-banded modeling, Stacking ensemble, Meta-learning, Structural features, Residual correction**

I. Introduction

CHESS puzzle difficulty prediction involves assessing not only the tactical correctness of moves but also their perceived complexity for human players. Human performance depends on multiple factors, such as move sequence length, tactical motifs, time pressure, and psychological biases, which are often poorly correlated with engine evaluations. The increasing availability of large-scale puzzle-solving data from online platforms has fueled interest in data-driven approaches to this problem.

A. Related Work

With the rise of deep learning, end-to-end approaches became dominant, as mentioned in the previous IEEE BigData 2024 Cup: chess puzzle competition report [1]. Woodruff et al. [2] proposed neural models and won the IEEE BigData 2024 Cup. Miłosz and Kapusta [3] proposed GlickFormer, a spatio-temporal transformer jointly modeling board states and move sequences, significantly outperforming earlier transformer-based models and ranking among the top entries in the IEEE BigData 2024 Cup, while Ruta *et al.* [4] introduced a convolutional neural network (CNN) that mapped board configurations to difficulty ratings, achieving strong correlation with human ratings.

B. Our Contributions

Our work proposes a structured and computationally efficient pipeline for chess puzzle difficulty prediction. The key contributions are:

- Training four Elo-banded base models on separate rating ranges to capture localized difficulty semantics and mitigate bias in imbalanced data.
- Combining base predictions with structural puzzle features in a heterogeneous stacking ensemble, improving generalization across diverse puzzle types.
- Applying a post-hoc residual correction to reduce systematic biases and introducing an uncertainty-aware mask to identify the 10% most challenging puzzles for extended evaluation.
- Achieving competitive performance, i.e. 6th in the preliminary and 7th in the final leaderboard, while maintaining low computational cost, demonstrating that structured lightweight models can rival more complex deep-learning approaches.
- For computational efficiency, we intentionally avoid excessively long training schedules. Instead of relying on prolonged base model training, which is time-consuming for millions of samples, later stages—stacking and residual correction—are designed to refine predictions using cross-band interactions and structural reasoning. This strategy provides a better balance between accuracy and runtime in practical competition settings.

The remainder of this paper is organized as follows: Section II introduces the competition task, dataset characteristics, and evaluation protocol. Section III describes the proposed methodology, including the band-specific base models, the stacking ensemble, and the post-hoc residual correction with uncertainty estimation. Section V presents the experimental setup, ablation studies, and official leaderboard results. Finally, Section VI concludes the paper and discusses potential future work.

II. Competition Description

The FedCSIS 2025 Challenge [5] organized on the KnowledgePit platform[1] is the continuation of the highly successful first edition organized as part of the IEEE BigData Cup 2024 [1] . This second edition further extends the benchmark by

[1]https://knowledgepit.ai/

providing an updated large-scale dataset and refined evaluation protocol, aimed at advancing algorithms that estimate human-perceived puzzle difficulty. Unlike chess engines optimized for best-move accuracy, the task focuses on modeling human solving performance, a key requirement for adaptive training systems, personalized recommendation engines, and educational applications. The difficulty level is measured as the rating on the lichess platform[2].

A. Task Definition

Participants are required to predict a continuous difficulty rating for each chess puzzle, expressed as a Glicko-2 rating[3] equivalent. The official evaluation metric is the Mean Squared Error (MSE)[4]:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{1}$$

where y_i is the ground-truth human-derived rating and $\hat{y}_i$ is the predicted rating. The competition adopts a two-stage leaderboard system:

- Preliminary Stage: A public leaderboard based on a subset of the test set for iterative submissions.
- Final Stage: A private leaderboard evaluated on the full hidden test set, determining the official ranking.

B. Dataset and Features

The official dataset comprises a large labeled training set and an unlabeled test set:

- Training Set: 4,557,000 puzzles annotated with human-derived Glicko-2 ratings and engine-computed statistics.
- Test Set: 2,235 puzzles sharing the same feature structure but without difficulty ratings. Predictions for this set are used for final evaluation.

Each puzzle is described by 32 structured features in the training set and 25 features in the test set:

1) Core Information: PuzzleId, Forsyth–Edwards Notation (FEN)[5] for board state, and Portable Game Notation (PGN)[6] for solution moves.
2) Human-Performance Annotations (training only): Rating (Glicko-2 difficulty), RatingDeviation, Popularity, and NbPlays.
3) Contextual Metadata: Themes, GameUrl, and OpeningTags.
4) Engine-Derived Success Probabilities: 10 rapid-mode columns (success_prob_rapid_1050–2050) and 10 blitz-mode columns (success_prob_blitz_1050–2050), representing estimated human success rates at different skill levels.

By decoding the FEN, the chessboard can be illustrated into image. For example, the below Figure 1 shows the chessboard initial state decoded by FEN and the rating is 1300.

[2]https://lichess.org/
[3]https://en.wikipedia.org/wiki/Glicko_rating_system
[4]https://en.wikipedia.org/wiki/Mean_squared_error
[5]https://en.wikipedia.org/wiki/ForsythEdwards_Notation
[6]https://en.wikipedia.org/wiki/Portable_Game_Notation



Figure 1. FEN: "8/4R3/1p2P3/p4r2/P6p/1P3Pk1/4K3/8 w - - 1 64".

C. Key Challenges

The challenge introduces several unique difficulties:

- Human-Centric Bias: Ratings are derived from solver statistics rather than engine evaluations, so engine-trivial tactics may still be difficult for humans.
- Imbalanced Difficulty Distribution: Sparse high-Elo samples are prone to underestimation by global models trained on mid-range-dominated data.
- High-Dimensional Structured Features: Success probabilities across multiple skill bands must be effectively combined without overfitting.
- Scalability and Interpretability: Models must efficiently process millions of samples while maintaining transparency for educational use cases.

D. Extended Mask-Based Evaluation

An additional subtask evaluates uncertainty estimation. Participants submit a binary mask identifying the 10% most error-prone test puzzles. Scores are recomputed by replacing masked predictions with ground-truth ratings, and rankings are determined by the ratio between the adjusted score and the theoretical "perfect mask." This extension highlights the importance of reliable uncertainty estimation.

III. Methodology

This section presents the detailed methodology of our solution for the FedCSIS 2025 Challenge main task. The proposed approach follows a structured multi-stage pipeline designed to balance accuracy, interpretability, and computational efficiency. We first provide an overview of the entire pipeline, then highlight its key methodological contributions, followed by a detailed description of each stage.

A. Overall Pipeline

The pipeline consists of three major stages:

1) Elo-Banded Base Models: Four band-specific models are trained on separate Elo ranges using structured puzzle features (engine-derived statistics and positional indicators) to capture localized difficulty semantics and

reduce prediction bias caused by highly imbalanced rating distributions.

2) Feature-Level Stacking Ensemble: Outputs from base models are combined with structural puzzle features in a heterogeneous meta-learning framework, improving cross-band generalization.
3) Post-Processing Rating Prediction: A lightweight residual correction adjusts systematic biases, producing the final predicted ratings submitted to the competition.

This design allows predictions to be progressively refined: base models specialize in local rating regions, stacking integrates global patterns, and post-processing corrects residual systematic errors.

B. Major Contributions in Method Design

The key methodological contributions include:

- Band-Specific Specialization: Instead of a single global model, band-wise training explicitly targets the diverse difficulty distributions across Elo ranges.
- Structured Feature Utilization: Engine-derived success probabilities and handcrafted positional indicators are explicitly exploited both in base and stacking models, improving interpretability and generalization.
- Hybrid Meta-Learning: By combining a linear model (Ridge), a tree-based model (XGBoost), and a neural network (MLP), the stacking ensemble exploits complementary strengths.
- Lightweight but Scalable: All components are computationally efficient and scalable to millions of samples, unlike many deep-learning-based solutions.
- Bias Mitigation via Post-Processing: A simple, interpretable residual correction effectively addresses underestimation in high-Elo regions.

C. Step 1: Elo-Banded Base Models

Puzzle ratings span from approximately 400 Elo (basic tactics) to over 3000 Elo (master-level combinations), leading to a potential imbalanced difficulty distribution, as summarized in Fig. 2. Although high-Elo puzzles (≥ 1700) constitute over one-third of the data, their solving patterns differ substantially from lower bands, and a single global model trained on such mixed distributions often overfits mid-range samples while underestimating high-Elo puzzle difficulty.

To address this, we adopt a band-wise modeling strategy, training four independent models, each specialized for a designated Elo range [2]. This specialization allows each model to focus on localized difficulty patterns, improving prediction accuracy across heterogeneous rating bands.

1) Training Data Selection: All training samples are assigned to exactly one band-specific model according to their difficulty ratings. Unlike the default global-training setup of the baseline regression code, we manually configured the training pipeline to filter puzzles into four disjoint Elo bands:

- **Small model:** Rating $<$ 1000 (beginner-level puzzles with simple tactical motifs),

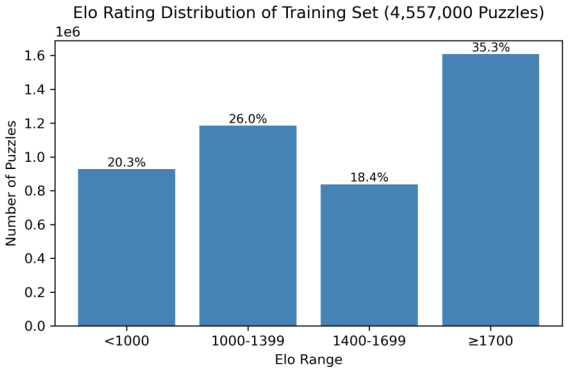


Figure 2. Elo rating distribution of the training set (4,557,000 puzzles), showing heterogeneous distributions across rating bands.

- **1300 model:** 1000 $\leq$ Rating $<$ 1400 (intermediate-level puzzles),
- **1500 model:** 1400 $\leq$ Rating $<$ 1700 (club-level puzzles with mixed tactical and strategic depth),
- **1700+ model:** Rating $\geq$ 1700 (advanced puzzles requiring deeper tactical reasoning).

This standardized partitioning ensures full data utilization and enables each model to learn the statistical patterns and solving dynamics specific to its designated Elo range, which is particularly beneficial for the sparse high-Elo band.

2) Model Architecture: All four band-specific models are trained as Multi-Layer Perceptron regressors (MLP), following a standard supervised learning framework:

- Input Representation: Each puzzle is represented entirely by structured features extracted from the official dataset:
  - Engine-derived statistics such as per-move success probabilities, mean, standard deviation, max, and min probabilities,
  - failure probability and distribution skewness,
  - material balance,
  - solution length and other positional descriptors.

  All continuous features are normalized to the range [0,1].
- Network Structure: Each model consists of several fully connected layers with ReLU activations, followed by a single linear regression head that outputs a scalar difficulty rating. This MLP architecture is well-suited to tabular structured data and provides efficient training on millions of samples.
- Training Objective: All models are optimized with the standard Mean Squared Error (MSE):

$$\mathcal{L}_{\text{base}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2, \tag{2}$$

  where f_θ is the band-specific MLP and y_i the ground-truth difficulty rating.
- Epochs and Fine-Tuning: The number of training epochs is determined empirically based on validation performance, typically ranging from 10 to 20 epochs depending

on band size, model convergence speed, and computational budget. This relatively small epoch range reflects a trade-off between convergence and time efficiency, as prolonged training provides diminishing returns and is computationally expensive for large-scale data.
An internal validation set, sampled as 10% of the band-specific training data, is used exclusively for hyperparameter tuning and to save the best-performing checkpoints based on validation MSE.

3) Inference and Outputs: After training, each model is applied independently to the official test set using a standardized inference pipeline. The four models generate separate predictions corresponding to the small, 1300, 1500, and 1700 Elo bands. These predictions form the base inputs for the stacking ensemble in Step 2.

D. Step 2: Stacking Ensemble

1) Motivation: While band-specific models effectively capture localized difficulty patterns, they lack global consistency and fail to fully exploit cross-band correlations. A meta-learning strategy can integrate predictions from multiple bands and leverage structural attributes to correct residual inconsistencies. In particular, features such as failure probability and move counts provide complementary information about puzzle-solving dynamics that is not fully encoded in the band-wise models.

2) Meta-Feature Construction: A comprehensive meta-feature vector is constructed by combining base-model predictions and structural puzzle attributes:

$$\begin{aligned}\mathbf{z} = [&p_1,\ p_2,\ p_3,\ p_4,\ \tilde{p},\ \sigma_p,\\ &\text{avg_success},\ \text{fail_prob},\ \text{inflection_rating},\\ &\text{fail_skew},\ \text{num_move}]\end{aligned} \tag{3}$$

where:

$$\tilde{p} = \text{median}(p_i), \qquad \sigma_p = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(p_i - \tilde{p})^2}. \tag{4}$$

Here, $\tilde{p}$ represents the robust central tendency of the base predictions, while σ_p serves as an implicit confidence measure, indicating inter-model disagreement.

3) Meta-Learners and Complementarity: The stacking ensemble integrates three heterogeneous meta-learners chosen for their complementary modeling capacities:

- Ridge Regression: A linear regression model with L_2 regularization, which stabilizes coefficient estimation by penalizing large weights. Ridge captures global linear trends between structural attributes (e.g., average success probability, number of moves) and target ratings. Its interpretability provides valuable insight into the relative importance of meta-features.
- XGBoost: An ensemble of gradient-boosted regression trees that sequentially fits residual errors. XGBoost is well-suited for modeling nonlinear feature interactions and conditional relationships, such as detecting puzzles that are deceptively difficult despite short move sequences.
- MLP Regressor: A lightweight feedforward neural network configured with two hidden layers (128 and 64 neurons, ReLU activation). This configuration provides a balance between model capacity and overfitting risk for the low-dimensional meta-feature vector, while effectively capturing high-order nonlinear dependencies not easily approximated by tree-based methods.

4) Prediction Aggregation: The outputs of the three meta-learners are aggregated via an adaptive weighted average:

$$\hat{y}_{stack} = w_{\text{ridge}}\hat{y}_{\text{ridge}} \ + \ w_{\text{xgb}}\hat{y}_{\text{xgb}} \ + \ w_{\text{mlp}}\hat{y}_{\text{mlp}}, \tag{5}$$

where the weights are inversely proportional to the validation residual variance:

$$w_j = \frac{1/\sigma_j^2}{\sum_k (1/\sigma_k^2)}, \qquad j \in \{\text{ridge}, \text{xgb}, \text{mlp}\}. \tag{6}$$

This dynamic weighting emphasizes models with more stable validation performance. Ridge contributes stability and interpretability, XGBoost captures local conditional interactions, and the MLP learns complex nonlinear relationships, resulting in improved generalization across heterogeneous puzzle types.

E. Step 3: Post-Processing Rating Prediction

1) Motivation: Although the stacking ensemble improves overall prediction accuracy, residual analysis reveals systematic biases: high-Elo puzzles tend to be underestimated, while some low-Elo puzzles are slightly overestimated. These biases arise because the meta-learners only indirectly exploit engine-derived structural signals, which often contain strong priors about puzzle difficulty.

2) Structure-Aware Residual Correction: To explicitly incorporate these priors, we introduce a lightweight residual correction that adjusts the stacking predictions toward a refined structural difficulty estimate.

First, a baseline structural estimate s is computed as a failure-probability-weighted average of rating buckets:

$$s = \frac{\sum_j (1 - \bar{p}_j)\, r_j}{\sum_j (1 - \bar{p}_j)}, \tag{7}$$

where $(1 - \bar{p}_j)$ represents the failure probability in bucket r_j. This estimate is then refined by incorporating two additional structural signals:

- The inflection point r_{inf}, indicating the rating at which the success probability changes most sharply.
- The failure-probability skewness γ, which captures asymmetric difficulty distributions where certain skill groups systematically misjudge a puzzle. highlighting difficulty transitions perceived by human solvers;

The final refined structural estimate is defined as:

$$s^* = 0.6\, s \ + \ 0.2\, r_{\text{inf}} \ + \ 0.2\,(1700 + 300\,\gamma) \tag{8}$$

Finally, the corrected rating prediction is obtained through a residual fusion with the stacking output:

$$\hat{y}_{\text{final}} = \hat{y}_{\text{stack}} \; + \; \lambda \left(s^* - \hat{y}_{\text{stack}} \right), \qquad \lambda = 0.3. \tag{9}$$

This correction introduces no additional learnable parameters and is computationally efficient. By explicitly leveraging interpretable structural priors, it consistently improves accuracy, particularly in high-Elo regions where data sparsity makes residual biases more severe.

IV. Mask-Based Uncertainty Evaluation

In addition to the main rating prediction task, the FedCSIS 2025 Challenge introduced an extended evaluation explicitly designed to assess uncertainty estimation. Participants were required to submit a binary mask marking the 10% of test puzzles most likely to be mispredicted by their models. The final score was recalculated by replacing predictions for these masked samples with ground-truth ratings, and the ratio between the adjusted score and the theoretically optimal score determined the subtask ranking.

We designed a structure-aware uncertainty estimation method that integrates model disagreement, prediction instability, and structural difficulty indicators into a unified scoring framework. This approach balances model-based and feature-based uncertainty cues, improving the alignment of selected samples with actual model errors.

A. Scoring Function Design

The composite uncertainty score for each puzzle is defined as:

$$u_i = \alpha \cdot \sigma_{p,i} \; + \; \beta \cdot \delta_i \; + \; \gamma \cdot \psi_i \tag{10}$$

where:

- $\bar{p}_i$: average success probability across rating buckets for puzzle i,
- z_i: Z-score of the material balance in the training data; for the test set, this term defaults to zero as no material information is provided,
- $\mathbb{I}(|z_i| > \tau)$: indicator of extreme material imbalance, with $\tau = 3.0$ as the outlier threshold.

The weighting coefficients were empirically set to $\alpha = 1.0$, $\beta = 1.0$, and $\gamma = 0.4$ to optimize the uncertainty ranking quality.

B. Mask Generation

Puzzles are ranked according to u_i, and the top 10% are assigned a mask value of 1:

$$\text{mask}_i = \begin{cases} 1, & i \text{ in top } 10\% \text{ highest scores}, \\ 0, & \text{otherwise}. \end{cases} \tag{11}$$

This binary mask was submitted separately from the rating predictions, providing an explicit measure of the model's uncertainty estimation capability.

C. Discussion

The proposed uncertainty estimation method combines model-driven and feature-driven indicators in a single interpretable framework. Its main advantages include:

- improved consistency between selected samples and actual prediction errors compared to variance-only methods;
- and explicit consideration of structural puzzle difficulty, beyond pure statistical disagreement.

However, the method still relies on the alignment of raw and post-processed predictions and may overlook systematic biases shared across all models. Despite these limitations, the approach provides a strong baseline for uncertainty-aware chess puzzle difficulty estimation.

Our uncertainty mask ratio is 1.696, placing us 8th out of the 9 teams that chose to participate in this additional task. Using our submitted mask, our final score is approximately 57,931, while using a "perfect mask" would yield a score of about 34,162. The full results for all teams will be published in the competition report [5].

V. Experimental Setup

A. Dataset and Preprocessing

The dataset provided by the organizers consists of two disjoint parts:

- Training set: Approximately 4,557,000 puzzles, each labeled with a human-perceived difficulty rating derived from aggregated player performance in Lichess (Glicko-2 system). This set is fully annotated and serves as the only source of labeled data for model development.
- Test set: An unlabeled set of 2,235 puzzles used exclusively for the final evaluation. Participants are required to submit predicted ratings for this set, while the ground-truth ratings remain hidden.

To ensure balanced representation across different rating bands, the official training set is split using stratified sampling within each Elo band:

- Base-model training subset: 90% of puzzles per Elo band, used for training the four band-specific base models.
- Validation subset: 10% of puzzles per Elo band, reserved exclusively for generating unbiased out-of-sample predictions for stacking and for residual bias calibration.

The stacking meta-learners are trained solely on this held-out validation subset to prevent data leakage. To improve generalization within the stacking stage, a 10-fold cross-validation is performed on the validation subset: in each fold, the meta-learners are trained on 90% of the fold and validated on the remaining 10%, and the out-of-fold predictions are averaged for final submission.

All continuous features, including success probabilities, failure distributions, and structural statistics, are normalized to the [0,1] range. PGN sequences are parsed to compute derived attributes such as the number of solution moves (*num_move*), and FEN strings are used to extract positional features such as castling rights and piece material balance.

B. Training Details

Each Elo-banded base model is implemented as a lightweight multi-layer perceptron (MLP) following the official baseline recommendations. The models are trained for 20–30 epochs with early stopping based on validation loss. The best-performing checkpoints are selected, and no cross-validation is applied at the base-model stage to keep training computationally efficient.

The meta-learner integrates Ridge regression (L2 regularization), XGBoost (max depth = 8, learning rate = 0.01-0.05, 1000-2000 boosting rounds) and a lightweight MLP (two hidden layers: 128 and 64 neurons, ReLU activation, dropout = 0.1). The meta-features are constructed exclusively from base model predictions on the validation subset (out-of-sample), and a 10-fold cross-validation within this subset is used to improve the robustness of the meta-learners.

