

Annals of Computer Science and Information Systems

Volume 49

# Communication Papers of the 21st Conference on Computer Science and Intelligence Systems (FedCSIS)

August 23–26, 2026. Riga, Latvia



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





# Annals of Computer Science and Information Systems, Volume 49

## 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,

*Rzeszów 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*

# Communication Papers of the 21<sup>st</sup> Conference on Computer Science and Intelligence Systems (FedCSIS)

Marek Bolanowski, Maria Ganzha,  
Marek Grzegorowski, Leszek Maciaszek,  
Marcin Paprzycki, Andrzej Paszkiewicz,  
Dominik Ślęzak

Annals of Computer Science and Information Systems, Volume 49  
Communication Papers of the 21<sup>st</sup> Conference on Computer Science and  
Intelligence Systems (FedCSIS)

ISBN 978-83-977088-4-6

ISSN 2300-5963

DOI 10.15439/978-83-977088-4-6

© 2026, Polskie Towarzystwo Informatyczne  
Ul. Solec 38/103  
00-394 Warsaw  
Poland

**Contact:** sekretariat@fedcsis.org  
<http://annals-csis.org/>

**Cover art:**

Michał Ożarowski,  
*Elbląg, Poland*

**Also in this series:**

Volume 48: Position Papers of the 21<sup>st</sup> Conference on Computer Science and  
Intelligence Systems (FedCSIS), **ISBN 978-83-977088-3-9**

Volume 47: Proceedings of the 21<sup>st</sup> Conference on Computer Science and Intelligence  
Systems (FedCSIS), **ISBN WEB: 978-83-977088-1-5, ISBN ART 978-83-977088-2-2**

Volume 46: Proceedings of the International Conference on Engineering, Technology  
and Applied Science Innovations, **ISBN 978-83-977088-0-8**

Volume 45: Communication Papers of the 20<sup>th</sup> Conference on Computer Science and  
Intelligence Systems (FedCSIS), **ISBN 978-83-973291-9-5**

Volume 44: Position Papers of the 20<sup>th</sup> Conference on Computer Science and  
Intelligence Systems (FedCSIS), **ISBN 978-83-973291-8-8**

Volume 43: Proceedings of the 20<sup>th</sup> Conference on Computer Science and Intelligence  
Systems (FedCSIS), **ISBN WEB: 978-83-973291-6-4, ISBN ART 978-83-973291-7-1**

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 19<sup>th</sup> 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 19<sup>th</sup> 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 19<sup>th</sup> 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 18<sup>th</sup> Conference on Computer Science and  
Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**

# This is the preface to Communication Papers volume

**D**EAR Reader it is our pleasure to present to you the Communication Papers of the 21th Conference on Computer Science and Intelligence Systems (FedCSIS 2026), which took place on 23-26 August, 2026, in Riga, Latvia.

In the context of the FedCSIS conference series, the communication papers were introduced in 2017, as a separate category of contributions. They report on research topics worthy of immediate communication. They may be used to mark a new research territory, or to describe work in progress, in order to quickly present it to the scientific community. They may also contain additional information, omitted from the earlier papers, or may present software tools and products in a research state.

FedCSIS 2026 was chaired by Oksana Nikiforova. Moreover, Andrejs Romānovs was the Chair of the Organizing Committee. Furthermore, this year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, The Faculty of Mathematics and Information Science of the Warsaw University of Technology, The Faculty of Electrical and Computer Engineering of the Rzeszów University of Technology, The Faculty of Electrical Engineering at the Riga Technical University and The Faculty of Engineering and Management of the University of Technology and Arts in Applied Sciences.

FedCSIS 2026 was technically co-sponsored by IEEE Poland Section, IEEE Serbia and Montenegro Section, IEEE Estonia Section, IEEE Latvia Section, IEEE Lithuania Section, IEEE South Africa Section, Serbia and Montenegro Section Computational Intelligence Society Chapter, IEEE Young Professionals Affinity Group, Novi Sad, Republic of Serbia, IEEE Romania Section WIE Affinity Group, Czechoslovakia Section Computer Society Chapter, Poland Section Computer Society (Gdańsk) Chapter, Estonia Section Computer Society Chapter, Latvia Section Computer Society Chapter, Lithuania Section Computer Society Chapter.

FedCSIS 2026 was sponsored by InfoSoftware Polska as well as MDPI Information, MDPI Electronics, MDPI Technologies, MDPI Applied Sciences journals. Moreover, FedCSIS 2026 has been conducted under patronages of: INFORMS Poland Chapter, Latvian Information and communications technology association (LIKTA), Naukowe Towarzystwo Informatyki Ekonomicznej, International Rough Set Society, Web Intelligence Consortium, Polish Neural Network Society, European Society for Fuzzy Logic and Technology, Polish Artificial Intelligence Society (PSSI). Finally, media patronage was provided by Forum Akademickie.