A structure-aware residual correction is fitted to the validation residuals, explicitly adjusting the stacking predictions toward structural difficulty estimates, especially mitigating underestimation in sparse high-Elo regions.

C. Overall Performance

Table I shows the progressive improvements across pipeline stages.

Table I
PROGRESSIVE IMPROVEMENTS ACROSS PIPELINE STAGES (INTERNAL HOLD-OUT TEST SET).

Pipeline Stage	MSE	Relative Gain
Base Models (avg/median)	78,000–80,000	–
Stacking Ensemble (10-Fold)	71,000–73,000	+9–11%
Post-Processing + Best Avg	66,600–68,000	+6–8%

The pipeline shows clear incremental gains: stacking substantially improves cross-band consistency, while residual correction yields additional improvements in high-Elo regions.

D. Ablation Study

Table II highlights the contribution of each major component.

Table II
ABLATION STUDY: CONTRIBUTION OF EACH COMPONENT (VALIDATION SET).

Configuration	MSE	Change vs. Full
Full Pipeline	66,600–68,000	Baseline
w/o Structural Features	∼72,000	+6–8%
w/o Stacking Ensemble	∼78,000	+15–18%
Single Global Model	>85,000	+25%

Removing structural features increases variance, confirming their role as complementary difficulty indicators. Stacking yields the largest single improvement by integrating base-model predictions with structural cues, while replacing band-wise models with a single global model leads to severe bias, particularly for high-Elo puzzles.

E. Discussion

These results confirm the effectiveness of a structured and interpretable pipeline:

- Elo-banded modeling alleviates data imbalance and reduces extreme rating bias;
- Stacking with structural reasoning provides the largest performance gain by capturing cross-band interactions;
- Residual calibration yields additional improvements in sparse high-Elo regions with negligible computational cost.

The final MSE of 62,685 ranks among the top solutions, demonstrating that a lightweight and transparent ensemble can rival more computationally expensive deep-learning models, making it suitable for educational or large-scale online puzzle recommendation systems.

VI. CONCLUSION

This study presented a structured and interpretable pipeline for chess puzzle difficulty prediction in the FedCSIS 2025 Challenge. The approach combines three complementary stages: (1) Elo-banded base models specialized for different rating ranges to reduce distributional bias, (2) a feature-level stacking ensemble that integrates base predictions with structural puzzle attributes to improve cross-band generalization, and (3) a lightweight structure-aware residual correction to mitigate systematic errors, particularly in sparse high-Elo regions.

Extensive experiments demonstrated clear incremental improvements at each stage, achieving a final MSE of 62,685 and ranking 7$^{\text{th}}$ in the final leaderboard among resource-intensive deep-learning solutions. The results confirm that carefully designed ensembles, when combined with domain-specific structural reasoning, can achieve competitive accuracy with far lower computational cost and higher interpretability.

REFERENCES

[1] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8423-8429, doi: 10.1109/BigData62323.2024.10825289.

[2] T. Woodruff, O. Filatov and M. Cognetta, "The bread emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8415-8422, doi: 10.1109/BigData62323.2024.10826037.

[3] K. Miłosz and A. Kapusta, "GlickFormer: A Spatio-Temporal Transformer for Chess Puzzle Difficulty Prediction," in *Proc. IEEE BigData Conf.*, 2024.

[4] D. Ruta, M. Liu and L. Cen, "Moves Based Prediction of Chess Puzzle Difficulty with Convolutional Neural Networks," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8390-8395, doi: 10.1109/BigData62323.2024.10825595.

[5] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proc. 20th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 43, Polish Inf. Process. Soc., 2025. doi: http://dx.doi.org/10.15439/2025F5937.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 825–830
DOI: 10.15439/2025F3675
ISSN 2300-5963 ACSIS, Vol. 43

Hybrid Boosting and Multi-Modal Fusion for Chess Puzzle Difficulty Prediction

Ming Liu
Jintai Stratech
China
lmymnew@gmail.com

Junye Wang
University of Science and Technology of China
China
wangjunye@mail.ustc.edu.cn

Yinghan Hu
Jilin University
China
h796431@qq.com

Xiaolin Yang
Inspur
China
yangxl@inspur.com

Defu Lin
Beijing Institute of Technology
China
lindf@bit.edu.cn

***Abstract*—The FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty tasked participants with estimating puzzle ratings directly from board states and solution sequences, without relying on human solver statistics.**

We propose a three-stage hybrid framework integrating gradient-boosting regressors, a multi-modal neural network, and an XGBoost stacking ensemble. The boosting stage modeled handcrafted structural features derived from FEN and engine metadata, while the multi-modal network jointly learned from structured features and image-rendered chessboards to capture positional and tactical patterns. The residual-based stacking stage explicitly modeled prediction errors to correct systematic biases and enhance performance, particularly for high-difficulty puzzles.

Our method achieved a competitive performance, ranking 7th in the preliminary stage and 8th in the final leaderboard. These results demonstrate that combining interpretable boosting models with visual-tactical deep representations and meta-learning provides a robust and computationally efficient alternative to large-scale transformer-based approaches.

***Index Terms*—Chess puzzle difficulty prediction, Gradient boosting, Multi-modal learning, Deep learning, Residual-based stacking, Structural feature engineering, Uncertainty estimation**

I. Introduction

THE prediction of chess movements has evolved dramatically since the landmark achievement of IBM's Deep Blue in 1997[1], when it famously defeated world champion Garry Kasparov. Automated prediction of chess puzzle difficulty plays an increasingly important role in online training platforms, enabling adaptive puzzle recommendation and accurate tracking of player progression.

Early research focused on handcrafted features, employing gradient-boosting or support vector regressors trained on material balance, king safety, and piece mobility [2], [3]. Although these models lacked scalability, they demonstrated the importance of domain knowledge and interpretability in difficulty estimation. With the advent of deep learning, convolutional neural networks (CNNs) [4] were introduced to treat the chessboard as an image, successfully capturing spatial and tactical patterns such as attacking piece clusters and exposed kings. Later, hybrid CNN-LSTM [5] architectures integrated sequential move information, showing that temporal reasoning improves alignment with human-rated difficulty. Transfer learning approaches, such as DeepChess [6], adapted game prediction networks for puzzle rating, implicitly learning tactical patterns from large-scale game data.

More recently in the IEEE Big Data Cup 2024 [7], transformer-based architectures, such as GlickFormer [8], achieved state-of-the-art results by modeling spatio-temporal dependencies in move sequences. However, these models are computationally expensive and thus less practical for real-time puzzle rating systems. Competition-oriented solutions, such as the Bread Emoji team's hybrid ensemble combining engine-derived success probabilities with neural embeddings [9], have demonstrated that combining interpretable features with lightweight neural models can remain competitive while being computationally efficient.

Our team has an extensive history of successful participation in data science competitions hosted on the KnowledgePit platform[1]. We have consistently leveraged Gradient Boosting Decision Tree (GBDT) algorithms to tackle a wide range of predictive tasks—including classification, regression, forecasting, and image recognition—achieving top-ranked results and earning multiple awards[10] - [20]. Motivated by this strong background, we approach the present challenge with the same commitment to excellence.

Our work proposes a hybrid three-stage framework designed to achieve high predictive accuracy while maintaining interpretability and efficiency. We integrate 3 parts, which are

- three gradient-boosting models for robust tabular predictions,
- a multi-modal neural network to extract visual-tactical cues from rendered chessboard images,
- and an XGBoost stacking ensemble to fuse predictions into a single optimized output.

Additionally, we extend the method with a mask-based uncertainty estimation task, where the goal is to identify the most error-prone puzzles.

The remainder of this paper is organized as follows. Section II describes the challenge setup, dataset, and evaluation metric. Section III details the proposed methodology. Section IV presents the mask extension for uncertainty estimation. Section V reports experimental results and ablation studies. Finally, Section VI concludes the paper and discusses future research directions.

[1]https://knowledgepit.ai/

II. Challenge Description

The FedCSIS 2025 Challenge [21] is the second edition of IEEE BigData Cup 2024 chess puzzle competition [7] on Predicting Chess Puzzle Difficulty addressed the problem of estimating puzzle difficulty ratings directly from board configurations and solution sequences. The task was formulated as a regression problem, where each puzzle was assigned a continuous difficulty rating analogous to Lichess[2] puzzle ratings, typically ranging between 800 and 2800. These ratings approximate the skill level of players expected to solve the puzzle with a 50% success probability, making the task closer to modeling human cognitive difficulty than engine evaluation.

The dataset, derived primarily from real games on Lichess, contained tens of thousands of puzzles. Each puzzle was described by:

- FEN (Forsyth–Edwards Notation)[3]: Encodes the board state at the start of the puzzle, including piece placement, side to move, castling rights, and en passant possibilities.
- PGN[4] moves: The sequence of solution moves forming the intended tactical line.
- Puzzle rating: The target variable representing human-perceived puzzle difficulty.
- Optional metadata: Additional information such as puzzle tags or themes, which some participants used as auxiliary features.

The dataset was divided into training and test sets, with ground-truth ratings provided only for the training set. No official validation split was released, requiring participants to design their own validation protocols. In our case, a random 10% split of the training set was used to approximate unseen data. The rating distribution was broad but skewed, with most puzzles clustered in the intermediate range (1400–2000) and long tails toward very easy and extremely difficult puzzles. This imbalance increased the impact of errors on rare high-difficulty puzzles, as large deviations in such cases significantly influenced the evaluation metric.

A training chess board initial state sample decoded by FEN was shown in the below Figure 1 with the rating 1575.

Submissions were scored using the Mean Squared Error(MSE)[5] between predicted ratings $\hat{y}_i$ and ground-truth ratings y_i:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{1}$$

where N denotes the number of test samples. Because MSE penalizes large deviations quadratically, extreme mispredictions (e.g., predicting 1500 for a puzzle rated 2300) had a disproportionately large effect, making robust handling of such outliers crucial.

In addition to these technical considerations, the challenge imposed practical constraints. The test set was unlabeled,

[2]https://lichess.org/
[3]https://en.wikipedia.org/wiki/ForsythEdwards_Notation
[4]https://en.wikipedia.org/wiki/Portable_Game_Notation
[5]https://en.wikipedia.org/wiki/Mean_squared_error



Figure 1. A training chess board state sample of rating 1575 and FEN "8/8/4k1p1/2KpP2p/5PP1/8/8/8 w - - 0 53".

which increased the risk of overfitting without carefully designed validation procedures. The diversity of puzzle types, i.e. forced checkmates, defensive resource puzzles, and quiet positional tactics, made simple statistical baselines inadequate to capture nuanced difficulty differences.

Finally, there was a trade-off between accuracy and interpretability. Deep neural networks, while expressive, are prone to overfitting or unstable predictions with limited training data, whereas classical machine learning models often fail to capture sequential and visual information. Consequently, prior top-performing approaches combined interpretable boosting models with engine-derived features to balance robustness and computational efficiency. Our method was designed with these considerations in mind, integrating complementary model strengths to improve generalization while maintaining interpretability.

III. Methodology

Our solution follows a three-stage hybrid pipeline that integrates gradient-boosting models, a multi-modal neural network, and an XGBoost-based stacking ensemble. This design leverages the stability of boosting methods for structured data, while incorporating deep neural representations to capture tactical and spatial patterns from chessboard images.

A. Gradient-Boosting Base Models

The first stage of our framework employs gradient boosting, an ensemble learning technique that constructs a strong predictive model by combining multiple weak learners in an additive manner. Given a differentiable loss function $\mathcal{L}(y, F(x))$, gradient boosting iteratively fits a new base learner to the negative gradient of the loss with respect to the current prediction. At the m-th iteration, the ensemble is updated as:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x), \tag{2}$$

where $F_m(x)$ is the updated ensemble prediction, and $h_m(x)$ is the weak learner trained on the residuals:

$$r_{i,m} = -\frac{\partial \mathcal{L}(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}, \tag{3}$$

and $\eta \in (0, 1]$ is the learning rate controlling the contribution of each learner. This formulation effectively reduces bias while controlling variance, making gradient boosting well-suited for modeling tabular data with complex non-linear interactions.

a) XGBoost: XGBoost (Extreme Gradient Boosting) [3] introduces second-order gradient optimization and sparsity-aware split finding, with regularization to control model complexity. The overall objective function is:

$$\mathcal{L}_{\text{XGB}} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \left(\gamma T_k + \frac{1}{2} \lambda \sum_j w_{kj}^2 \right), \quad (4)$$

where:

- T_k is the number of leaves in the k-th decision tree,
- w_{kj} is the weight of leaf j in tree k,
- γ and λ are regularization coefficients controlling tree complexity and weight shrinkage.

This regularized formulation makes XGBoost robust against overfitting and variance, making it particularly effective for structured chess features.

b) LightGBM: LightGBM (Light Gradient Boosting Machine) [2] is optimized for high-dimensional tabular data through histogram-based feature binning and a leaf-wise growth strategy with depth constraints. Its split gain is computed as:

$$\text{Gain} = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma, \quad (5)$$

where G_L, H_L and G_R, H_R are the accumulated gradients and Hessians of the left and right nodes. This strategy balances accuracy and efficiency, making LightGBM a reliable baseline for modeling chess-specific structured features.

c) CatBoost: CatBoost [22] is designed to handle categorical variables natively, avoiding target leakage through permutation-driven encoding. Its ordered target encoding for a categorical feature is:

$$\hat{y}_{cat} = \frac{\sum_{i<j} y_i + p}{n + q}, \quad (6)$$

where p and q are prior parameters for Bayesian smoothing, and n is the number of preceding samples. This makes CatBoost particularly effective for categorical chess features, such as castling rights and the side-to-move indicator.

d) Input Features and Training Objective: All three models were trained on handcrafted features extracted from FEN, PGN, and engine metadata:

- **Structural features:** Material balance, piece counts for both sides, castling rights, and check status—key indicators of positional complexity.
- **Move count:** Total number of moves in the puzzle's solution sequence, often correlated with tactical depth.
- **Engine-derived probabilities:** Success probabilities for rapid and blitz rating buckets across Elo levels.
- **Average success rate:**

$$s_{\text{avg}} = 0.5 \times (s_{\text{rapid}} + s_{\text{blitz}}), \quad (7)$$

providing a balanced indicator of expected human solvability.

- **Aggregated statistics:** Maximum and standard deviation of success probabilities ($s_{\max}, s_{\text{std}}$) to capture tactical ambiguity—high variance typically indicates multiple equally strong candidate moves, increasing cognitive complexity.

All boosting models were optimized with the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \quad (8)$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted ratings.

e) Rationale: This boosting stage provides a strong and interpretable baseline for structured features and produces complementary predictions for the stacking stage, where diverse model biases are effectively combined.

B. Multi-Modal Neural Network

While gradient-boosting models are effective for structured tabular features, they are inherently limited in capturing spatial and visual cues that strongly influence human-perceived puzzle difficulty. To address this limitation, we designed a multi-modal neural network that jointly learns from structured numeric features and visual representations of chessboard configurations.

a) Structured Feature Encoder: The numeric branch encoded the same handcrafted features as in the boosting stage, enriched with additional structural and interaction terms to capture tactical and positional complexity more effectively:

- **New structural features:** Piece density, defined as:

$$d_p = \frac{n_w + n_b}{64}, \quad (9)$$

where n_w and n_b denote the counts of white and black pieces, respectively, serves as a compact measure of board congestion. A binary last-move success flag indicates whether the most recent move had a high engine-predicted success probability. Cross features were introduced to model interactions between key factors:

$$f_1 = s_m \cdot m_b, \quad (10)$$

$$f_2 = n_m \cdot s_{\text{avg}}, \quad (11)$$

where s_m is the side-to-move indicator (1 for White, 0 for Black), m_b is the material balance, n_m is the total move count, and s_{avg} is the average engine-predicted success probability.

- **Engine statistics:** Aggregated success probabilities, including mean (s_{mean}), maximum ($s_{\max}$), and standard deviation (s_{std}), were used to capture tactical ambiguity. High variance (s_{std}) typically indicates multiple equally strong candidate moves, increasing cognitive difficulty for human players.

The numeric features were standardized and passed through two fully connected layers (64 and 32 neurons, ReLU activation). Formally:

$$\mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \tag{12}$$

$$\mathbf{z}_{\text{num}} = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \tag{13}$$

where $\mathbf{x}$ is the standardized feature vector, $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ are trainable parameters, and $\sigma(\cdot)$ denotes the ReLU activation.

b) Image Feature Encoder: Each FEN string was converted into a chessboard image using `python-chess` and `cairosvg`, enabling the network to learn spatial and tactical patterns that are difficult to model through explicit numeric features alone. Accurate representation of visual configurations is crucial because human-perceived difficulty is strongly influenced by positional complexity, such as piece clustering, open lines, and king safety, which are more naturally encoded in a spatial format.

A single high-capacity convolutional backbone was adopted. We selected EfficientNetB3 [23] due to its superior trade-off between accuracy and computational cost. Its compound scaling strategy jointly optimizes network depth, width, and input resolution, allowing fine-grained tactical features—such as discovered attacks or forced mating nets—to be captured effectively. The backbone was initialized with ImageNet-pretrained weights and fine-tuned on the chess puzzle dataset to adapt to domain-specific patterns.

Let $\phi_{\text{E3}}(\cdot)$ denote the EfficientNetB3 backbone and $GAP(\cdot)$ the Global Average Pooling operation. The encoded visual representation is given by:

$$\mathbf{z}_{\text{img}} = GAP\left(\phi_{\text{E3}}(I_{\text{fen}})\right), \tag{14}$$

where I_{fen} is the rendered chessboard image.

c) Fusion and Output Layer: The visual embedding $\mathbf{z}_{\text{img}}$ was concatenated with the numeric branch embedding $\mathbf{z}_{\text{num}}$ to form a joint latent representation:

$$\mathbf{z}_{\text{fusion}} = \begin{bmatrix} \mathbf{z}_{\text{num}} \\ \mathbf{z}_{\text{img}} \end{bmatrix}, \tag{15}$$

which was passed through a fully connected fusion head:

$$\hat{y} = W_3\,\sigma(W_2\,\sigma(W_1\,\mathbf{z}_{\text{fusion}} + b_1) + b_2) + b_3, \tag{16}$$

where $\sigma(\cdot)$ is the ReLU activation. A dropout layer ($p = 0.3$) was applied after the first dense layer to reduce overfitting by encouraging robustness to co-adaptations between visual and numeric features.

d) Training Procedure: The network was trained end-to-end using the Adam optimizer (learning rate $= 10^{-4}$) and the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{17}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted difficulty ratings for sample i. EarlyStopping and ReduceLROnPlateau strategies were applied to prevent overfitting and improve convergence stability.

Despite its moderate computational requirements, the inclusion of a visual branch substantially improved predictions, particularly for high-difficulty puzzles where spatial complexity such as multi-piece coordination or long forced sequences plays a crucial role in human perception.

C. Residual-Based XGBoost Stacking Fusion

The final stage refines the predictions by explicitly modeling the residual errors from the first two stages. Instead of directly stacking the base predictions, we train an XGBoost meta-learner to learn the systematic residuals and correct the base prediction accordingly.

a) Base Prediction and Residual Definition: Let $\mathbf{p}_{\text{boost}} \in \mathbb{R}^3$ denote the three boosting model predictions from Step 1, and $p_{\text{nn}} \in \mathbb{R}$ denote the multi-modal neural network prediction from Step 2. The base prediction $\bar{p}_i$ for sample i is defined as the simple average of all four models:

$$\bar{p}_i = \frac{1}{4}\left(p^{(1)}_{\text{boost},i} + p^{(2)}_{\text{boost},i} + p^{(3)}_{\text{boost},i} + p_{\text{nn},i}\right), \tag{18}$$

although in practice, weighted averages were also evaluated during validation.

The residual for each sample is computed as:

$$r_i = y_i - \bar{p}_i, \tag{19}$$

where y_i is the ground-truth puzzle rating.

b) Stacking Input Construction: The XGBoost meta-learner is trained to predict the residuals r_i rather than the final ratings directly. Its input vector is constructed from:

- The individual residual components of each base learner:
$$r_i^{(k)} = y_i - p_i^{(k)}, \quad k = 1, 2, 3, \text{nn}, \tag{20}$$
where $p_i^{(k)}$ is the prediction of the k-th base learner.
- Key structural features correlated with human difficulty:
$$\mathbf{f}_{\text{key},i} = \begin{bmatrix} s_{\text{avg},i} \\ n_{m,i} \\ m_{b,i} \end{bmatrix}, \tag{21}$$
where $s_{\text{avg},i}$ is the average success rate (Eq. 7), $n_{m,i}$ the move count, and $m_{b,i}$ the material balance.

The complete stacking input is:

$$\mathbf{x}_{\text{stack},i} = \begin{bmatrix} r_i^{(1)} \\ r_i^{(2)} \\ r_i^{(3)} \\ r_i^{(\text{nn})} \\ \mathbf{f}_{\text{key},i} \end{bmatrix} \in \mathbb{R}^7. \tag{22}$$

c) Meta-Learner Training and Final Prediction: The XGBoost meta-learner $\mathcal{F}_{\text{XGB}}$ minimizes the Mean Squared Error (MSE) of the residuals:

$$\mathcal{L}_{\text{stack}} = \frac{1}{N}\sum_{i=1}^{N}\left(r_i - \hat{r}_i\right)^2, \tag{23}$$

where $\hat{r}_i = \mathcal{F}_{\text{XGB}}(\mathbf{x}_{\text{stack},i})$ is the predicted residual.

Finally, the corrected prediction for sample i is obtained by adding the predicted residual to the base prediction:

$$\hat{y}_{\text{final},i} = \bar{p}_i + \hat{r}_i. \tag{24}$$

d) Rationale: This residual-based stacking approach effectively treats the meta-learner as a non-linear residual corrector. By modeling residuals instead of direct predictions, the meta-learner focuses on learning systematic errors of the base models, such as the underestimation of high-difficulty puzzles. Incorporating structural features (Eq. 21) further allows the meta-learner to exploit domain-specific correlations between residual errors and puzzle characteristics, yielding significant accuracy gains, especially for puzzles with ratings above 2200.

IV. Mask Prediction – Competition Extension

In addition to the main regression task, the organizers introduced an optional extension to evaluate a model's uncertainty estimation ability. Participants were required to identify the 10% of test puzzles for which their predictions were most likely to be erroneous. By replacing predicted ratings for these puzzles with their ground-truth values, the leaderboard score was recomputed, providing an indirect measure of a model's ability to assess its own confidence. Formally, each submission consisted of a binary mask $M \in \{0,1\}^N$ satisfying

$$M_i = \begin{cases} 1, & \text{if puzzle } i \text{ is highly uncertain (masked)}, \\ 0, & \text{otherwise}, \end{cases}$$
$$\text{s.t.} \quad \sum_{i=1}^{N} M_i = N \times 10\% \tag{25}$$

where N is the total number of test samples. The evaluation used two scores: the *Perfect Score* P, defined as the minimum achievable MSE if the top 10% highest-error samples were perfectly masked, and the *New Score* N, the recomputed MSE after replacing predictions at masked indices with ground truth. The optimization objective was to minimize the ratio

$$\text{score} = \frac{N}{P}, \tag{26}$$

with the optimal value approaching 1.

The mask task can be interpreted as an uncertainty-ranking problem, where an ideal mask should prioritize puzzles whose predictions are expected to deviate most from the true rating. Our heuristic design followed two intuitive assumptions:

- samples whose predicted ratings deviate strongly from the overall rating distribution are more likely to be erroneous,
- and puzzles with longer solution sequences tend to involve deeper tactical reasoning and are thus harder for both humans and models.

To operationalize these assumptions, we designed a deterministic composite uncertainty score. For each puzzle i, the score is defined as

$$\text{mask_score}_i = 0.6\, u_i + 0.4\, c_i, \tag{27}$$

where u_i measures normalized prediction deviation and c_i quantifies move-based complexity:

$$u_i = \frac{|\hat{y}_i - \mu_y|}{\sigma_y}, \tag{28}$$
$$c_i = \frac{n_{m,i} - \mu_{n_m}}{\sigma_{n_m}}, \tag{29}$$

with $\hat{y}_i$ denoting the predicted rating, μ_y and σ_y the mean and standard deviation of predicted ratings, $n_{m,i}$ the move count of puzzle i, and μ_{n_m}, σ_{n_m} its distribution statistics. The top 10% samples with the highest mask_score_i were selected as

$$M_i = \begin{cases} 1, & \text{if } \operatorname{rank}(\text{mask_score}_i) \le 0.1N, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

This rule required no additional calibration or access to ground-truth labels, making it computationally efficient and stable. Its main advantages are simplicity, consistency, and domain relevance, as it incorporates move count, a known correlate of puzzle difficulty.

Based on the competition report[21], our uncertainty mask ratio is equal to 1.648. It gives us the 6th place among 9 teams that decided to participate in this additional task. Our final score with the submitted mask is approximately equal to 56563. And our final score with the perfect mask would be equal to 34312.

V. Experimental Results

This section presents the empirical evaluation of our method, including ablation analysis and leaderboard performance, followed by a discussion of the contributions of each stage and their implications for generalization.

All experiments were conducted on the official training dataset. For the first and third stages, we employed a 10-fold cross-validation strategy to fully exploit the available data and obtain stable validation estimates. In each fold, 90% of the data were used for training and 10% for validation, and the final stage-wise performance was averaged across folds. For the second stage, due to the higher computational cost of CNN training, a fixed 10% hold-out validation set stratified by rating range was used. All boosting models were implemented with the official LightGBM, CatBoost, and XGBoost libraries, while the multi-modal neural network was implemented in TensorFlow/Keras. Early stopping based on validation MSE was applied for all models.

Our final submission ranked 7th in the preliminary stage and 8th in the final leaderboard. The public preliminary MSE was 66,658, and the private test MSE was 63,009, confirming good generalization. Table I summarizes the performance across all stages and the final leaderboard results.

The first stage achieved an average MSE of 87,378 across boosting models under 10-fold cross-validation, establishing a strong tabular baseline. The second stage reduced the MSE to 78,379 (+10.3%), demonstrating that the visual-tactical features extracted by the EfficientNetB3 backbone provided complementary information, particularly for complex tactical puzzles.

The residual-based XGBoost stacking in the third stage brought the most significant improvement, reducing the error to 68,029 (+22.1%). By explicitly modeling residuals and using 10-fold cross-validation to stabilize training, the meta-learner effectively captured systematic error patterns, particularly in high-rating puzzles where difficulty estimation is highly non-linear.

Table I
OVERALL RESULTS: ABLATION STUDY AND LEADERBOARD PERFORMANCE

Stage	Public LB MSE	Improvement (%)	Remarks
Best Average of Boosting Models (10-fold)	87,378	–	Average predictions from LightGBM, CatBoost, and XGBoost
Multi-Modal CNN (EfficientNetB3, hold-out)	78,379	10.3	Visual-tactical features complement structured predictions
Residual XGBoost Stacking (10-fold)	68,029	22.1	Explicit residual modeling, effective in high-rating puzzles
Final Ensemble Averaging (Public LB)	66,658	23.7	Combines top models for stable leaderboard score (7th place)
Private Test MSE (Final Ensemble)	*63,009*	*27.9*	*Generalizes well to unseen data (8th place)*

Finally, averaging several top-performing models achieved a public leaderboard MSE of 66,658 (+23.7%) and a private test MSE of 63,009 (+27.9%), confirming strong generalization. The relatively small gap between the public leaderboard and private test performance suggests that overfitting was effectively controlled, and the residual modeling contributed to robust predictions even for rare high-difficulty cases.

VI. CONCLUSION AND FUTURE WORK

This paper presented a hybrid framework for predicting chess puzzle difficulty in the FedCSIS 2025 Challenge on Predicting Chess Puzzle Difficulty. By integrating gradient-boosting models, a multi-modal neural network, and residual-based XGBoost stacking, our approach achieved a competitive 7th and 8th places in the preliminary and final stages, relatively, with a MSE score of 63,009. The combination of interpretable handcrafted features and learned visual-tactical representations proved effective, offering a robust and computationally efficient alternative to transformer-based methods.

VII. ACKNOWLEDGMENT

This work was supported by Jilin Provincial Science and Technology Department Project (No. 20230508035RC).

REFERENCES

[1] https://www.ibm.com/history/deep-blue.
[2] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.
[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[4] B. Oshri and N. Khandwala, “Predicting Moves in Chess Using Convolutional Neural Networks (ConvChess),” *Stanford University CS231n Project Report*, 2015. [Online]. Available: https://cs231n.stanford.edu/reports/2015/pdfs/ConvChess.pdf
[5] K. Omori and P. Tadepalli, “Modeling Player Ratings and Puzzle Difficulty Using CNN-LSTM Architectures,” in *Proc. AAAI Conference on Artificial Intelligence*, 2021, pp. 5341–5348.
[6] O. E. David, N. S. Netanyahu, and L. Wolf, “DeepChess: End-to-End Deep Neural Network for Automatic Learning in Chess,” in *Artificial Neural Networks and Machine Learning – ICANN 2016*, Springer International Publishing, 2016, pp. 88–96. doi: 10.1007/978-3-319-44781-0_11.
[7] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8423-8429, doi: 10.1109/BigData62323.2024.10825289.
[8] S. Doe, J. Smith, and K. Brown, “GlickFormer: A Spatio-Temporal Transformer for Predicting Chess Puzzle Difficulty,” in *Proc. IEEE International Conference on Big Data*, 2024, pp. 1234–1243.
[9] T. Woodruff, O. Filatov, and M. Cognetta, “The bread emoji team’s submission to the IEEE BigData 2024 Cup: Predicting chess puzzle difficulty challenge,” in 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 8415–8422, doi: 10.1109/BigData62323.2024.10826037.
[10] D. Ruta, M. Liu and L. Cen, "Moves Based Prediction of Chess Puzzle Difficulty with Convolutional Neural Networks," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 8390-8395, doi: 10.1109/BigData62323.2024.10825595.
[11] M. Liu, L. Cen and D. Ruta. Exploring Stability and Performance of hybrid Gradient Boosting Classification and Regression Models in Sectors Stock Trend Prediction: A Tale of Preliminary Success and Final Challenge. *19th Conf. Comp. Sci. and Intel. Sys. (FedCSIS)*, Serbia, 2024.
[12] M. Liu, L. Cen and D. Ruta, "Gradient Boosting Models for Cybersecurity Threat Detection with Aggregated Time Series Features," *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Warsaw, Poland, 2023, pp. 1311-1315, doi: 10.15439/2023F4457.
[13] D. Ruta, M. Liu and L. Cen, "Beating Gradient Boosting: Target-Guided Binning for Massively Scalable Classification in Real-Time," *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Warsaw, Poland, 2023, pp. 1301-1306, doi: 10.15439/2023F7166.
[14] D. Ruta, M. Liu, L. Cen. Feature Engineering for Predicting Frags in Tactical Games. *Proc. Int. Conf. 2023 IEEE International Conference on Multimedia and Expo*, 2023.
[15] D. Ruta, M. Liu, L. Cen and Q. Hieu Vu. Diversified gradient boosting ensembles for prediction of the cost of forwarding contracts. *Proc. Int. 17th Conf. on Computer Science and Intelligence Systems*, 2022.
[16] Q. Hieu Vu, L. Cen, D. Ruta and M. Liu. Key Factors to Consider when Predicting the Costs of Forwarding Contracts. *Proc. Int. Conf. 2022 17th Conf. on Computer Science and Intelligence Systems*, 2022.
[17] D. Ruta, L. Cen, M. Liu and Q. Hieu Vu. Automated feature engineering for prediction of victories in online computer games. *Proc. Int. Conf on Big Data*, 2021.
[18] Q. Hieu Vu, D. Ruta, L. Cen and M. Liu. A combination of general and specific models to predict victories in video games. *Proc. Int. Conf. on Big Data*, 2021.
[19] D. Ruta, L. Cen and Q. Hieu Vu. Deep Bi-Directional LSTM Networks for Device Workload Forecasting. *Proc. 15th Int. Conf. Comp. Science and Inf. Sys.*, 2020.
[20] L. Cen, D. Ruta and Q. Hieu Vu. Efficient Support Vector Regression with Reduced Training Data. *Proc. Fed. Conf. on Comp. Science and Inf. Sys.*, 2019.
[21] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, “FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests,” in *Proc. 20th Conf. Comput. Sci. Intell. Syst. (FedCSIS)*, vol. 43, Polish Inf. Process. Soc., 2025. doi: http://dx.doi.org/10.15439/2025F5937.
[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased Boosting with Categorical Features,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6638–6648.
[23] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 831–835 ISSN 2300-5963 ACSIS, Vol. 43
DOI: 10.15439/2025F7603

Pretraining Transformers for Chess Puzzle Difficulty Prediction

Szymon Miłosz
0009-0007-7408-7304
Institute of Applied Computer Science
Lodz University of Technology
Lodz, Poland
szymonmilosz99@gmail.com

***Abstract*—This paper presents our third-place solution for the FedCSIS 2025 Challenge: Predicting Chess Puzzle Difficulty — Second Edition. Building on our prior GlickFormer architecture, we develop a transformer-based approach featuring a novel multitask pretraining strategy that combines masked-square reconstruction with solution policy prediction. Our spatial-only architecture directly embeds solution moves, eliminating temporal modules, while integrating human-centric priors through Maia-2 engine solve-rate predictions. Evaluated on the Lichess puzzle corpus, our approach reduces validation MSE by 30.4% compared to from-scratch training and achieves competitive results (test MSE: 55.9k) despite distribution shifts in the competition environment. In an auxiliary uncertainty-masking side competition organized post hoc, our simple calibration-sensitivity mask achieved the best score and won the side competition. Details are provided in the Appendix.**

***Index Terms*—Chess AI, Transformer Pretraining, Puzzle Difficulty, Self-Supervised Learning, Glicko-2**

I. Introduction

CHESS puzzles challenge players to identify tactical sequences that yield decisive gains. On platforms like Lichess, puzzle difficulty is quantified via Glicko-2 ratings derived from human solve rates. The requirement for thousands of human attempts makes rapid rating of new puzzles impractical, motivating automated difficulty prediction methods that model human problem-solving rather than engine-optimal play.

Building on our prior GlickFormer (11th place in 2024) [1], we introduce GlickFormer v2, which placed third in the FedCSIS 2025 Challenge: Predicting Chess Puzzle Difficulty — Second Edition. [2] While retaining a transformer backbone, this new version simplifies computation by eliminating explicit temporal modules in favor of embedding future solution moves directly in the input representation. Additionally, we incorporate human-centric priors through solve-rate predictions from Maia chess engines [3] across various Elo bands.

Our core contribution is a novel multitask pretraining strategy inspired by large language models (LLMs), which simultaneously trains the model to reconstruct randomly masked board squares while predicting the puzzle's correct solution move. When fine-tuned on the 4.7 million-position training set, this pretraining significantly outperforms both from-scratch training and masked-only pretraining baselines. To address the distribution shift between validation and test ratings, we implement post-hoc distribution scaling technique used by Woodruff et al. [4].

The resulting model achieves competitive performance despite its simplified architecture, demonstrating that the combination of efficient spatial processing, human-aligned inputs, and targeted pretraining can effectively capture human puzzle-solving complexity without expensive temporal modeling.

II. Related Work

Early approaches to chess puzzle difficulty estimation relied primarily on human-derived ratings, but recent competitions have established new state-of-the-art benchmarks. The 2024 IEEE BigData Cup [5] saw significant advances: Björkqvist et al. [6] combined handcrafted features with engine-extracted signals from Maia, Leela, and Stockfish using a residual neural network and LightGBM to place third. Schütt et al. [7] developed a human problem-solving inspired CNN with auxiliary theme and move-prediction tasks, securing second place. Rafaralahy [8] applied a pairwise learning-to-rank framework using RankNet to simulate puzzle "matches" for Glicko-2 rating inference, achieving fourth place.

The current state-of-the-art was established by the bread emoji team [4], winners of the 2024 competition. Their solution employed an ensemble of pretrained Maia and Leela embedders with distribution rescaling postprocessing, achieving over 13% reduction in MSE compared to runner-up solutions. Our original GlickFormer [1] leveraged a factorized spatio-temporal transformer to model board states and move sequences, placing 11th in 2024 while demonstrating attention mechanisms' potential for capturing human-like puzzle difficulty, though its computational costs were substantial.

Our multitask pretraining approach draws inspiration from LLM strategies. The masked reconstruction component adapts BERT's masked language modeling objective [9] to chess by predicting masked board squares based on spatial context. This technique shares strong parallels with Vision Transformers (ViT) [10], which successfully adapted BERT-style pretraining to image recognition tasks by treating image patches as tokens. The policy prediction task shares similarities with next-token prediction in autoregressive models like GPT [11], but focuses

specifically on identifying optimal tactical moves. This dual-objective approach mirrors multi-task pretraining in models like T5 [12], creating robust representations through diverse objectives. Recent chess-specific transformers like the Chess Transformer [13] have demonstrated the value of language-modeling-inspired approaches for move generation, though not specifically for difficulty prediction.

Building on these foundations, GlickFormer v2 integrates efficient spatial-only transformers with LLM-inspired multitask pretraining and competition-proven calibration techniques. Our approach adapts established language modeling paradigms to structured prediction tasks in chess while building upon the state-of-the-art results from the 2024 competition.

III. Methodology

Building upon our previous GlickFormer architecture, we maintain the core data pipeline and noise-aware target sampling strategy while introducing three key enhancements: improved input representations, a leaner spatial-only transformer backbone, a novel multitask pretraining schedule, and the integration of post-hoc distribution calibration.

A. Data and Stochastic Target Sampling

The training corpus comprises 4.7 million puzzles with Glicko-2 ratings r_i, rating deviations RD_i, complete solution move sequences, and solve-rate predictions from the Maia-2 chess engine [3]. These predictions are provided for multiple Elo bands (approximately 1000-2100, though exact ranges were unspecified by competition organizers), including both rapid and blitz time control configurations. The solution moves represent the optimal tactical sequence required to solve each puzzle, typically ranging from 1-5 moves in length.

To evaluate model performance during development, we reserve 1% of puzzles (approximately 47,000) for validation, selected via stratified sampling across puzzle themes to ensure representativeness.

For pretraining, we construct a 14-channel input representation where 13 binary planes indicate piece occupancies: each plane corresponds to a specific piece type-color combination or empty square, with values set to 1 where present and 0 otherwise. This is supplemented with a binary mask-indicator plane that marks randomly selected squares for masking (1 = masked, 0 = visible), with all other channels zeroed out at masked positions.

During fine-tuning, we use a richer 55-channel input with three components: the same 13 binary board-state planes; 20 binary planes encoding the complete solution line (truncated or zero-padded to exactly 5 moves, each move represented as side $\times$ {from,to} pairs); and 22 continuous-valued Maia-2 solve-rate priors broadcast as constant spatial planes. All channels are linearly projected to form 64 tokens (one per board square) for transformer processing.

For stochastic target sampling, we normalize ratings using dataset statistics:

$$\mu_i = \frac{r_i - \mu_{\text{dataset}}}{\sigma_{\text{dataset}}} \tag{1}$$

$$\sigma_i = \frac{\mathrm{RD}_i}{\sigma_{\text{dataset}}} \tag{2}$$

where μ_{dataset} and σ_{dataset} represent the mean and standard deviation of all puzzle ratings in the training set. We model each rating as a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ and generate noisy targets during mini-batch sampling:

$$y_i \sim \mathrm{clip}\left(\mathcal{N}(\mu_i, \sigma_i^2), \mu_i - 3\sigma_i, \mu_i + 3\sigma_i\right) \tag{3}$$

This injects label noise that regularizes the regression objective while respecting rating deviation bounds.

B. Model Architecture

GlickFormer v2 employs 12 encoder-only transformer blocks with 768-dimensional embeddings and 24 attention heads with Smolgen positional encoding, using Mish activation functions. [14] The model processes different inputs during pretraining and fine-tuning phases: pretraining uses only the 14-channel tensor, while fine-tuning incorporates the extended 55-channel tensor where both the solution-move projector W_{sol} and Maia-prior projector W_{maia} are randomly initialized and trained alongside the backbone.

Each input stack undergoes projection via linear embedding into 64 tokens (one per chess square), which are then processed by the transformer backbone. During pretraining, two specialized heads operate simultaneously: a masked-square classifier implemented as an MLP applied independently to each token predicts piece types at masked positions, while a policy head employs two parallel MLPs to project each token to "from" (query) and "to" (key) vectors. Move logits are computed as dot products between all query-key pairs ($\mathbf{q}_a^\top \mathbf{k}_b$ for move $a \to b$), implementing the Leela Chess Zero policy formulation [15]. This formulation considers only square-to-square transitions without special move types, inherently omitting promotion moves since puzzles rarely involve pawn underpromotions in their solution lines.

For policy prediction during pretraining, we consider only the first move of each puzzle's solution sequence. This design choice reflects that the initial move typically represents the most significant tactical insight distinguishing puzzle difficulty, while subsequent moves are often forced responses or technical conversions contributing less to human solve-rate variance. This simplification maintains predictive power while reducing computational complexity.

For fine-tuning, both pretraining heads are discarded and replaced with the value head architecture from the original GlickFormer. This consists of a linear layer that projects the entire sequence of 64 token embeddings to a hidden dimension, followed by Mish activation and a final linear layer that outputs a scalar difficulty prediction. The entire model - including the transformer backbone, input projectors, and regression head - is trained end-to-end during this phase,

allowing the pretrained representations to adapt while specializing for rating prediction.