During FedCSIS 2026 four keynote speakers delivered

lectures providing a broader context for the conference participants. These presentations were:

- Ginter, Filip, TurkuNLP, Department of Computing University of Turku, Finland and ELLIS Institute, Finland  
*Keynote title: The Challenges of Tracing Meaning at Scale across Multilingual Historical Texts*
- Pechenizkiy, Mykola, Department of Mathematics and Computer Science, Eindhoven University of Technology, Netherlands  
*Keynote title: Advancing the Science of Safe and Trustworthy AI Evaluation*
- Prodan, Radu, Edge AI at the Department of Computer Science, University of Innsbruck, Austria  
*Keynote title: Distribute and Learn 1 Billion*
- Sabou, Marta, Vienna University of Economics and Business (WU), Austria  
*Keynote title: Neuro-Symbolic Knowledge Engineering: Emerging Paradigms and Challenges*

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

- Atiquzzaman, Mohammed, School of Computer Science, University of Oklahoma, USA  
*EnsembleClassifier: Using Machine Learning to Select Exact Packet Classification Structures for Partitioned Rule Subsets*
- Blum, Christian, Artificial Intelligence Research Institute (IIIA-CSIC), Spain  
*Contribution title: Preserving Optimization Algorithm Expertise by means of Executable Algorithm Knowledge Graphs: A Worked Example on the TSP*
- Skowron, Andrzej, Systems Research Institute Polish Academy of Sciences, Poland  
*Contribution title: Evolution of Approximation Spaces in Rough Sets: From Relational Systems to Complex Adaptive Systems*
- van der Aalst, Wil M.P., RWTH Aachen University, Germany  
*Contribution title: Mining Bureaucracy Debt: Process Intelligence for Evidence-Based Deregulation*

At the time, when you are reading this text, videos of the keynote presentations and of invited contributions, delivered during the FedCSIS 2026 conference, are already available on the official conference website ([www.fedcsis.org](http://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.

This year, the structure of the conference remained the same as last year. FedCSIS 2026 had a single Main Track with 5 Topical Areas, which was supported by 16 Thematic Sessions and an Industrial Track. Such a 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 2026) 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.

Articles are ordered alphabetically by the last name of the first author. The specific Topical Area or Thematic Session that given contribution was associated with is listed in the article metadata. Each contribution, found in this volume, was refereed by at least two referees.

Delivery of FedCSIS 2026 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 2026 Senior Program Committee, and members of the FedCSIS 2026 Program Committee (a total of more than 300 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 Andrejs Romānovs, Megija Krista Milūne, Rihards Bobkovs, Uldis Karlovs-Karloviskis, Kristaps Babris, Veronika Tatjana Vucāne, Zane Bičevska 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 22nd Conference on Computer Science and Intelligence Systems (FedCSIS 2027), which will take place in Gdynia, Poland, on September 12-15, 2027.

*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*

**Grzegorowski, Marek**, *Samsung Research and University of Warsaw, Poland*

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

**Paprzycki, Marcin**, *Systems Research Institute Polish Academy of Sciences and University of Technology and Arts in Applied Sciences, Poland*