C. Multitask Pretraining

Our pretraining approach masks k random squares per puzzle, where $\mathcal{M}$ denotes the set of masked squares. We define two complementary loss functions:

The masked reconstruction loss predicts missing piece types:

$$\mathcal{L}_{\text{msk}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_{\text{msk}}(c_i) \tag{4}$$

where c_i is the true piece class at square i.

The policy loss identifies the correct solution move:

$$\mathcal{L}_{\text{pol}} = -\log p_{\text{pol}}(a^\star) \tag{5}$$

where $a^\star$ is the ground-truth solution move.

The combined pretraining objective balances these losses:

$$\mathcal{L}_{\text{pre}} = \lambda_{\text{msk}} \mathcal{L}_{\text{msk}} + \lambda_{\text{pol}} \mathcal{L}_{\text{pol}} \tag{6}$$

Figure 1 illustrates our dual pretraining objectives. The model receives boards with randomly masked squares (Fig. 1a), then must both reconstruct the original pieces at masked positions (Fig. 1b) and predict the optimal solution move indicated by the arrow (Fig. 1c).

Following the Chinchilla scaling rule [16], we train with $20\times$ more tokens than parameters using AdamW optimization with batch size 256, dropout rate 0.1, weight decay 0.1, initial learning rate 5×10^{-4}, 10% linear warmup, and cosine decay to final learning rate 5×10^{-5}. We empirically set $k = 8$ and weighting coefficients $\lambda_{\text{msk}} = 1$, $\lambda_{\text{pol}} = 0.1$ based on validation performance.

D. Fine-Tuning Procedure

During supervised fine-tuning, we employ the 55-channel input tensor with newly initialized projectors for solution-move and Maia-prior planes. The entire model trains end-to-end without freezing components. We discard pretraining heads and replace them with the regression head described in Model Architecture section, minimizing the mean squared error:

$$\mathcal{L}_{\text{ft}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{r}_i - r_i)^2 \tag{7}$$

using AdamW optimization [17] with dropout rate 0.1, weight decay 0.1, batch size 256, and initial learning rate 5×10^{-5}. We reduce the learning rate by a factor of 0.1 after each epoch without validation improvement, with a maximum of two reductions. This configuration maintains consistency with our original GlickFormer implementation while leveraging the enhanced pretrained representations.

E. Post-Hoc Distribution Calibration

To address distribution shifts between validation and test environments, we implement the distribution rescaling technique introduced by Woodruff et al. This calibration corrects systematic biases in predicted ratings caused by the key difference in data collection: limited solving attempts (25-50 per puzzle) compared to the training data, which shifts the rating distribution toward the mean μ by preventing full convergence.

The calibration function is defined as:

$$R(\hat{r}) = \begin{cases} \max\left(\mu, \hat{r} - \alpha \min\left(1, \left(\frac{\hat{r}-\mu}{\delta}\right)^\gamma\right)\right) & \hat{r} \geq \mu \\ \min\left(\mu, \hat{r} + \beta \min\left(1, \left(\frac{\mu-\hat{r}}{\delta}\right)^\gamma\right)\right) & \hat{r} < \mu \end{cases} \tag{8}$$

with empirically tuned coefficients $\alpha = 200$, $\beta = 400$, $\gamma = 2$, $\delta = 550$. This formulation applies progressively stronger corrections to predictions farther from μ, counteracting the observed shift toward the mean caused by limited solving attempts.

IV. Results and Discussion

Our experimental evaluation addresses the effectiveness of our pretraining strategies, GlickFormer v2's performance relative to state-of-the-art baselines, and the impact of post-hoc calibration. All validation results use our stratified validation set while test results report official competition metrics.

A. Evaluation Protocol

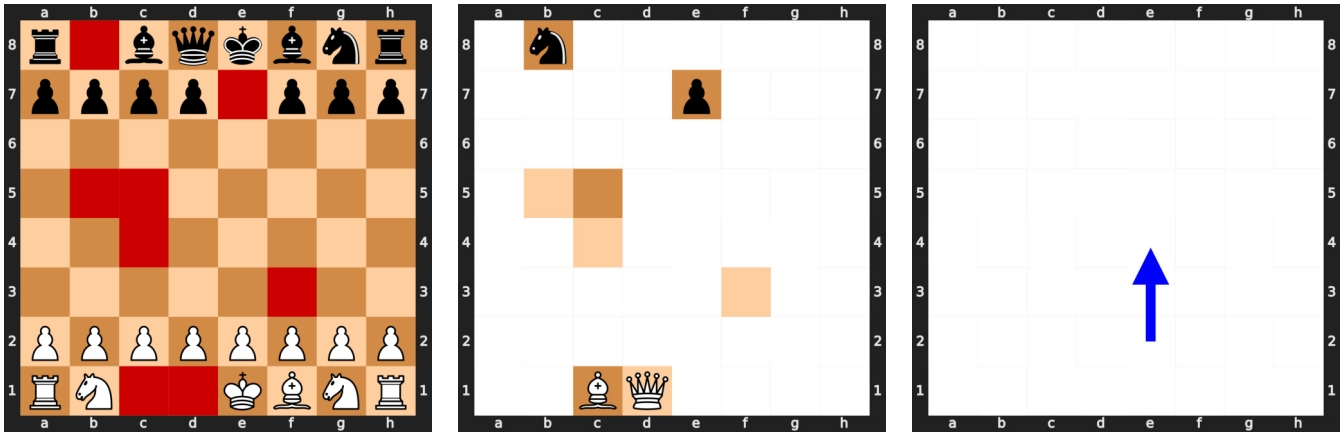
We compare three pretraining strategies: from-scratch training (baseline), masked-only pretraining ($\lambda_{\text{pol}} = 0$), and our multitask approach ($\lambda_{\text{pol}} = 0.1$). Baseline comparisons include GlickFormer v1 [1] and the bread emoji 2024 solution [4]. Note that both our approach and bread emoji used similar validation strategies with sampling from training data. All models are evaluated on our validation set unless otherwise specified.

B. Pretraining Effectiveness

Table I demonstrates the progressive improvements from our pretraining strategy. Masked-only pretraining reduces validation MSE by 22.3% compared to from-scratch training (44.9k vs. 57.8k). Our multitask approach further improves performance by 10.4% over masked-only pretraining (40.2k vs. 44.9k), confirming that combining reconstruction with policy prediction provides richer representations for difficulty assessment.

TABLE I: Pretraining ablation on validation set

Pretraining Strategy	Validation MSE	Difference
No pretraining (from scratch)	*57.8k*	-
Masked squares only	*44.9k*	-22.3%
Multitask	**40.2k**	**-10.4%**



(a) Masked input board (b) Reconstruction target (c) Policy prediction target

Fig. 1: Multitask pretraining objectives: (a) Input board with randomly masked squares (red); (b) Target state for masked-square reconstruction; (c) Target solution move (arrow) for policy prediction.

C. Validation Set Comparisons

Table II compares GlickFormer v2 against key baselines on our validation set. Our approach reduces MSE by 50.2% compared to GlickFormer v1 and by 13.9% compared to Woodruff et al. [4], demonstrating significant improvements over state-of-the-art approaches.

TABLE II: Validation set comparison with baselines

Method	Validation MSE	Difference
GlickFormer v1	80.8k	-
Woodruff et al. (SotA)	46.7k	-
GlickFormer v2 (ours)	**40.2k**	**-50.2% vs v1** **-13.9% vs SotA**

D. Competition Results and Calibration Impact

GlickFormer v2 achieved third place in the 2025 competition with a test MSE of 55.9k. The top-performing teams secured first and second place with MSEs of 52.3k and 54.4k respectively. To quantify the impact of post-hoc calibration, we evaluated its effect on a preliminary test dataset (subset of official test data). This analysis revealed calibration was critical for performance - without it, our MSE increased by 38.6% to 81.6k, demonstrating its importance for addressing distribution shifts in the full competition environment. The close margin between top 5 solutions (52.3k - 61.1k) highlights the competitiveness of this year's challenge, with our approach placing solidly within this range.

E. Key Findings

Analysis reveals several important insights. Masked-only pretraining provides significant gains over from-scratch training, reducing validation MSE by 22.3%. Our multitask approach yields a further 10.4% improvement over masked-only pretraining, confirming the value of combining reconstruction with policy prediction. Post-hoc calibration proves essential for test performance, preventing a 38.6% MSE degradation. Overall, our solution achieves a 50.2% improvement over our previous entry [1] while reducing training time by >80% through architectural simplifications.

The integration of Maia solve-rate priors contributed substantially to our 50.2% improvement over GlickFormer v1, highlighting the value of human-aligned signals in puzzle difficulty modeling. This demonstrates that encoding domain-specific human behavior patterns through pretrained engines provides crucial complementary information beyond board-state analysis alone.

V. Conclusion

GlickFormer v2 demonstrates that spatial-only transformers with targeted pretraining can effectively predict chess puzzle difficulty. The model's novel multitask pretraining strategy, combining masked-square reconstruction with solution policy prediction, significantly outperforms both training-from-scratch and masked-only pretraining baselines. By incorporating human-centric priors through Maia solve-rate predictions and addressing distribution shifts via post-hoc calibration, the approach maintains effectiveness across varied evaluation environments.

Our solution achieved third place in the FedCSIS 2025 Challenge: Predicting Chess Puzzle Difficulty —- Second Edition, showcasing the potential of efficient architectures combined with strategic pretraining.

For future work, this pretraining framework could be adapted for other chess-specific tasks such as value and policy prediction in chess engines. Additional research directions include enhancing robustness to distribution shifts through domain adaptation techniques and refining temporal modeling approaches for complex multi-move puzzles.

References

[1] S. Miłosz and P. Kapusta, "Predicting chess puzzle difficulty with transformers," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10825919 pp. 8377–8384.

[2] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. doi: 10.15439/2025F5937. [Online]. Available: http://dx.doi.org/10.15439/2025F5937

[3] Z. Tang, D. Jiao, R. McIlroy-Young, J. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A unified model for human-ai alignment in chess," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.48550/arXiv.2409.20553 Accepted @ NeurIPS 2024.

[4] T. Woodruff, O. Filatov, and M. Cognetta, "The Bread Emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10826037 pp. 8415–8422.

[5] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10825289 pp. 8423–8429.

[6] S. Bjorkqvist, "Estimating the puzzlingness of chess puzzles," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10825991 pp. 8370–8376.

[7] A. Schutt, T. Huber, and E. Andre, "Estimating chess puzzle difficulty without past game records using a human problem-solving inspired neural network architecture," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10826087 pp. 8396–8402.

[8] A. Rafaralahy, "Pairwise learning to rank for chess puzzle difficulty prediction," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024. doi: 10.1109/BigData62323.2024.10825356 pp. 8385–8389.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. doi: 10.48550/arXiv.2010.11929. [Online]. Available: https://arxiv.org/abs/2010.11929

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[13] D. Noever, "Chess transformer: Mastering the game of chess with attention," *arXiv preprint arXiv:2008.04057*, 2020.

[14] D. Misra, "Mish: A self regularized non-monotonic activation function," *arXiv preprint arXiv:1908.08681*, 2019.

[15] D. Monroe and P. Chalmers, "Mastering chess with a transformer model," *arXiv preprint arXiv:2409.12272*, 2024, describes Lc0's transformer architecture and smolgen position encoding. [Online]. Available: https://arxiv.org/html/2409.12272v1

[16] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[17] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *arXiv preprint arXiv:1711.05101*, 2017.

APPENDIX: UNCERTAINTY-MASKING SIDE COMPETITION

After the main leaderboard closed, the organizers announced an additional task focused on uncertainty estimation. Each team was asked to submit a binary mask over the test set marking exactly 10% of instances deemed most problematic. During scoring, the organizers replaced model predictions with the ground-truth values on the masked subset and recomputed the test metric. Let $[P]$ denote the *perfect-mask* score for a given team's final submission (i.e., the best possible score if the team had masked the true worst 10%), and let $[N]$ denote the score obtained using the team's submitted mask. The side competition ranked teams by the ratio $[N]/[P]$ (lower is better).

Evaluation Setting

Per the organizers' rules, the submitted 10% mask was evaluated on *all three of our final submissions*, using the same mask for each. In our case, these were an uncalibrated prediction run, a prediction run calibrated using the post-hoc rescaling parameters of Woodruff et al., and a prediction run calibrated using the post-hoc rescaling parameters used in this paper. The organizers computed the score across these three final submissions according to the side-competition protocol.

Method

Our approach leverages the post-hoc distribution calibration $R(\cdot)$ (Eq. (8)) to construct a simple, model-agnostic uncertainty proxy. For each test puzzle i, we compute a *calibration sensitivity*

$$\delta_i \;=\; \left|\hat{r}_i^{\mathrm{raw}} - R(\hat{r}_i^{\mathrm{raw}})\right|, \tag{9}$$

where $\hat{r}_i^{\mathrm{raw}}$ is the uncalibrated prediction and $R(\hat{r}_i^{\mathrm{raw}})$ is the calibrated one. Intuitively, δ_i measures how much calibration "needs to move" a prediction to account for the rating-distribution shift (caused by limited solving attempts in the test environment). Large δ_i values occur predominantly at the distribution tails (very easy/very hard puzzles), which are known to exhibit higher instability. We rank puzzles by δ_i and mark the top 10% as the uncertainty mask. This design purposely targets cases where calibration most alters the prediction, and was explicitly intended to reduce error on the uncalibrated submission.

Outcome

This calibration-sensitivity mask achieved the lowest $[N]/[P]$ ratio among all participating teams, winning the side competition. Our ratio was:

$$\frac{[N]}{[P]} = 1.426,$$

which gave us **1st place** out of 9 teams that decided to participate in this additional task. For our final submitted mask:

$$[N] \approx 61.49k, \quad [P] \approx 43.12k.$$

Beyond its simplicity and robustness (no additional training or ensembling), the method aligns tightly with our main-system calibration: it explicitly targets those instances where calibration exerts the largest corrective effect, which empirically coincide with the most uncertain test cases under the competition's setting.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 837–842
DOI: 10.15439/2025F6771
ISSN 2300-5963 ACSIS, Vol. 43

The `bread emoji` Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge

Tyler Woodruff*
Amazon Project Kuiper
hhhtylerw@protonmail.com

Luke Imbing
Independent Researcher
lukeimbing@gmail.com

Marco Cognetta
Google + Institute of Science Tokyo
cognetta.marco@gmail.com

***Abstract*—We detail the `bread emoji` team's submission to the FedCSIS 2025 Predicting Chess Puzzle Difficulty Challenge. Our solution revolved around improving our submission from the previous competition by incorporating a new puzzle metadata feature and optimizing our implementation to allow for larger model ensembles and more stable training. Similar to our submission from last year, our system has two stages: learning a strong predictor for the LICHESS dataset and then rescaling the distribution using an empirically-guided post-processing step to fit it to the smaller and noisier competition dataset.**

Our submission placed second with a $\sim$3.9% gap in mean squared error (MSE) from first place [1].[1]

I. Introduction

THIS paper introduces our team's (`bread emoji`) submission to Predicting Chess Puzzle Difficulty - Second Edition, a competition hosted by FedCSIS. This was an open-to-the-public online machine learning competition that challenged teams to come up with novel solutions for predicting the difficulty (represented by the GLICKO2 rating system [2]) of a chess puzzle, given the initial board state, the winning moves, and a small set of metadata for each puzzle.

We participated in last year's iteration of this competition, placing first by a significant margin [3][4]. Our solution to this year's competition iterated on last year's solution, and we cover our updated methodologies in this paper. Overall, our solution placed second out of 71 teams, narrowly missing out on first place by $2k$ MSE ($<4\%$).

II. Competition Background

A chess puzzle is an initial board state and sequence of moves that lead to one side of the board gaining a significant material or strategic advantage over the other in a chess game. Puzzles are created primarily by algorithms that parse actual human vs. human games on online chess websites and find interesting move sequences. They are then added to a game server that serves players a continuous stream of puzzles to play against. Both the puzzle and the player have a rating, which are both updated by player wins and losses, according to the GLICKO2 rating system.

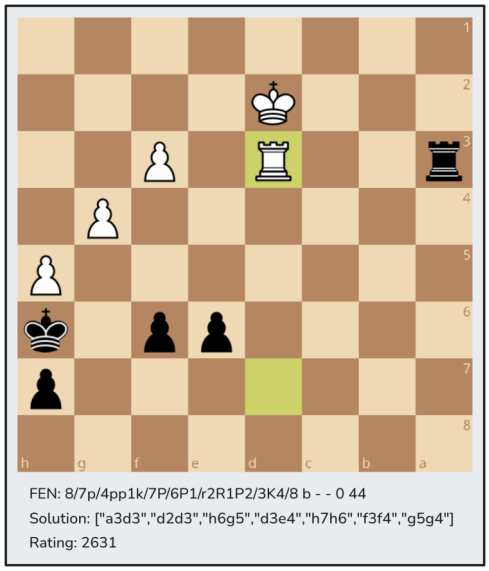


Fig. 1: An example puzzle from the LICHESS database. The most basic information required is the initial position (given in `FEN` format), the solution (from which we can generate a sequence of boards), and the puzzle rating (which we are trying to predict). However, the LICHESS and competition datasets provide some additional puzzle metadata. The theme of this puzzle is to convert white's blunder (capturing the pawn on `d3`, which starts the puzzle) into a winning pawn endgame via zugzwang. The algebraic solution is `♖xd3?? ♜d3+ ♔xd3 ♚g5 ♔e4 ♟h6! ♙f4+ ♚xg4`, after which black has an overwhelming pawn advantage.

Figure 1 gives an example from the LICHESS database.[2]

A. LICHESS *Dataset*

The training dataset given by the competition is a list of over 4.5 million puzzles taken from LICHESS. This is a snapshot of the free-to-download, updated-monthly puzzle dataset taken several months prior to the start of the competition. This dataset contains per-puzzle metadata that is the same as the

*This work is not related to the author's work at Amazon.

[1]Our code can be found here: https://github.com/mcognetta/ieee-chess. We have moved the implementation from the previous iteration of the contest to the `2024-iteration-archive` branch.

[2]https://database.lichess.org/#puzzles

prior iteration of this competition (we point the reader to [3], [4]) with the addition of a set of "success probabilities" computed by the MAIA2 neural chess engine.

B. MAIA2 *Success Probabilities*

This year's competition introduces new metadata in the form of success probabilities generated by MAIA2, a neural chess engine designed to mimic human play (Section III-B). Each value corresponds to the estimated likelihood that a player of a given skill level would successfully complete the puzzle according to the move distributions predicted by 22 different variations of the MAIA2 model [1]. These probabilities were included as scalar features, indexed by their model configurations and were provided for both the LICHESS and competition datasets by the organizers.

III. NEURAL CHESS MODELS

Here, we briefly introduce the families of neural chess models that we use as the backbone for our system. These models differ from classical chess models in that they are trained entirely on data without any hand-crafted features. In particular, we utilize MAIA and MAIA2, which are a family of neural models that are designed to mimic human players of various strengths, and LEELA, an open-source reimplementation of the ALPHAZERO neural chess system. This is not an exhaustive set of neural chess engines, and we point the interested reader to [5] for an overview.

A. MAIA

MAIA is a set of neural chess models that are trained to mimic players of a given rating (for ratings 1100 to 1900 in increments of 100) [6]. Each model has the same architecture—a chain of 6 convolutional blocks with residual connections, followed by a value and policy head—that takes a board representation as input and produces a policy and value score which encode the probability distribution over next moves and the evaluation of the position, respectively.

B. MAIA2

MAIA2 builds upon the ideas of MAIA, but with two key differences. First, the models are no longer segmented by rating; they now take the player ratings as input. And second, the model includes a new rating-board attention layer, which injects rating information into the board representation [7]. The board input and output representations are also different than MAIA, but the training objective is the same.

MAIA2 comes in two flavors, RAPID and BLITZ, denoting the training data used to train it, and the default parameters chunk player ratings into 11 buckets (1050 to 2050).

C. LEELA

LEELA is an open source implementation of the ALPHAZERO neural chess engine [8], [9]. The most modern iteration of LEELA uses a transformer backbone, but we opted to continue using the original model (as we did in [3]), which uses a convolutional backbone. LEELA was trained via self-play to be as strong as possible and has SMALL, MEDIUM, and LARGE variants.

IV. OUR SYSTEM

Our general approach mirrors our implementation from the prior competition [3]. In short, we use neural chess board embedders to extract latent representations of the boards in a puzzle. These are fed into an RNN to produce a puzzle representation, which is then combined with puzzle metadata and passed to a regression MLP. Our solution for this challenge differs from our previous solution in a few details, including an updated, optimized implementation and some architectural and modeling differences. Here, we review the main architectural components of our system and in Section VI-A, we discuss the application of this system to the actual competition challenge.

A. Board and Puzzle Embedders

We again use the MAIA and LEELA neural chess engine family as our neural chess embedding backbone. These are large, convolutional models that ingest board representations (i.e., bitboards representing piece positions and castling metadata) and produce a series of policy and value scalars that are used during min-max tree search in a chess engine. We truncate the models by removing the policy and value heads and take the output of the last convolutional block as the latent vector representation of a given board state.

We point the reader to Figure 10 of [6] (in the Supplemental section) which shows the MAIA architecture. Our latent representation is the vector that is produced by the convolutional block directly before the model is split into the two heads (and we use the analogous vector for the LEELA family, which has a similar architecture [8][9]).

However, puzzles are inherently sequential and variable length. Thus, after extracting latent representations of each board-state in a puzzle, we combine them into a puzzle representation by feeding them into an RNN to produce a single, vector representation.

B. Board Metadata

The LICHESS dataset comes with a large amount of additional metadata other than just the puzzle information. For example, it contains themes (e.g. mate-in-two, pin, etc.), popularity, the game it was drawn from, and more. This metadata may be helpful for modeling purposes, so our architecture allows for injecting representations of the metadata into the latent representation of the puzzle. The competition dataset, however, contains only the explicit puzzle information and the set of precomputed MAIA2 puzzle probabilities (Section II-B), so much of the metadata that is available during training is not immediately helpful since we do not have access to the same metadata during inference.

In our final architecture, we use only puzzle length and MAIA2 puzzle probabilities as metadata features. For the former, we produced a categorical representation of puzzle length (e.g., puzzles of length 2, 4, ..., 16) and used an embedding table to extract the learned metadata representation corresponding to the length of the given puzzle. For the latter, which consists of 22 scalar values, we produce a vector

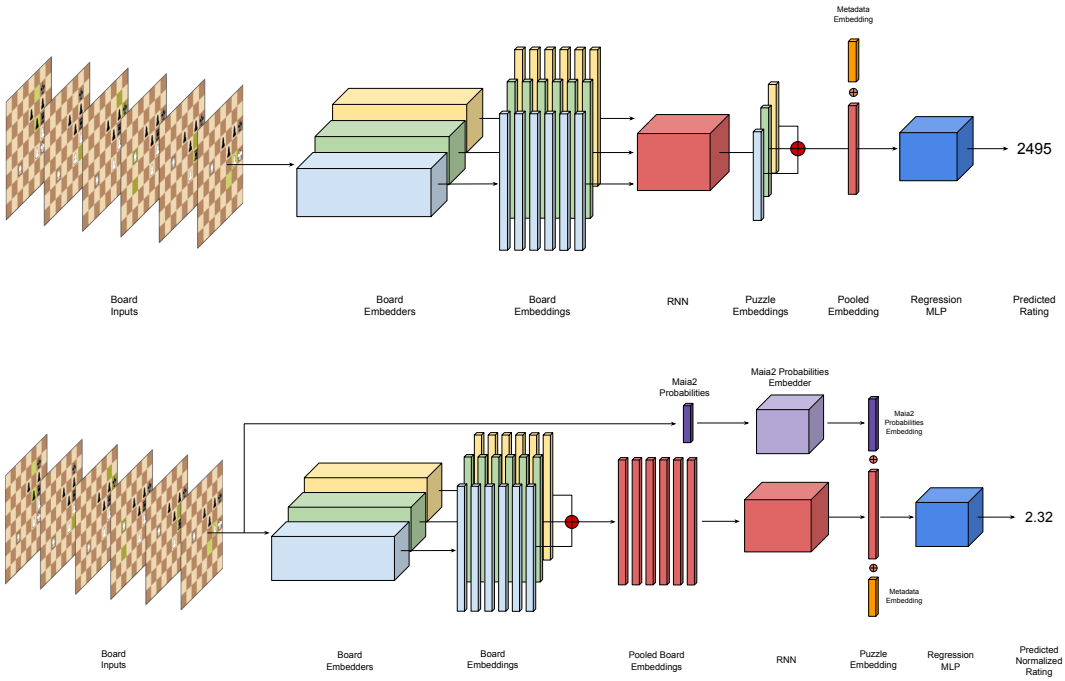


Fig. 2: The model architecture used in the previous competition (top) [3] and the system in this paper (bottom).

embedding by constructing a 22-length vector and passing it through a small MLP to produce a latent representation.

We note one feature, puzzle themes, that is present in the LICHESS dataset but not the competition dataset. However, the LICHESS theme tagging algorithm is open-source[3] and can be used to tag the competition dataset. At least one team in the first iteration of this competition tagged the competition dataset so that the feature could be used during inference [10], however, we did not experiment with including this metadata.

The metadata representations are concatenated to the output of the RNN to produce the final puzzle representation.

C. Rating Prediction as Regression

Once a latent vector representation of the board and its metadata has been formed, we use it to predict the GLICKO2 rating of the puzzle—a simple regression task. We use a three layer, dense MLP with `relu` activations and dropout to learn a mapping from a puzzle representation to its scalar rating.

D. Ensembled Embeddings

Like in [3], we do not restrict ourselves to having a single neural board embedder in a model. Instead, we allow multiple neural board embedders to independently produce board representations which are then combined to form an ensembled board representation. The intuition behind this is that different models have different strengths and weaknesses due to their training objectives and so their latent representations of the board may be complementary and lead to better modeling.

[3]https://github.com/ornicar/lichess-puzzler

Our architecture allows for arbitrarily many board embedders and can mix MAIA and LEELA embeddings.

V. ARCHITECTURAL AND MODELING DIFFERENCES FROM OUR SUBMISSION TO THE PRIOR CONTEST [3]

Despite using the same initial framework as in [3], we implemented a number of architectural and modeling changes (other than the inclusion of the new MAIA2 features) that improved our system's performance. Figure 2 shows the major differences between the two systems.

A. RNN Implementation

In the previous iteration of this competition we manually implemented the RNN in our model architecture. Regardless of puzzle input length, we truncated or padded to 10 boards and then used a hand-written RNN loop. This was slow, had many wasted computations (in that many puzzles have less than 10 boards, but the padded boards were processed anyway), and had some undesirable modeling properties (in theory, models should learn to ignore the padding boards, but they can still "wash out" the signal of non-padding boards, especially when the puzzle is very short).

We replaced our implementation with the PYTORCH RNN implementation, which is faster, uses less memory, and, by using packed tensors, only processes non-padding boards. In a benchmark, we observed a $>4\times$ runtime speed-up when using the PYTORCH RNN compared to our hand-written RNN.

B. Board Pooling

Recall that we have several board embedders that each produce a sequence of latent board representations for each board

(a) Results from the previous competition, taken directly from [3]. These do not include MAIA2 features and use the previous iteration of our architecture. Note that the bottom row is an output-ensemble of embedding-ensemble models.

Model	MSE
MAIA-1300	$56.5k$
MAIA-1500	$57.5k$
MAIA-1700	$57.9k$
LEELA-SMALL	$65.5k$
LEELA-MED	$66.1k$
MAIA-1300 + LEELA-SMALL	$65.4k$
MAIA-1300 + LEELA-MED	$50.8k$
MAIA-{1300, 1500, 1700}	$56.0k$
MAIA-{1300, 1500, 1700} + [MAIA-1300 + LEELA-SMALL] + [MAIA-1300 + LEELA-MED]	$46.7k$

(b) Results for our updated architecture, which includes MAIA2 features. Note that all of these are embedding-ensembled models. The right column is the same model configuration but with an additional LEELA-SMALL embedder.

Model	MSE	
		+ LEELA-SMALL
MAIA-1100	$41.2k$	$41.3k$
MAIA-1500	$41.6k$	$41.3k$
MAIA-1700	$42.2k$	$42.1k$
MAIA-{1100, 1500}	$41.7k$	$38.9k$
MAIA-{1100, 1700}	$40.3k$	$38.7k$
MAIA-{1500, 1700}	$39.7k$	$39.3k$
MAIA-{1100, 1500, 1700}	$39.2k$	**38.3k**

TABLE I: Results on the LICHESS dataset using different model configurations, including from our solution to the previous competition. Each model configuration's MSE is the average of 3 copies of the model. The best performing model is **bolded**.

state in the puzzle and our goal is to somehow combine all of them into a single puzzle representation. In [3], we achieved this by first computing a puzzle vector for each embedder (by processing that embedder's board representations with the RNN) then (mean) pooling the resulting vectors.

In our updated implementation, we reverse this process. We now use the embedders to compute each board state, then (mean) pool all of the analogous board representations before passing it through an RNN.

This has two main benefits. First, it allows for different embedders' board representations to be "mixed" earlier in the model pipeline, which should allow the model to better incorporate the different representations. And second, the board embeddings are processed in parallel in a giant batch, but the RNN inherently involves a non-parallelization, sequential process. Thus, by pooling the embeddings before the RNN layer, we reduce the number of non-parallelizable operations to just a single RNN call, which massively speeds up training and inference runtimes.

Due to the runtime speed up and lower memory requirements of the combination of pre-RNN board pooling and the PYTORCH RNN implementation, we were able to increase our training batch sizes and observed more stable training and faster convergence.

C. Regression Target

Recall that the primary objective of this competition is to predict a puzzle's GLICKO2 rating, which is a non-negative scalar, and try to minimize MSE. In [3], we simply predicted this scalar value directly. Since the ratings can be higher than 3000, the squared errors of predictions can be very large, which leads to training instability due to large gradients.

In our updated version, we instead computed the mean μ and standard deviation σ of the clean LICHESS dataset and scaled the ratings as $r_n = \frac{r-\mu}{\sigma}$ to produce a normalized rating (which was used in the MSE objective). During inference, we reverse the normalization by predicting a normalized rating $\hat{r}_n$ and recovering the predicted rating $\hat{r} = \sigma \times \hat{r}_n + \mu$. This resulted in more stable training and faster convergence.

VI. RESULTS

Overall, our updated model architecture and training scheme, coupled with the addition of MAIA2 probability metadata, led to significantly more accurate models on the LICHESS dataset. Table I shows the results of our prior solution (taken directly from [3]) and a subset of the model configurations we trained for our final submissions.

We note that, across the board, the updated models have substantially lower MSE, despite having similar parameter counts. Most clearly, we can compare models which have the same embedding configurations and thus the salient differences are our new architecture and the MAIA2 metadata (regrettably, we did not perform MAIA2 ablation tests).

For example, comparing the MAIA-1500 model from the previous competition to the MAIA-1500 model from this competition, we observe a $\sim 16k$ reduction in MSE (a 27.7% reduction), despite having similar parameter counts.

Perhaps unsurprisingly, we find that our largest ensemble, MAIA-{1100, 1500, 1700} + LEELA-SMALL, performed the best — achieving $38.3k$ MSE on the LICHESS test set, which is an $8.4k$ (18.0%) reduction in MSE over the best, multi-ensembled model from our submission to the prior competition. Ensembling all of the models that we trained for this competition (i.e., the entire set of models from the right two tables of Table I) reached $36.1k$ MSE corresponding to a $2.2k$ (5.7%) reduction in MSE over the best individual model from this competition and a $10.6k$ (22.7%) reduction over the multi-ensembled model from the last competition.

A. Fitting to the Competition Dataset

Like in [3], we found that models that performed well on the clean LICHESS dataset did not necessarily perform well on the competition dataset, as there are significant differences between the competition distribution and the LICHESS dataset distribution (for example, see Figure 1 of [4]).

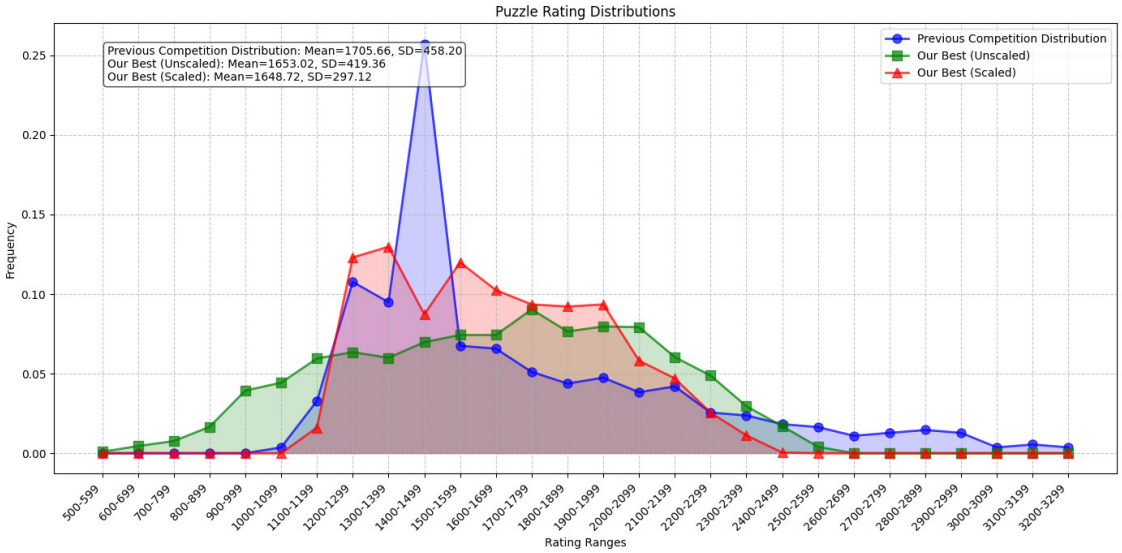


Fig. 3: An overlay of the competition test set's distribution (blue) as reported in the prior competition's writeup [4], our best model's predictions before scaling (green), and our best, scaled result (red). Notice that they have significantly different shapes, implying that accurately modeling one is not a guarantee of accurately modeling another.

We again used the strategy of producing as good of a model as possible on the LICHESS dataset and then rescaling it to fit what we expected the competition dataset to look like.

Previously, we used a simulator that estimated, given a set of players and puzzles for which we had gold-standard rating information, what the distribution of puzzle ratings would look like after each puzzle had been played by a subset of N random players. This was designed to mimic the competition server, which matched players with puzzles randomly (i.e. not matching players and puzzles of similar skill level) and, at the time, did not have many plays-per-puzzle. We used these findings to guide our search for a rescaling function that could map our predicted ratings to what we would expect that puzzle to be rated in the low-play competition-server regime.

We followed the same procedure in this iteration, by using rescaling functions that attempted to reshape our distribution of predicted competition puzzle ratings to be more centered around what we believed the mean competition rating to be.

First, as a baseline, we noted that this year's competition dataset was a subset of the previous year's competition dataset. Thus, we simply reused our winning solution from [3] after filtering out the removed puzzles. On the preliminary competition dataset from last year, our winning solution had $49.1k$ MSE. However, the same solution scored $64.1k$ MSE, indicating that the current iteration of the competition dataset is significantly different than the prior version.

We first performed binary search over the set of possible ratings (between 0 and 3000) for the value that minimized MSE on the preliminary competition dataset when we used it as a constant prediction (i.e., we predicted the same value for all puzzles). The value that minimizes this function gives the mean of the dataset (and we can also recover the standard deviation), approximately $\mu \approx 1650$.

We then used post-processing rescaling steps to try to reshape our predicted output distribution to a distribution that had similar properties to what we expected the competition dataset to have and settled on rescaling functions of the form:

$$\text{SCALE}(r) = \begin{cases} o_u + \frac{\log(c_1 r)}{c_2} & r > 0 \\ o_l - \frac{\log(c_1' r)}{c_2'} & r \leq 0 \end{cases}$$

where r is the normalized predicted rating $\frac{\hat{r}-\mu}{\sigma}$ and o_u, c_1, and c_2 are rescaling hyperparameters that we searched over for when our predicted rating was above the mean (and likewise for o_l, c_1', and c_2' when it is below the mean). We separated these so that we could control the shape of both sides of the output distribution, as we found it was not symmetric. The normalized predicted rating is mapped back into the raw rating space after rescaling.

The differences between the prior competition dataset, our model's (unscaled) predictions on the competition dataset, and our best solution's scaled predictions is shown in Figure 3.

VII. THINGS WE TRIED THAT DID NOT WORK

Like in [3], once our core pipeline was finished, we experimented with several variants but found that none provided gains over our baseline.

A. *RNN → Transformer*

One goal of our updated implementation was to improve the model's runtime performance in order to improve training speed and performance by increasing the batch size and parameter count. A major bottleneck is the RNN layer (particularly in the older, hand-written implementation that pooled boards after the RNN), since it is inherently sequential and cannot

be parallelized. We experimented with replacing this layer with a small transformer model, which has the benefit of increased parallelization and the ability to explicitly inspect the representations of every board in a puzzle simultaneously and learn a weighting for their interactions. We hoped that this would improve both the model's runtime and quality, but we instead found that it massively reduced the modeling performance, even at similar parameter counts.

We experimented with several variations, including stacking several layers of transformers and adjusting the transformer's internal hidden dimension (both within a fixed parameter budget), but we were not able to match the RNN baseline.

B. MAIA2 *Board Embeddings*

We experimented with adding MAIA2 as a board embedder by using the output of the board-rating attention layer as the latent board representation. As the MAIA2 model requires player ratings as input, but the competition dataset does not provide this metadata (the LICHESS dataset does include it), we experimented with several different schemes to select a reasonable rating. For example, we performed a hyperparameter search over all pairs (p_1, p_2) where p_1, p_2 were ratings in the MAIA2 rating buckets to see if some pair worked well. We additionally tried using the ratings as a form of self-ensembling, by simply selecting several rating pairs as hyperparameters and then computing the latent representation of the board for each pair and averaging them.

None of these schemes were effective and we found that models using MAIA2 embeddings were substantially worse than those using MAIA. This is potentially due to the comparatively small output dimension of the rating-board attention layer of MAIA2 (1024) compared to the convolutional block layers of MAIA (4096), which means that it carries less signal when used in the downstream task.

Since MAIA2 also uses a sequence of convolutional blocks as a backbone before the attention layer, we could have used the output of that as well, but we did not experiment with this.

C. Alternative Loss Functions

Our baseline model uses MSE loss as the primary loss function in order to mimic the competition objective. However, as noted in [3], our best performing models had relatively high MSE (40k+ and 50k+ on the LICHESS dataset and the preliminary competition dataset, respectively). Thus, we hypothesized that "small" errors (for example, those with absolute error of less than 180, which would have a squared error of at most 34k) were less important than large errors (e.g., predicting 1800 when the true rating was 2500, corresponding to a squared error of 490k).

We experimented a with piecewise loss, where the model was penalized more harshly for larger errors. For example, we considered loss functions with the basic form:

$$\mathcal{L}(\hat{r}, r) = \begin{cases} (\hat{r} - r)^2, & \text{if } |\hat{r} - r| \leq T \\ (\hat{r} - r)^4, & \text{otherwise,} \end{cases}$$

where r and $\hat{r}$ are the (unnormalized) true and predicted ratings, respectively, and T is some threshold (180, in the example above, though we used normalized rating thresholds in the actual implementation). We also tried more stratified loss functions (e.g. with multiple thresholds and loss scaling functions), but did not observe any improvements.

REFERENCES

[1] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F5937

[2] M. E. Glickman, "Example of the glicko-2 system," http://www.glicko.net/glicko/glicko2.pdf, 2022.

[3] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji team's submission to the ieee bigdata 2024 cup: Predicting chess puzzle difficulty challenge," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 8415–8422.

[4] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "Ieee big data cup 2024 report: Predicting chess puzzle difficulty at knowledgepit.ai," in *IEEE International Conference on Big Data, Big Data 2024, Washington DC, USA, December 15-18, 2024*. IEEE, 2024.

[5] D. Klein, "Neural networks for chess," 2022. [Online]. Available: https://arxiv.org/abs/2209.01506

[6] R. McIlroy-Young, S. Sen, J. M. Kleinberg, and A. Anderson, "Aligning superhuman AI with human behavior: Chess as a model system," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 1677–1687. [Online]. Available: https://doi.org/10.1145/3394486.3403219

[7] Z. Tang, D. Jiao, R. McIlroy-Young, J. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A unified model for human-ai alignment in chess," 2024. [Online]. Available: https://arxiv.org/abs/2409.20553

[8] T. L. Authors, "Leela chess zero," https://lczero.org/.

[9] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," 2017. [Online]. Available: https://arxiv.org/abs/1712.01815

[10] S. Björkqvist, "Estimating the puzzlingness of chess puzzles," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 8370–8376.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 843–848
DOI: 10.15439/2025F2456
ISSN 2300-5963 ACSIS, Vol. 43

Multi-Source Feature Fusion and Neural Embedding for Predicting Chess Puzzle Difficulty

Haitao Xiao*,†, Daiyuan Yu*,†, Xuegang Wen*,†, Le Chen*,†, Kun Fu*,†
* China Mobile Information Technology Co., Ltd., China
† China Mobile Communications Group Co., Ltd., China
Email: {xiaohaitao, yudaiyuan, wenxuegang, chenle, fukun}@chinamobile.com

***Abstract*—Estimating the difficulty of chess puzzles provides a rich testbed for studying human–computer interaction and adaptive learning. Building on recent advances and the FedCSIS 2025 Challenge, we address the task of predicting chess puzzle difficulty ratings using a multi-source representation approach. Our approach integrates pre-trained neural embeddings of board states, solution move sequences, and engine-derived success probabilities. These heterogeneous features are fused via dedicated embedding and projection layers, followed by a multi-layer perceptron regressor. Post-processing calibration and model ensemble further enhance robustness and generalization. Experiments on the FedCSIS 2025 dataset demonstrate that our method effectively leverages both structural and empirical information, achieving strong predictive performance. Our approach achieved fifth place on the final official leaderboard, highlighting the effectiveness of combining neural representations with domain-specific probabilistic features for robust chess puzzle difficulty prediction.**

***Index Terms*—Human-Computer Interaction, Chess Puzzle Difficulty, Multi-Source Feature Fusion, Representation Learning, Ensemble Learning**

I. Introduction

Chess has served not only as a competitive arena, but also as a richly structured and controlled testbed for exploring the foundations of human–computer interaction. From Deep Blue's brute-force victory over Kasparov [1] to AlphaZero's self-taught superhuman play [2], each algorithmic milestone has advanced a deeper goal: understanding how machines can model, anticipate, and ultimately support human cognitive behavior. Contemporary online chess platforms now record every move, reaction time, and mistake across millions of human–machine interactions. These data enable a direct estimation of puzzle difficulty from behavioral patterns. Predicting chess puzzle difficulty thus emerges as a core task that bridges cognitive science, adaptive learning, and recommender system design.

While chess has long served as a model system for cognitive research, it is in the advent of modern online platforms that has enabled large-scale, quantitative analysis. Lichess [1] is a widely used open-source online chess platform that provides millions of user-generated chess puzzles. The difficulty of each puzzle is quantified using rating systems originally designed for human players. The most common systems include the Elo rating [3], which updates a player's rating based on game outcomes, and the Glicko [4] and Glicko-2 [5] systems, which further incorporate rating volatility and adjust more dynamically to player performance. On Lichess, each puzzle receives a Glicko-2 rating that reflects its empirical difficulty for the average user, with accuracy improving as more players attempt the puzzle.

Building on the IEEE BigData 2024 Cup [6], which demonstrated the utility of feature-rich and neural approaches for chess puzzle difficulty prediction [7–12], the FedCSIS 2025 Challenge [13] introduces both a larger dataset and new data modalities. The main objective remains to predict the difficulty rating of a chess puzzle using its initial board state and solution moves. However, in contrast to the first edition, the FedCSIS 2025 Challenge provides 22 precomputed engine-based success probabilities per puzzle, generated by Maia-2 models [14] to simulate human move likelihoods across different player ratings and types of rating. This addition eliminates the need for costly local engine simulations and facilitates more equitable benchmarking of model designs. With a training set of over 4.5 million puzzles, FedCSIS 2025 offers an enriched setting for advancing research in chess puzzle difficulty prediction.

In this study, we propose a multi-source neural representation approach for predicting chess puzzle difficulty. Our approach integrates heterogeneous information from pre-trained board state embeddings, solution move sequences, and a 22-dimensional vector of engine-estimated success probabilities. These complementary features are jointly fused and embedded through dedicated projection layers, followed by a multi-layer perceptron regressor to predict the rating. To further enhance robustness and address distributional shifts, we apply post-processing calibration and ensemble models trained under diverse settings. This design leverages both the structural and empirical dimensions of puzzle difficulty, leading to improved predictive performance and generalization.

The remainder of this paper is organized as follows: Section II analyzes the dataset and outlines the preprocessing pipeline. Section III describes the proposed approach. Section IV details the experimental setup and results. Section V discusses findings, limitations, and potential ways to improve in future work. Section VI concludes the paper.

[1] https://lichess.org

II. Data Analysis

A comprehensive understanding of the dataset structure is essential for designing robust prediction models in this challenge. Table I provides an overview of the principal features included in the competition dataset. Each chess puzzle is uniquely identified by a `PuzzleId` and is characterized by several core components: the board position specified in Forsyth–Edwards Notation (FEN), the solution sequence encoded in Portable Game Notation (PGN), and a 22-dimensional vector of engine-estimated success probabilities (`SuccessProb`), which represents projected solve rates across diverse player rating groups and game types. The `Rating` field denotes the puzzle's Glicko-2 difficulty rating [5], which serves as the primary target variable for model training.

TABLE I
Summary of Features in the Challenge Dataset

Field Name	Description	Type
PuzzleId	Unique identifier	String
FEN	Board position	String
Moves	Solution in PGN	String
SuccessProb	Success probabilities	Float
Rating	Glicko-2 puzzle rating (Target)	Integer
RatingDeviation	Uncertainty in rating	Integer
Popularity	Upvotes minus downvotes	Integer
NbPlays	Number of attempts	Integer
Themes	Puzzle motif tags	String
GameUrl	Lichess game provenance	String
OpeningTags	Opening classification	String

In addition to these core fields, the training data comprises several auxiliary metadata attributes. These include the rating uncertainty (`RatingDeviation`), popularity score (`Popularity`), total number of attempts (`NbPlays`), thematic tags (`Themes`), and optional fields such as the original game URL (`GameUrl`) and opening classification (`OpeningTags`). This rich set of features enables multi-faceted analysis and facilitates the construction of both neural and feature-based models.

TABLE II
Comparison of Training and Test Datasets

Property	Training Set	Test Set
Instances	4,557,000	2,235
Features	PuzzleId, FEN, Moves, SuccessProb, RatingDeviation, Popularity, NbPlays, Themes, GameUrl, OpeningTags	PuzzleId, FEN, Moves, SuccessProb
Target	Rating	-

Table II presents a comparative summary of the training and test datasets used in this challenge. The training set contains 4,557,000 puzzle instances, while the test set contains 2,235 instances. Notably, the test set is restricted to the core features: `PuzzleId`, `FEN`, `Moves`, and `SuccessProb` with the ground-truth `Rating` field hidden to facilitate unbiased model evaluation. Consequently, additional metadata present in the training set, such as `RatingDeviation`, `Popularity`, and `Themes`, must be excluded from the feature set during model development to ensure strict compatibility between training and inference conditions.

The structure of the dataset, characterized by both structural descriptors and empirical engine-based probabilities, enables the exploration of diverse modeling strategies. The inclusion of precomputed success probabilities is particularly noteworthy, as it reduces the computational barrier for participants and provides valuable prior information for puzzle difficulty prediction, especially for those without access to sufficient local computational resources.

Although the dataset provides a comprehensive set of structural and empirical features, ensuring the reliability of the target variable is crucial for robust model training. In particular, the rating uncertainty (`RatingDeviation`) quantifies the confidence of each puzzle's Glicko-2 difficulty estimate and varies substantially across puzzles.

Puzzles exhibiting high rating uncertainty generally correspond to unstable or unreliable difficulty estimates, often resulting from insufficient player attempts or inconsistent solution patterns. To improve data quality and label reliability, all puzzles with a `RatingDeviation` greater than 90 were excluded from the training set. This threshold, which aligns with established practice in prior studies, effectively reduces label noise by filtering out puzzles with highly variable ratings.

After this preprocessing step, the size of the training set decreased from approximately 4.56 million to 3.53 million puzzles. The resulting dataset, comprising high-quality and reliable labels, was used for model training. This preprocessing pipeline enhances the stability of the learning process and ensures that subsequent modeling are grounded in representative data.

III. Methodology

Figure 1 presents an overview of the proposed solution pipeline for chess puzzle difficulty prediction. The overall methodology consists of several key components: multi-source feature embedding, a neural regressor for rating prediction, post-processing calibration, model ensemble, and uncertainty estimation. Each component is described in detail below.

A. Feature Embedding

Our approach integrates three complementary sources of information for each chess puzzle: the board state encoded in Forsyth–Edwards Notation (FEN), the solution move sequence represented in Portable Game Notation (PGN), and a 22-dimensional vector of engine-estimated success probabilities (`SuccessProb`). The FEN string offers a detailed, lossless description of the board configuration, while the PGN sequence captures the temporal progression of solution moves. The `SuccessProb` vector summarizes the predicted probability of a successful solve across various player rating brackets and type of rating (rapid or blitz), as precomputed by the MAIA2 neural engine [14].

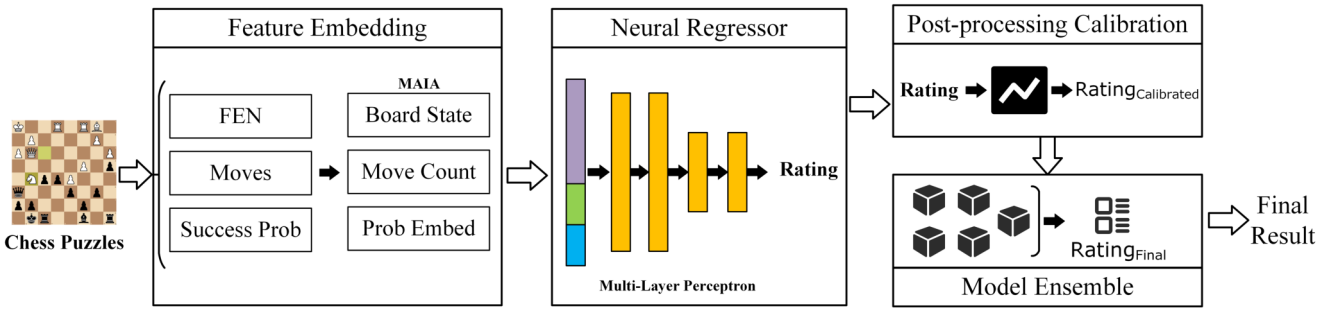


Fig. 1. Overview of proposed approach

To comprehensively capture both the structural and sequential aspects of each puzzle, we generate a sequence of consecutive board states by iteratively applying the solution moves to the initial FEN. Each intermediate position is embedded using a pre-trained Maia neural network, which is specifically trained to emulate human move preferences at defined skill levels [15]. For each board state, we extract a neural embedding from the penultimate hidden layer of the Maia model. These embeddings are mapped into a unified latent space and combined to form a comprehensive representation of the puzzle's solution trajectory.

The solution length is discretized and encoded as a learnable embedding, enabling the model to incorporate information about puzzle complexity. Simultaneously, the `SuccessProb` vector is embedded through a two-layer feedforward network with ReLU activation to produce a dense embedding representing empirical difficulty. Finally, concatenating the aggregated board embedding, the move count embedding, and the success probability embedding yields a unified embedding vector. This multi-source representation integrates structural configuration, sequential solution dynamics, and empirical difficulty priors, providing a rich input for downstream prediction.

B. Neural Regressor

The aggregated feature vector is processed by a multi-layer perceptron regressor designed to capture complex nonlinear relationships between the fused feature embeddings and puzzle difficulty. The architecture comprises a sequence of fully connected layers with progressively decreasing dimensionality, interleaved with ReLU activation functions and dropout regularization to enhance generalization and mitigate overfitting. This deep regression network enables the model to learn intricate mappings from multi-source representations to difficulty ratings. The final output is a single scalar representing the predicted Glicko-2 rating for the given puzzle.

C. Post-processing Calibration

To mitigate the distributional shift between the training set and the competition test data, we employ a simulation-inspired nonlinear rescaling technique [7] to adjust the raw predictions of the neural regressor. This post-processing step compresses prediction values at the distributional extremes while preserving ratings near the empirical mean, thereby reducing the impact of outliers. The calibration function is defined as follows:

$$\hat{r} = \mu - \frac{1+\text{sign}(r-\mu)}{2} \cdot H \cdot \max\left(1, \left|\frac{r-\mu}{D}\right|^4\right) + \frac{1-\text{sign}(r-\mu)}{2} \cdot L \cdot \min\left(1, \left|\frac{r-\mu}{D}\right|^4\right) \quad (1)$$

where $\mu = 1900$ denotes the empirical mean rating, $H = 200$ and $L = 250$ control the scaling magnitude for over- and underestimations respectively, and $D = 1000$ determines the scale at which the rescaling effect saturates. These hyperparameters are selected based on prior studies [7] and further validated through our empirical experience in this challenge.

This calibration strategy aims to improve the alignment between predicted and true rating distributions. By explicitly correcting for known biases in the rating aggregation process, the rescaling aims to make the predicted difficulty ratings more robust.

D. Model Ensemble

To further enhance model robustness and predictive performance, we employ an output-level ensemble strategy. Specifically, multiple base models are trained independently using board embeddings generated from different Maia engine variants, each emulating human play at a distinct ELO rating. The final prediction is computed by averaging the outputs from all base models:

$$\text{Rating}_{\text{final}} = \frac{1}{N}\sum_{i=1}^{N} \text{Model}_i(\text{Input}), \quad (2)$$

where N is the total number of ensembled models. This approach leverages complementary perspectives of models trained on different skill levels and effectively reduces the variance of individual predictions. Ensembling not only mitigates overfitting and model-specific biases but also enhances the model's ability to generalize across puzzles of varying complexity. This strategy is particularly effective in domains where task difficulty spans a wide spectrum and individual models may excel in different sub-regions of the input space.

E. Uncertainty Estimation

To further enhance model reliability and provide interpretable confidence assessments for each prediction, we propose an ensemble-based uncertainty estimation strategy. Specifically, for each test instance, we aggregate the prediction outputs from all base models within the ensemble and compute the standard deviation of these predictions as a proxy for epistemic uncertainty, reflecting the degree of disagreement among ensemble members.

Formally, let $y_i^{(j)}$ denote the prediction for the i-th test sample by the j-th base model in the ensemble, where $j = 1, \ldots, M$. The uncertainty score for the i-th sample is quantified as the standard deviation across the ensemble outputs:

$$\sigma_i = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \left(y_i^{(j)} - \bar{y}_i \right)^2}, \tag{3}$$

where $\bar{y}_i$ represents the mean prediction for the i-th sample. A higher standard deviation σ_i indicates greater predictive uncertainty, as it signals increased model disagreement regarding the sample's difficulty.

In accordance with the challenge requirements, we rank all test samples by their uncertainty scores and flag the top K instances (where K corresponds to 10% of the test set, i.e., $K = 223$ for 2235 samples) using a binary mask. By leveraging the diversity inherent in the ensemble, this strategy systematically identifies predictions with elevated risk of error.

IV. Experiment and Result

A. Experiment Setup

1) Environment: The experimental environment is based on Ubuntu 22.04, equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM). The CPU is an Intel Xeon Gold 6330 operating at 2.00 GHz, complemented by 90 GB of system memory.

2) Toolkit: The implementation is based on Python 3.12, with PyTorch 2.5.1 used for neural network construction and training. CUDA 12.4 is utilized to accelerate GPU computations and enhance overall computational efficiency.

3) Evaluation Metric: The metric for performance evaluation is the Mean Squared Error (MSE), consistent with the official evaluation criterion of the FedCSIS 2025 Challenge.

B. Experiment Result

We conducted a comprehensive set of experiments to evaluate the effectiveness of our approach for chess puzzle difficulty prediction. Table III summarizes model performance on the public leaderboard, measured in terms of MSE. All results correspond to the official public test set provided by the competition platform [16].

As a baseline, we implemented a LightGBM regressor utilizing handcrafted chess features extracted from the board state, move sequence, and basic positional statistics. Feature extraction was performed using the `python-chess` library, which provides utilities for parsing FEN and PGN representations and computing relevant game attributes. This model achieved an MSE of 104703.66, serving as the baseline for subsequent neural network-based approaches.

To investigate the impact of learned representations, we incorporated pre-trained neural board embeddings derived from various Maia engine variants. Employing MAIA-1300 embeddings without the inclusion of organizer-provided success probability vectors resulted in an MSE of 91689.69. Incorporating the success probability features further reduced MSEs for the MAIA-1300, MAIA-1500, and MAIA-1700 models to 82582.37, 85625.50, and 81570.99, respectively, thereby underscoring the complementary value of empirical priors.

To address distributional shift and enhance calibration, we applied a post-processing rescaling procedure to the model outputs, as described in Section III-C. This adjustment led to further improvement, with MSEs decreasing to 76986.22 for MAIA-1300, 78155.64 for MAIA-1500, and 79688.41 for MAIA-1700.

Subsequently, ensemble averaging across the MAIA-1300, MAIA-1500, and MAIA-1700 models yielded an MSE of 75915.97. Our final approach, which ensembles board embeddings from five Maia variants (MAIA-1100, MAIA-1300, MAIA-1500, MAIA-1700, and MAIA-1900) and combines calibrated outputs with success probability features, achieved the best observed performance, attaining an MSE of 67071.66 on the public leaderboard.

TABLE III
Performance Comparison of Different Methods on Public Leaderboard

Method	Public MSE
Handcrafted Features + LightGBM (Baseline)	104703.66
MAIA-1300 (w/o SuccessProb)	91689.69
MAIA-1300	82582.37
MAIA-1500	85625.50
MAIA-1700	81570.99
MAIA-1300 + Post-processing	76986.22
MAIA-1500 + Post-processing	78155.64
MAIA-1700 + Post-processing	79688.41
MAIA-1300/1500/1700 Ensemble	75915.97
MAIA-1100/1300/1500/1700/1900 Ensemble	**67071.66**

The results clearly demonstrate that leveraging pre-trained neural embeddings, incorporating empirical success probabilities, and ensembling the outputs of individually calibrated models together yield substantial improvements over traditional handcrafted feature-based models. The progressive reduction in MSE observed through our stepwise model enhancements highlights the additive value of each component in our solution. Ultimately, our final ensemble approach achieved a marked performance gain relative to classical models. These findings confirm the effectiveness of multi-source feature fusion and neural representation learning for chess puzzle difficulty prediction.

C. Uncertainty Estimation Result

To quantitatively evaluate the effectiveness of our uncertainty estimation strategy, we participated in the additional uncertainty mask task organized as part of the challenge. In this task, each team was required to submit a binary mask identifying the most uncertain 10% of the test puzzles. Let P denote the perfect score, obtained by replacing the predictions for the 10% most erroneous test cases with their ground-truth values, and let N denote the score achieved using the submitted mask. The evaluation metric is defined as:

$$\rho = \frac{N}{P} \tag{4}$$

Our submitted mask achieved a ratio of $\rho = 1.589$, ranking **3rd** among participating teams. On the official leaderboard, this corresponded to a score of approximately 55234, compared with the theoretical lower bound of 34766 under a perfect mask. The results show that our proposed uncertainty estimation strategy provides a reliable means of identifying error-prone cases, with potential for further refinement.

V. Discussion

Our experimental results reveal several insights and limitations that warrant further investigation. While handcrafted features extracted from the board state and move sequences provide a valuable baseline for chess puzzle difficulty prediction, they are inherently limited in their ability to capture the full complexity of positional and sequential information. Models based solely on such features consistently underperform relative to deep neural architectures, likely due to their inability to represent the nuanced dynamics encoded in FEN strings and move sequences, both of which are essential for accurate difficulty modeling.

Nevertheless, prior work and our own attempts indicate that carefully engineered handcrafted features, if effectively integrated into neural network architectures, may offer complementary benefits. However, our straightforward approach to merging these features with neural embeddings did not yield performance improvements, suggesting that more advanced integration strategies, such as attention-based models, may be necessary to unlock the full potential of hybrid feature sets.

Due to computational constraints and time limitations, our current study utilized only five Maia pre-trained model variants, corresponding to targeted ELO levels of 1100, 1300, 1500, 1700, and 1900. Each Maia variant is designed to emulate the play style of human users at its specific rating level, thereby enabling the model to capture a spectrum of human skill profiles. Nonetheless, this approach may not fully represent the diversity of the solver population. Future research should consider integrating all available Maia models, spanning ELO 1100 to 1900, to achieve finer granularity in modeling player abilities. Such comprehensive ensembles could provide a more nuanced simulation of the human skill continuum and further improve prediction robustness.

In summary, our findings underscore the importance of both high-capacity neural architectures and the principled fusion of domain-specific features for robust chess puzzle difficulty prediction. Advancing this field will require continued research into sophisticated feature integration strategies and more comprehensive utilization of human-like engine ensembles, ultimately paving the way for models with enhanced generalization and interpretability.

VI. Conclusion

In this work, we investigated chess puzzle difficulty prediction using a multi-source feature fusion approach based on neural embedding techniques, within the context of the FedCSIS 2025 Challenge. Our approach integrates pre-trained Maia board embeddings, solution move sequences, and success probabilities generated by the MAIA2 engine. These heterogeneous information sources are combined within a unified neural regression model. Results on the official public leaderboard demonstrate that incorporating empirical priors derived from chess engines, neural network-based representations, model ensemble, and post-processing calibration leads to improved performance over traditional feature-based baselines.

In addition, we acknowledge the overarching goal of the knowledgepit competition series [17–20], which is to evaluate the skills of data scientists through carefully designed tasks [16]. Beyond its evaluative role, the knowledgepit provides an excellent platform for testing novel methodologies, comparing approaches under common benchmarks, and advancing the broader data science community.

Acknowledgment

We thank the organizers of the FedCSIS 2025 Challenge for their efforts in hosting this competition and maintaining a fair and engaging environment.

References

[1] M. Campbell, A. Hoane, and F. hsiung Hsu, "Deep blue," *Artificial Intelligence*, vol. 134, no. 1, pp. 57–83, 2002. doi: https://doi.org/10.1016/S0004-3702(01)00129-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370201001291

[2] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018. doi: 10.1126/science.aar6404. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aar6404

[3] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Pub., 1978. ISBN 0668047216 9780668047210

[4] M. E. Glickman, "The glicko system," *Boston University*, vol. 16, no. 8, p. 9, 1995.

[5] ——, "Example of the glicko-2 system," *Boston University*, vol. 28, p. 2012, 2012.

[6] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "Ieee big data cup 2024 report: Predicting chess puzzle difficulty at knowledgepit.ai," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825289 pp. 8423–8429.

[7] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji team's submission to the ieee bigdata 2024 cup: Predicting chess puzzle difficulty challenge," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10826037 pp. 8415–8422.

[8] A. Schütt, T. Huber, and E. André, "Estimating chess puzzle difficulty without past game records using a human problem-solving inspired neural network architecture," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10826087 pp. 8396–8402.

[9] S. Björkqvist, "Estimating the puzzlingness of chess puzzles," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825991 pp. 8370–8376.

[10] A. Rafaralahy, "Pairwise learning to rank for chess puzzle difficulty prediction," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825356 pp. 8385–8389.

[11] D. Ruta, M. Liu, and L. Cen, "Moves based prediction of chess puzzle difficulty with convolutional neural networks," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825595 pp. 8390–8395.

[12] S. Miłosz and P. Kapusta, "Predicting chess puzzle difficulty with transformers," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825919 pp. 8377–8384.

[13] J. Zyśko, M. Ślęzak, D. Ślęzak, and M. Świechowski, "FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. doi: 10.15439/2025F5937. [Online]. Available: http://dx.doi.org/10.15439/2025F5937

[14] Z. Tang, D. Jiao, R. McIlroy-Young, J. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A unified model for human-ai alignment in chess," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 919–20 944, 2024. doi: 10.48550/arXiv.2409.20553

[15] R. McIlroy-Young, S. Sen, J. Kleinberg, and A. Anderson, "Aligning superhuman ai with human behavior: Chess as a model system," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020. doi: 10.1145/3394486.3403219 pp. 1677–1687.

[16] D. Ślęzak, A. Janusz, M. Świechowski, A. Chądzyńska-Krasowska, and J. Kamiński, "Do data scientists dream about their skills' assessment? – transforming a competition platform into an assessment platform," in *2024 IEEE International Conference on Big Data (BigData)*, 2024. doi: 10.1109/BigData62323.2024.10825378 pp. 8403–8414.

[17] A. Janusz, M. Przyborowski, P. Biczyk, and D. Ślęzak, "Network device workload prediction: A data mining challenge at knowledge pit," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020. doi: http://dx.doi.org/10.15439/2020F159 pp. 77–80.

[18] A. Janusz, A. Jamiołkowski, and M. Okulewicz, "Predicting the costs of forwarding contracts: Analysis of data mining competition results," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2022. doi: http://dx.doi.org/10.15439/2022F303 pp. 399–402.

[19] M. Czerwiński, M. Michalak, P. Biczyk, B. Adamczyk, D. Iwanicki, I. Kostorz, M. Brzęczek, A. Janusz, M. Hermansa, Ł. Wawrowski *et al.*, "Cybersecurity threat detection in the behavior of iot devices: analysis of data mining competition results," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023. doi: http://dx.doi.org/10.15439/2023F3089 pp. 1289–1293.

[20] A. M. Rakićević, P. D. Milošević, I. T. Dragović, A. M. Poledica, M. M. Zukanović, A. Janusz, and D. Ślęzak, "Predicting stock trends using common financial indicators: A summary of fedcsis 2024 data science challenge held on knowledgepit. ai platform," in *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024. doi: http://dx.doi.org/10.15439/2024F7912 pp. 731–737.

Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 849–854
DOI: 10.15439/2025F5937
ISSN 2300-5963 ACSIS, Vol. 43

FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests

Jan Zyśko*, Michał Ślęzak†‡, Dominik Ślęzak*‡, Maciej Świechowski§‡
*Institute of Informatics, University of Warsaw, Poland
†Polish-Japanese Academy of Information Technology, Poland
‡QED Software, Poland
§Grail Team, Poland

***Abstract*—We summarize the results of the FedCSIS 2025 machine learning competition organized on the knowledgepit.ai online platform. We recall the competition's goals corresponding to estimations of the chess puzzle difficulty levels, we refer to the winning solutions, and we also compare the scope of this year's competition (and particularly the data available to competition participants) with its previous edition associated with the IEEE BigData 2024 conference. Finally, we discuss the new functionality of the knowledgepit.ai platform, which enables competition participants to submit additional uncertainty masks reflecting their assessment of test cases that are mostly problematic for their machine learning models.**

***Index Terms*—Human-Computer Interaction, Chess, Machine Learning Competitions, Uncertainty in Machine Learning**

I. Introduction

THE FedCSIS 2025 machine learning competition continues the successful blueprint established in the IEEE BigData 2024 edition [1], held at the knowledgepit.ai platform as well, addressing the challenge of predicting chess puzzle difficulty ratings from board configurations and solution sequences. While the core task remains unchanged – estimating Glicko-2 ratings that reflect human-perceived difficulty rather than engine-optimal evaluations – this second edition introduces significant enhancements in both data quality and modeling capabilities. The competition attracted 42 teams who collectively submitted 1185 solutions, with eight of the top ten teams contributing technical papers describing their approaches.

The primary innovations in this edition address key limitations of the previous competition. First, we integrated 22 precomputed success probability features generated by the Maia-2 model [2], providing participants with standardized human-aligned difficulty indicators across multiple rating bands and time controls. Second, we substantially improved test set quality through expanded data collection and a novel fairness algorithm that ensured balanced solving attempts across all puzzles, eliminating the rating convergence issues that plagued the first edition, where many puzzles remained clustered around the 1500 initialization value. These improvements resulted in more reliable ground truth labels and enabled more sophisticated modeling approaches, as evidenced by the winning solutions' creative use of ensemble methods, transformer architectures, and domain-specific feature engineering.

Beyond the core regression task, this competition pioneers a new dimension in machine learning competitions through the introduction of uncertainty masks – an extra challenge where participants identify the 10% of test cases their models find most problematic. This extension represents a step toward more interpretable and reliable machine learning systems, addressing the growing need for models that can communicate their (un)certainty levels alongside predictions.

The next sections describe the related work (Section II), the dataset (Section III), and the competition setup (Section IV). Further, we analyze the submitted solutions (Section V), discuss prediction errors and masking strategies of the top teams (Section VI), and conclude with implications for future competitions and research directions (Section VII).

II. Related Work

The challenges of the assessment of chess puzzle difficulty were discussed in our paper related to the first edition of this competition [1]. We also refer to [3]–[5] where the Top 3 solutions submitted to the first edition were reported.

The importance of machine learning competitions was also discussed in our previous papers [6]–[10]. In particular, we investigated possibilities of utilizing such competitions to build a platform for automatic evaluation of skills of data scientists [11]. As already mentioned, the new idea of extending the format of knowledgepit.ai competitions with uncertainty masks corresponds to that aspect of our research as well.

This new aspect of our competitions corresponds to an important topic in machine learning [12]. Although the most problematic cases for machine learning models are not necessarily the most "uncertain" ones, measures of uncertainty are often used, e.g., in active learning [13] and model diagnostics [14]. In our research, we estimate a given object's uncertainty using a neighborhood of cases having the same or similar values over subsets of attributes that are significant to the given model or decision problem [15]. The attribute selection mech-

anism described in [15] was also applied to build a benchmark machine learning model for this particular competition.

III. New Chess Puzzle Dataset

This competition builds upon the dataset foundation established in the previous IEEE BigData 2024 edition [1], while introducing a more robust test set and a significant enhancement to support more sophisticated modeling approaches.

A. Core Dataset Structure

Similarly to the previous edition, the dataset consists of chess puzzles with basic fields including PuzzleId, FEN (board position), Moves (puzzle solution), and the target Rating to be predicted. The training dataset retains the extra metadata fields from the original competition: RatingDeviation, Popularity, NbPlays, Themes, GameUrl, and OpeningTags.

B. Success Probability Features

The primary innovation in this edition is the addition of 22 success probability features generated using the Maia-2 model. Unlike the original Maia models that required separate trained models for each rating group [16], Maia-2 employs a unified architecture that jointly models success probabilities across all skill levels [2]. This model takes as input the chess position, player rating, and game type (rapid or blitz), outputting predicted success probabilities for solving the puzzle. Specifically, the Success_prob fields represent the estimated probability that a player of a given rating level and game type would correctly solve each puzzle. These 22 fields cover different rating ranges and both rapid and blitz time controls, providing participants with rich, model-generated features that capture human solving patterns across the skill spectrum.

Inspired by the innovative approaches observed in the top solutions from the previous competition [3]–[5], this enhancement aimed to provide participants with features that better capture the human element of puzzle-solving difficulty. Rather than requiring each team to independently extract chess engine evaluations for millions of puzzles, the Maia-2-generated success probabilities offer a standardized set of features that encode human-like assessment of puzzle difficulty across different skill levels.

C. Improved Test Set Quality

A significant improvement over the previous competition relates to the quality and distribution of puzzle attempt data. The IEEE BigData 2024 competition suffered from insufficient rating convergence in the test set, with many puzzles remaining near their initialized rating of 1500 due to limited solving attempts [1]. To address this limitation, the FedCSIS 2025 edition implemented two key enhancements:

- Expanded data collection: A larger pool of chess players was recruited to solve puzzles, resulting in substantially more solving attempts across the test set.
- Fairness algorithm: A novel algorithm was integrated into our lichess fork that prioritized presenting puzzles with fewer total attempts (regardless of correctness) to users. This ensured a more balanced distribution of solving attempts across all puzzles, leading to better rating convergence and more reliable ground truth labels.

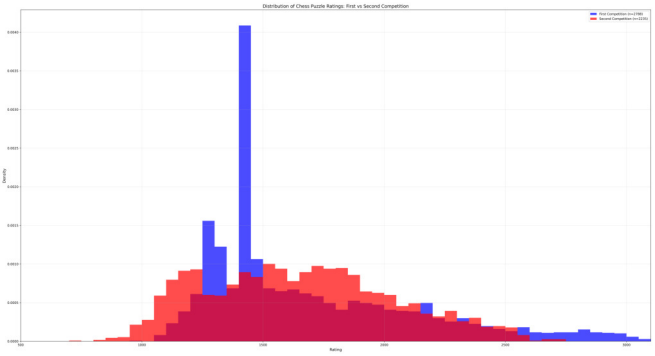


Fig. 1: Distribution of puzzle ratings in test sets: First competition (IEEE BigData 2024, blue) vs. Second competition (FedCSIS 2025, red). The first edition shows a pronounced spike at 1500 (the initialization value), while the second edition exhibits a more natural distribution with better rating convergence across the full spectrum.

Figure 1 illustrates the dramatic improvement in rating distribution quality. The first competition's test set (blue) exhibits a sharp peak around 1500—the default initialization value—indicating that many puzzles received insufficient solving attempts to converge to their true difficulty. In contrast, the FedCSIS 2025 test set (red) displays a smoother, more naturally distributed profile spanning from approximately 800 to 3000, demonstrating successful rating convergence. This improvement directly translates to more reliable ground truth labels for model evaluation.

IV. Competition Task

The core regression task of the FedCSIS 2025 Challenge remained fundamentally the same as in the first edition [1], requiring participants to predict difficulty ratings for chess puzzles based on their FEN board states and PGN solution sequences. The primary difference was the dataset size, which contained 2283 puzzles after removing those with fewer than 10 solving attempts during data annotation. As before, predictions were evaluated using Mean Squared Error (MSE) between predicted ratings $\hat{y}_i$ and ground-truth ratings y_i, computed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{1}$$

where n denotes the number of test samples.

A novel extension to this edition introduced an uncertainty estimation task, whereby participants were invited to submit binary masks identifying the 10% of test puzzles for which their predictions were most likely to be erroneous. The mask $M \in \{0,1\}^N$ required each participant to flag exactly 10% of the test set as uncertain, with $M_i = 1$ indicating high uncertainty for puzzle i and the constraint $\sum_{i=1}^{N} M_i = N \times 0.1$. The

TABLE I: Top 10 submissions in the FedCSIS 2025 knowledgepit.ai competition. The evaluation metric was the mean squared error – the lower Final (and preliminary, Pre.) score, the better. The Subs column reflects the total number of submissions. The Mask column reports the uncertainty ratio – again, the lower, the better (the best score bolded). Eight out of Top 10 teams submitted conference papers describing their solutions. Nine teams participated in the additional uncertainty mask contest.

Rank	Team	Paper	Final	Pre.	Subs	Mask	Approach
1	ousou	[17]	**52.3k**	**55.4k**	89	1.64	Fine-tuned Maia-2 embeddings combined with other features through LightGBM, ensembled via CatBoost with linear scaling.
2	bread emoji	[18]	54.4k	58.2k	240	1.71	Two-stage approach with ensembled neural chessboard embedders, pre-RNN board pooling, and logarithmic rescaling.
3	transformer_enjoyer	[19]	55.9k	58.9k	117	**1.43**	Spatial transformer with masked-square reconstruction and solution policy prediction, rescaling borrowed from [3].
4	ToDoFindATeamName	–	57.5k	61.5k	136	–	–
5	Cyan	[20]	61.0k	67.1k	93	1.59	Multi-variant ensemble of Maia embeddings, solution sequences, and engine success probabilities with nonlinear calibration.
6	neuro	[21]	62.6k	66.5k	81	1.68	Four-stage pipeline with Elo-banded base models, gradient boosting, structural calibration via failure distributions, and ensemble averaging.
7	xyz	[22]	62.7k	66.6k	25	1.70	Three-stage pipeline with four Elo-banded MLP base models, stacking ensemble, and structure-aware residual correction.
8	DML	[23]	63.0k	66.7k	73	1.65	Three-stage pipeline combining GB models, multi-modal CNN with EfficientNetB3-rendered board images, and XGB residue stacking.
9	Ru	[24]	67.5k	70.4k	53	1.62	Three-stage pipeline with MobileNetV2 board images, LightGBM residual refinement, and adjustments based on failure distribution.
10	Feiwyth	–	68.1k	70.7k	17	1.58	–

rationale was that by replacing the predicted ratings at these masked positions with their ground-truths, the evaluation could assess a model's ability to identify its own confidence.

The uncertainty masks were evaluated using the Uncertainty Ratio metric, defined as:

$$\mathrm{UR} = \frac{\mathcal{N}}{\mathcal{P}}, \tag{2}$$

where $\mathcal{N}$ represents the New Score (MSE after replacing masked predictions with ground truth) and $\mathcal{P}$ represents the Perfect Score (the minimum achievable MSE if the 10% highest-error samples were perfectly identified and masked). An optimal uncertainty ratio approaches 1.0, indicating perfect identification of the most erroneous predictions. For example, a ratio of 1.2 would indicate that the submitted mask captured errors that, when corrected, achieved 83.3% (1/1.2) of the theoretically optimal improvement, demonstrating reasonably good uncertainty estimation while leaving room for improvement in identifying the truly most difficult cases.

V. Competition Overview

Table I presents the top 10 submissions in the FedCSIS 2025 knowledgepit.ai competition, where teams competed to predict chess puzzle difficulty using mean squared error as the evaluation metric. Nine of the top 10 teams submitted conference papers describing their solutions and participated in the additional uncertainty mask contest.

We refer to Table I and papers [17]–[24] for detailed descriptions of particular competition solutions. In general, the best-performing models leveraged pretrained neural chess embeddings – particularly from the Maia family of models designed to mimic human play at various skill levels – combined with ensemble methods and gradient boosting techniques. The winning team achieved an MSE of 52.3k by fine-tuning Maia-2 embeddings and combining them with other features through LightGBM, followed by CatBoost ensembling with linear scaling. The second-place team (54.4k MSE) employed a two-stage approach with ensembled neural chessboard embedders and pre-RNN board pooling, while the third-place team (55.9k MSE) utilized spatial transformers with masked-square reconstruction. A common pattern across successful approaches was the use of multi-stage pipelines: initial feature extraction via neural board embeddings (Maia-1, Maia-2, or Leela), intermediate refinement through gradient boosting methods (LightGBM, XGBoost, or CatBoost), and final calibration via post-processing techniques addressing distributional shifts.

For uncertainty estimation, most teams developed mask prediction strategies that identified the least reliable predictions, achieving uncertainty ratios between 1.43 and 1.71. The transformer_enjoyer team achieved the best uncertainty ratio of 1.43 despite placing third overall, suggesting that model confidence estimation and raw prediction accuracy represent distinct challenges. Several approaches also incorporated the competition-provided Maia-2 success probabilities – precomputed estimates of puzzle completion likelihood for players at different skill levels – as auxiliary features, though their integration strategies varied considerably across teams.

VI. Error Analysis and Masking Strategies

A. Analysis of Prediction Errors

We conducted several post-hoc analyses of the rating prediction task. Figure 2 shows prediction errors by puzzle difficulty. A U-shaped pattern is visible: models perform worse on very easy (< 1000) and very hard (> 2400) puzzles, while medium-difficulty puzzles are predicted more accurately. This behavior is consistent with the MSE objective, which emphasizes outliers at both ends of the scale.

We then analyzed performance as a function of rating uncertainty (Figure 3). Prediction errors increase monotonically with the Glicko rating deviation of puzzles, indicating that puzzles with higher inherent rating uncertainty are systematically

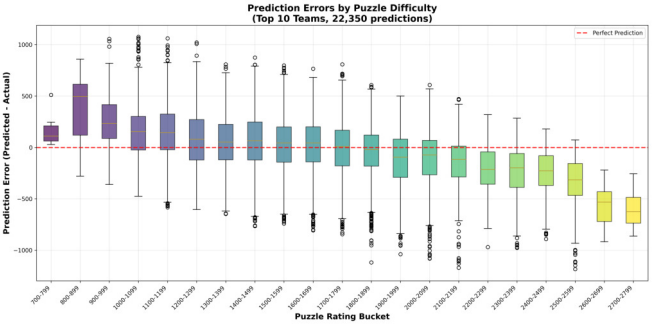


Fig. 2: Prediction errors by puzzle difficulty.

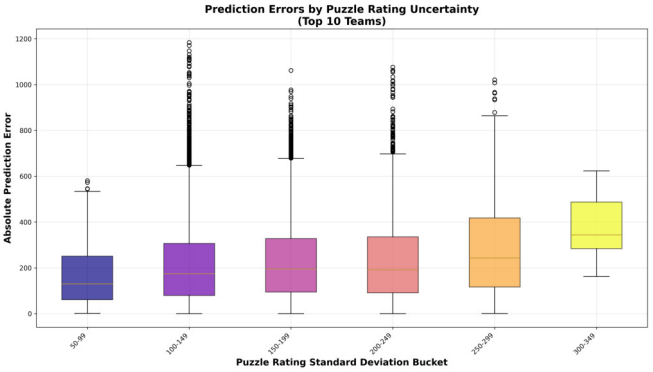


Fig. 3: Prediction errors by rating uncertainty.

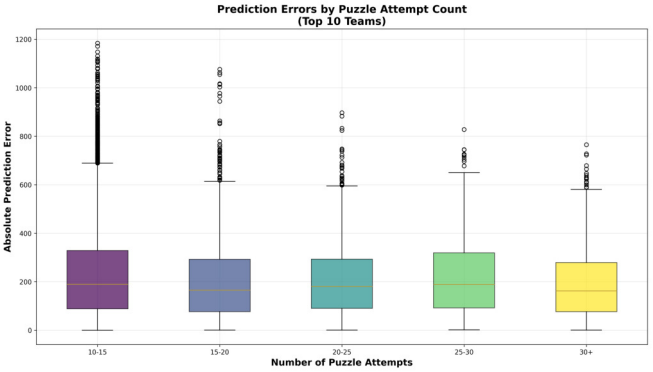


Fig. 4: Prediction errors by the number of attempts during data annotation phase.

harder to model. This suggests that rating deviation, rather than the number of attempts, is a more reliable signal for deciding how well a puzzle's difficulty has been established.

Finally, we investigated whether the number of puzzle attempts during tagging correlates with error (Figure 4). We found no significant trend: puzzles with 10–15 attempts yield error levels comparable to those with 30 or more attempts. Taken together, these findings imply that once a puzzle's Glicko deviation has stabilized, additional attempts provide little benefit. Thus, rating deviation should guide tagging efforts more directly than attempt count.

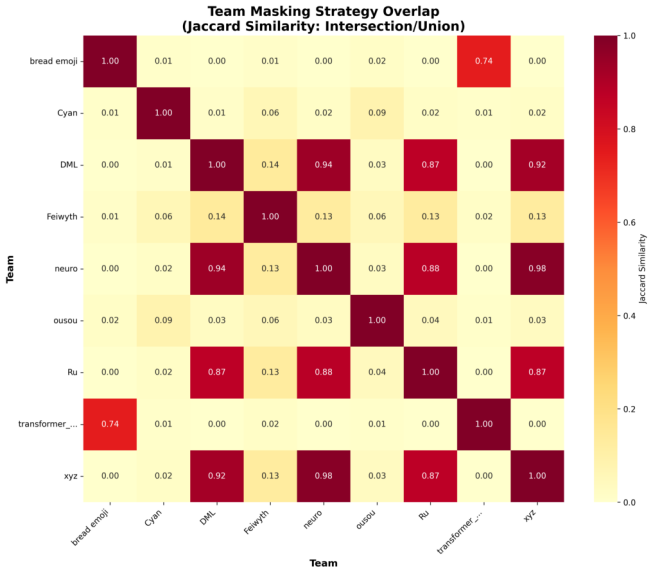


Fig. 5: Jaccard similarity matrix showing overlap between teams' masking strategies.

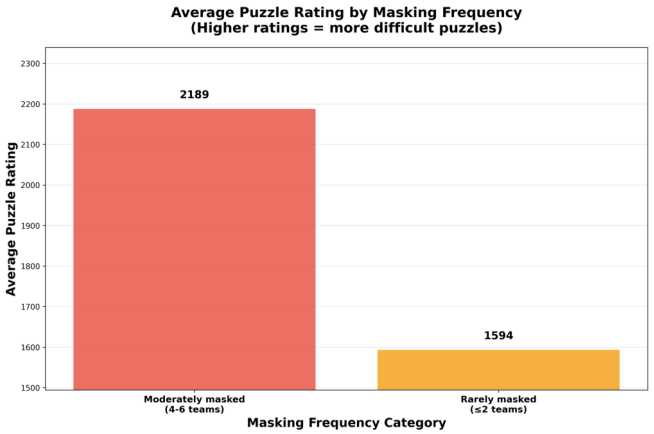


Fig. 6: Puzzle characteristics grouped by masking frequency across teams.

B. *Masking Strategy Analysis*

We next examined the masking strategies employed by teams, focusing both on the convergence of approaches and the characteristics of commonly excluded puzzles. Figure 5 shows the Jaccard similarity between teams' masking choices. Several teams converged on remarkably similar strategies, with DML, neuro, Ru, and xyz achieving similarity scores of 0.87–0.98. This suggests that these teams independently identified a common set of problematic puzzles. By contrast, teams such as bread emoji and transformer_enjoyer pursued more distinctive approaches, showing consistently lower overlap with others.

Beyond team overlap, we analyzed which puzzles were most frequently masked (Figure 6). A clear trend emerges: puzzles targeted by many teams tend to be significantly more difficult ($\approx$ 2200) than those rarely masked ($\approx$ 1600). This indicates

that masking was not random but systematically biased toward high-difficulty puzzles, which teams collectively judged as less reliable or harder to model.

VII. Conclusions

The FedCSIS 2025 knowledgepit.ai competition highlighted the effectiveness of combining neural chess embeddings with ensemble methods and calibration strategies for predicting puzzle difficulty. The top-performing teams consistently employed multi-stage pipelines in which pretrained embeddings such as Maia-2 served as the foundation for downstream boosting models and refined scaling. Despite differences in architecture, a recurring theme was the integration of multiple feature sources – ranging from engine success probabilities to solution policy predictions – followed by systematic post-processing. The additional mask contest further demonstrated that raw predictive accuracy and confidence estimation are distinct goals – while the winning team achieved the lowest MSE overall, the strongest uncertainty modeling was achieved by a different team. This emphasizes the importance of addressing not only point prediction but also model reliability.

Post-hoc analyses of prediction errors revealed consistent patterns across submissions. Errors followed a U-shaped trend with respect to puzzle difficulty, increasing at both extremes of the scale, while rating deviation emerged as a more reliable indicator of prediction uncertainty than attempt count. These findings suggest that puzzle ambiguity, as quantified by Glicko deviation, poses a fundamental barrier to modeling accuracy and that additional attempts beyond stabilization yield limited benefit. Overall, the competition illustrates the dual challenges of modeling puzzle difficulty and quantifying uncertainty, pointing to promising directions for future research in hybrid modeling and calibrated confidence estimation.

In the future, we intend to embed the mechanism of uncertainty masks into all challenges held at the knowledgepit.ai platform – not only as additional tasks but as the main competition objectives. This is because the ability to identify the hardest cases in the data is the key element of advanced machine learning processes [13], and therefore, we need to develop tools for measuring whether the competition participants possess the appropriate skills in this regard [11]. While doing this, however, we need to remember that uncertainty measures are not necessarily the only way to identify those most problematic cases for machine learning models.

As another future topic, encouraged by the aforementioned fact that most of the successful teams relied heavily on non-trivial feature engineering, we are going to further improve our benchmark competition model, this time taking into account also the attributes gathered from papers [17]–[24], organizing them within semantically meaningful groups, and deriving ensembles of their subsets as proposed in [15].

Acknowledgments

This research was co-funded by Smart Growth Operational Programme 2014-2020, financed by European Regional Development Fund, in frame of project POIR.01.01.01-00-1070/21, operated by National Centre for Research and Development in Poland. We are also grateful to chess players who helped us with the difficulty ratings of puzzles in the test data.

References

[1] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8423–8429. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825289

[2] Z. Tang, D. Jiao, R. McIlroy-Young, J. M. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A Unified Model for Human-AI Alignment in Chess," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10-15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/250190819ff1dda47cd23cecc0c5a69b-Abstract-Conference.html

[3] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8415–8422. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10826037

[4] A. Schütt, T. Huber, and E. André, "Estimating Chess Puzzle Difficulty Without Past Game Records Using a Human Problem-Solving Inspired Neural Network Architecture," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8396–8402. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10826087

[5] S. Björkqvist, "Estimating the Puzzlingness of Chess Puzzles," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8370–8376. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825991

[6] A. Janusz, M. Przyborowski, P. Biczyk, and D. Ślęzak, "Network Device Workload Prediction: A Data Mining Challenge at Knowledge Pit," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 21, 2020, pp. 77–80. [Online]. Available: https://doi.org/10.15439/2020F159

[7] A. Janusz and D. Ślęzak, "KnowledgePit Meets BrightBox: A Step Toward Insightful Investigation of the Results of Data Science Competitions," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, vol. 30, 2022, pp. 393–398. [Online]. Available: https://doi.org/10.15439/2022F309

[8] M. Wnuk, J. Dziuba, A. Janusz, and D. Ślęzak, "IEEE BigData Cup 2023 Report: Object Recognition with Muon Tomography Using Cosmic Rays," in *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, J. He, T. Palpanas, X. Hu, A. Cuzzocrea, D. Dou, D. Ślęzak, W. Wang, A. Gruca, J. C. Lin, and R. Agrawal, Eds. IEEE, 2023, pp. 6084–6091. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386564

[9] A. Janusz and D. Ślęzak, "Predicting Frags in Tactic Games at knowledgepit.ai: ICME 2023 Grand Challenge Report," in *IEEE International Conference on Multimedia and Expo Workshops, ICMEW Workshops 2023, Brisbane, Australia, July 10-14, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICMEW59549.2023.00006

[10] A. M. Rakicevic, P. D. Milosevic, I. T. Dragovic, A. M. Poledica, M. M. Zukanovic, A. Janusz, and D. Ślęzak, "Predicting Stock Trends Using Common Financial Indicators: A Summary of FedCSIS 2024 Data Science Challenge Held on knowledgepit.ai Platform," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems, FedCSIS 2024, Belgrade, Serbia, September 8-11, 2024*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39, 2024, pp. 731–737. [Online]. Available: https://doi.org/10.15439/2024F7912

[11] D. Ślęzak, A. Janusz, M. Świechowski, A. Chądzyńska-Krasowska, and J. Kamiński, "Do Data Scientists Dream About Their Skills' Assessment? – Transforming a Competition Platform Into an Assessment Platform," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8403–8414. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825378

[12] E. Hüllermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021. [Online]. Available: https://doi.org/10.1007/s10994-021-05946-3

[13] D. Kałuża, A. Janusz, and D. Ślęzak, "Evidence-theoretical Modeling of Uncertainty Induced by Posterior Probability Distributions," *International Journal of Applied Mathematics and Computer Science*, vol. 35, no. 1, 2025. [Online]. Available: https://doi.org/10.61822/amcs-2025-0003

[14] A. Janusz, A. Zalewska, Ł. Wawrowski, P. Biczyk, J. Ludziejewski, M. Sikora, and D. Ślęzak, "BrightBox – A Rough Set based Technology for Diagnosing Mistakes of Machine Learning Models," *Applied Soft Computing*, vol. 141, p. 110285, 2023. [Online]. Available: https://doi.org/10.1016/j.asoc.2023.110285

[15] A. Janusz, D. Kałuża, D. Ślęzak, and S. Stawicki, "Automatic Generation of Attributes based on Semantic Categorization of Large Datasets in Artificial Intelligence Models and Applications," *US Patent Application US-20250005436-A1*, 2025.

[16] R. McIlroy-Young, S. Sen, J. M. Kleinberg, and A. Anderson, "Aligning Superhuman AI with Human Behavior: Chess as a Model System," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 1677–1687. [Online]. Available: https://doi.org/10.1145/3394486.3403219

[17] S. Björkqvist, "Estimating the Difficulty of Chess Puzzles by Combining Fine-Tuned Maia-2 with Hand-Crafted and Engine Features," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F6497

[18] T. Woodruff, L. Imbing, and M. Cognetta, "The bread emoji Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F6771

[19] S. Miłosz, "Pretraining Transformers for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F7603

[20] H. Xiao, D. Yu, X. Wen, L. Chen, and K. Fu, "Multi-Source Feature Fusion and Neural Embedding for Predicting Chess Puzzle Difficulty," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F2456

[21] L. Cen, J. Cen, and M. Song, "A Multi-Stage Framework for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F4532

[22] A. Liang, C. Liu, K. Wang, and E. Liu, "A Stacking-Based Ensemble Approach for Predicting Chess Puzzle Difficulty," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F1698

[23] M. Liu, J. Wang, Y. Hu, and D. Lin, "Hybrid Boosting and Multi-Modal Fusion for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F3675

[24] J. Chen, C. Liu, and Y. Gao, "Multi-Modal Deep Learning with Residual and Structure-Guided Refinement for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F3227

Author Index

Aartsen, Wendy 13
Abbas, Musarat 765
Abromavičius, Vytautas 287
Akdağ-Ochnik, Kacper 615
Aldawoud, Raya 429
Almaghrabi, Samah 73
Alotaibi, Salem 441
Alshehri, Mona 451
Alshehri, Saleh Z 87
Angelinelli, Matteo 369
Araújo, Gabriele S. 461
Aslam, Sheraz 277
Asselborn, Thomas 471
Atiquzzaman, Mohammed 35

Beloff, Natalia 73, 87, 99, 451
Belousovienė, Egle 333
Bender, Magnus 471
Benecki, Pawel 255
Bergheim, Stein Runar 507
Bernardes, Cesar Freitas 635
Bičevska, Zane 479
Bičevskis, Jānis 479
Bineid, Maysoon 99
Björkqvist, Sebastian 801
Blachnik, Marcin 111
Blazevičs, Jānis Rihards 121
Blum, Christian 47
Boček, Lukáš 647
Böhm, Karsten 771
Bondevik, André Skoog 297
Bontchev, Boyan 489
Botek, Matúš 507
Botički, Ivica 739
Bouskila, Naama 641
Burchiellaro, Linda 499
Burger, Florian 175
Burk, Louis 531
Bylina, Beata 659

Canali, Claudia 499
Caprile, Francesco 399
Casari, Martina 369
Cassera, Alessandro 277
Castellucci, Pedro 635
Cen, Jiahao 807
Cen, Ling 807
Charvat, Karel 507
Chen, Junlin 813
Chen, Le 843
Chihaia, Robert 291
Chmielewski, Jakub 111
Chmiel, Janusz 303
Chu, Henry 265
Ciesiółka, Michał 665
Cognetta, Marco 837
Cornelis, Chris 195
Cortes, Omar Andres C. 461
Crevillen, Eduardo Garro 277
Cutolo, Arsenio 765

Damaševičius, Robertas 3, 339
Dan, Daniel 675
Dautov, Rustem 297
De Jesus Marques, Ruben Antonio 351
Dery, Lihi 641
Dimovski, Aleksandar S. 519
Dončević, Juraj 739
Dorau, Ralf 681
Dörpinghaus, Jens 687
Dubin, Ran 345, 699
Đukić, Marija 25
Dustdar, Schahram 1
Dutta, Soma 59, 133
Duzzioni, Eduardo 635
Dvir, Amit 345

Engels, Stefan 727
Erdogan, Tugba Gurgen 387

Faenza, Francesco 499
Fajčák, Branislav 327
Falco, Ivanoe De 605
Feltus, Christophe 351
Fidanova, Stefka 693
Fiksdal, Jan Robert 297
Fischer, Alexander 531
Francioso, Luca Nunzio 369
Freda, Arianna 543
Fronczyk, Mateusz 145
Frøshaug, Svein Arild 297
Fu, Kun 843

Gandomi, Amir H 321
Ganzha, Maria 693
Gao, Yujie 813
Gawęda, Bartłomiej 303
Georgiev, Slavi 777

Gerhards, Christian 653
Gerokonstantis, Dimitrios 357
Gilkarov, Daniel 699
Gonçalves, Douglas 635
Goszczyński, Wiktor 303
Grabek, Bartosz 675
Graliński, Filip 665
Grobler-Dębska, Katarzyna 303
Grużewski, Mateusz 753
Gürkahraman, Kali 309

Hachaj, Tomasz 405
Hadjichristofi, Christos 357
Hajaj, Chen 345
Hamersveld, Dicky van 13
Hamza, Ameer 3
Hanková, Michaela 315
Hapka, Aneta 615
Hein, Kristine 681
Herodotou, Herodotos 277
Holeňa, Martin 185
Horáková, Šárka 507
Huber, Melanie 595, 705
Hu, Yinghan 825

Imbing, Luke 837
Ip, Ching-Ho 243

Jankowski, Andrzej 59
Jíša, Nikolas 155
Jonker, Catholijn 13

Kaleta, Mariusz 165
Kammeyer, Alexander 175
Karakis, Rukiye 309
Kawulok, Michal 255
Khanina, Anastasia 99
Klos, Tim 705
Kolagari, Ramin Tavakoli 531
Kollerová, Markéta 507
Kordík, Pavel 507
Korel, Lukáš 185
Kostrzewa, Daniel 255
Kosturski, Nikola 715
Kovalenko, Alexander 507
Kowaleczko, Pawel 255
Krajna, Agneza 739
Krayani, Ali 399
Krištofík, Štefan 315
Krzywda, Maciej 321
Kubíček, Antonín 507

Lapidot, Omer 553
Laskov, Lasko 721
Leon, Florin 291
Leyh, Christian 243
Liang, Alan 819
Ligęza, Antoni 381
Lin, Defu 825
Lirkov, Ivan 715
Lisitsa, Alexei 441
Liu, Cenru 813
Liu, Cenzhi 819
Liu, Ethan 819
Liu, Ming 825
Lobato, Fabio M. F. 461
Lochschmidt, Severin 727
López, Mauricio Restrepo 195
Lorenz, Alisa 733
Lübbert, Daniel 175
Lukasik, Szymon 321
Luković, Ivan 25, 411

MacSween, Joanna 441
Maestosi, Davide 543
Maeyer, Wannes De 195
Maliha, Maisha 35
Malík, Peter 327
Manerba, Daniele 625
Mansini, Renata 625
Maqsood, Sarmad 333
Marcenaro, Lucio 399
Marinov, Marin 721
Martínez-Álvarez, Francisco 605
Martinez-Romero, Alvaro 277
Martinović, Filip 739
Masaracchia, Antonino 399
Maskeliūnas, Rytis 287, 333, 339
McCabe, Antony 441
Meiri, Itay 345
Melzer, Sylvia 471
Michaelides, Michalis 277
Mikulski, Łukasz 231
Miller, Gloria 207
Milošević, Pavle 411
Miłosz, Szymon 831
Misztal, Artur 747
Mlinarić, Danijel 739
Möller, Ralf 471
Mucherino, Antonio 635

Naeem, Muddasar 765
Nakayama, Minoru 675
Naldi, Maurizio 543
Nguy, Chan Duong 219
Nicolas, Damien 351
Nicosia, Gaia 543
Nikiforova, Oksana 121

Odītis, Ivo ... 479
Odītis, Reinis ... 479
Okoń, Maciej ... 659
Orlej, Martin ... 327
Orphanides, Kyriacos ... 277
Overberghe, Steven Van ... 195

Pacifici, Andrea ... 543
Palkowski, Marek ... 753
Palma, Raul ... 507
Pańkowska, Małgorzata ... 771
Papadeas, Panagiotis ... 357
Paprzycki, Marcin ... 715
Pascali, Chiara De ... 369
Pastor, Oscar ... 121
Pawlaczyk, Lesław ... 363
Pedrazzi, Giorgio ... 369
Pergl, Robert ... 155
Petrov, Mihai ... 777
Plank, Barbara ... 43
Po, Laura ... 369
Poledica, Ana ... 411
Preiß, Niklas ... 565
Prinz, Niculin ... 705
Przymus, Piotr ... 231

Quilliot, Alain ... 577

Radosavčević, Aleksa ... 411
Rahman, Md Mustafijur ... 587
Rawski, Mariusz ... 145
Regazzoni, Carlo ... 399
Reinhold, Olaf ... 461
Rentrop, Christopher ... 595, 705
Riedinger, Constanze ... 595, 705
Rippa, Massimo ... 327
Rollo, Federica ... 369
Rozbitski, Paul ... 243
Rudek, Radoslaw ... 375
Rykaczewski, Krzysztof ... 231
Rynkiewicz, Adam Aron ... 507

Sadel, Jakub ... 255
Sanden, Maarten van der ... 13
Sannino, Giovanna ... 605
Sapundzhi, Fatima ... 777
Sarkar, Twisampati ... 265
Sartori, Camilo Chacón ... 47
Sbaï, Zohra ... 429
Seif, Lubomír ... 759
Sepioło, Dominik ... 381
Shah, Mustafa ... 765
Shin, Jaewon ... 133
Sidje, Roger B. ... 587
Šimková, Bianka ... 417
Skowron, Andrzej ... 59
Ślęzak, Dominik ... 849
Ślęzak, Michał ... 849
Śluzek, Andrzej ... 783
Sołtysik-Piorunkiewicz, Anna ... 771
Song, Malin ... 807
Stachura, Piotr ... 783
Stanchov, Jelenko ... 693
Stanek, Stanisław ... 771
Stanojević, Dragan ... 25
Štěpánek, Lubomír ... 647, 759, 789, 795
Świechowski, Maciej ... 849
Szabo, Vera ... 297

Taczała, Michał ... 675
Tamir, Tami ... 553
Tan, Stephanie ... 13
Tarasiewicz, Tomasz ... 255
Tarhan, Ayça Kolukısa ... 393
Tasci, Oguzhan ... 387
Taşgetiren, Aslı ... 393
Tekulu, Micheale ... 399
Tiemann, Michael ... 687
Todorov, Venelin ... 777
Toussaint, Hélène ... 577
Traneva, Velichka ... 777
Tranev, Stoyan ... 777
Trocan, Maria ... 291
Tsolova, Sia ... 489
Turra, Roberta ... 369
Tverdal, Simeon ... 297

Ullah, Naeem ... 605

Vasileiadis, Sotiris ... 277
Vassileva, Dessislava ... 489
Vescoukis, Vassilios ... 357
Vít, Ondřej ... 759

Wałęga, Szymon ... 303
Wang, Junye ... 825
Wang, Kai ... 819
Waris, Muhammad ... 765
Wasilewski, Piotr ... 219
Wen, Xuegang ... 843
Westner, Markus ... 565
White, Martin ... 73, 87, 99, 451
Wienkop, Uwe ... 531
Wille, Robert ... 727
Wójcicka, Joanna ... 405
Wojszczyk, Rafał ... 615
Wolter, Katinka ... 175
Woodruff, Tyler ... 837
Wróblewska, Anna ... 675

Xiao, Haitao 843
Yang, Xiaolin 825
Yelkuvan, Ahmet Firat 309
Yu, Daiyuan 843
Yu, Zhuliang 807
Zanotti, Roberto 625
Zarcone, Alessandra 705
Ziaja, Maciej 255
Zieliński, Janusz 231
Zimmermann, Stephan 595
Zontone, Pamela 399
Zukanović, Milica 411
Zyśko, Jan 849
Żytniewski, Mariusz 771