**Paszkiewicz, Andrzej**, *Rzeszów University of Technology, Poland*

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



# Communication Papers of the 21<sup>st</sup> Conference on Computer Science and Intelligence Systems (FedCSIS)

August 23–26, 2026. Riga, Latvia

---

## TABLE OF CONTENTS

---

---

### COMMUNICATION PAPERS

---

|                                                                                                                                                                        |           |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Phase-Dependent Drivers and Cumulative Institutional Impacts of Digital Transformation Adoption in Saudi Public Universities</b>                                    | <b>1</b>  |
| <i>Saleh Z Alshehri, Natalia Beloff, Martin White, Imran Khan</i>                                                                                                      |           |
| <b>A multithreaded Java implementation of Monte Carlo simulation for Value-at-Risk estimation</b>                                                                      | <b>11</b> |
| <i>Paweł Borowiecki, Beata Bylina</i>                                                                                                                                  |           |
| <b>INCLUDE-AI and G.I.U.L.I.A.: A Multi-Level European and Territorial Framework for Inclusive, Trustworthy, and Gender-Aware Artificial Intelligence in Education</b> | <b>17</b> |
| <i>Matteo Ciaschi, Daniel Dan</i>                                                                                                                                      |           |
| <b>Predicting Centrality Measures in Complex Networks Utilizing Markov Chains</b>                                                                                      | <b>25</b> |
| <i>Anchal Gera, Jens Dörpinghaus, Robert Rockenfeller</i>                                                                                                              |           |
| <b>Potential of Found Establishment Data Linkage for the Social Sciences</b>                                                                                           | <b>37</b> |
| <i>Christian Gerhards</i>                                                                                                                                              |           |
| <b>A Holistic CPS for Smart Precision Livestock Farming: Multi-Actor Collaborative Decision-Making Across Ecosystem Domains</b>                                        | <b>45</b> |
| <i>Sonia Hajri-Gabouj, Imen Harbaoui, Rabaa Youssef</i>                                                                                                                |           |
| <b>Semantic Matching of IT Vocational Training Offers with the ESCO Ontology</b>                                                                                       | <b>53</b> |
| <i>Katerina Kostadinovska, Kristine Hein</i>                                                                                                                           |           |
| <b>Incorporation of Min/Max, With/Without and Pairing Restrictions in Column Generation Solution Algorithms for Aircrew Rostering</b>                                  | <b>59</b> |
| <i>George Kozanidis, Nikos Polychronopoulos, Andreas Gavranis</i>                                                                                                      |           |
| <b>Software System for Optimization of Metallic Nanostructures</b>                                                                                                     | <b>67</b> |
| <i>Rossen Mikhov, Vladimir Myasnichenko, Leoneed Kirilov, Nikolay Sdobnyakov, Stefka Fidanova</i>                                                                      |           |
| <b>A Reproducible Zero Trust Testbed for Dynamic Trust Evaluation Using Emulation and Virtualization Tools</b>                                                         | <b>73</b> |
| <i>Tiberius Nkinyili, Vincent Omwenga, Solomon Ogara</i>                                                                                                               |           |
| <b>Cosine Colorization for Enhancement of Magnetic Resonance Angiography (MRA)</b>                                                                                     | <b>81</b> |
| <i>Modupe Odusami, Robertas Damaševičius, Olusola Abayomi-Alli, Rytis Maskeliūnas</i>                                                                                  |           |
| <b>Data collection, aggregation, and analysis in the Internet of Everything environment – a case study</b>                                                             | <b>89</b> |
| <i>Andrzej Paszkiewicz, Andrzej Gołębek, Piotr Sowa, Mateusz Kozik, Marek Bolanowski</i>                                                                               |           |
| <b>Parsing Hierarchical Document Structure with Graph Neural Networks</b>                                                                                              | <b>93</b> |
| <i>Thomas Reiser, Veronika Vasileva, Volker Riediger, Alexander Esser, Petra Steiner, Jan Jürjens</i>                                                                  |           |

|                                                                                                                         |            |
|-------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Still Waiting for the Shock: AI's Limited Impact on Early-Career Vacancies, Skills and Tasks</b>                     | <b>101</b> |
| <i>Stefan Speckesser, Lei Xu</i>                                                                                        |            |
| <b>Iterative Inductive Learning for Generalized Modeling of Technical Systems</b>                                       | <b>111</b> |
| <i>Balázs Ádám Toldi, András Földvári, Zsigmond Pap, András Pataricza</i>                                               |            |
| <b>Acceptance of Generative AI in Higher Education in Software Engineering Education—A Systematic Literature Review</b> | <b>119</b> |
| <i>Veronika Vasileva, Thomas Reiser, Jan Jürjens</i>                                                                    |            |
| <b>Author Index</b>                                                                                                     | <b>127</b> |

# Phase-Dependent Drivers and Cumulative Institutional Impacts of Digital Transformation Adoption in Saudi Public Universities

Saleh Z Alshehri  
0000-0002-0702-3382  
dept. Computer Science and Information Systems  
King Khalid University  
Abha Saudi Arabia  
dept. of Informatics  
University of Sussex  
Brighton, United Kingdom  
Sa2142@sussex.ac.uk

Natalia Beloff, Imran Khan, and Martin White  
0000-0002-8872-7786  
0000-0002-5732-4685  
0000-0001-8686-2274  
dept. of Informatics  
University of Sussex  
Brighton, United Kingdom  
{N.Beloff,Imran.Khan,M.White}@sussex.ac.uk

**Abstract**—Digital transformation (DX) adoption in Saudi public universities is shaped by centralised governance and national policy, notably Vision 2030, which existing Western models often fail to capture. This study examines and extends the Digital Transformation Adoption framework (DXA-SPU) through a sequential explanatory mixed-methods design, combining structural equation modelling of survey data with thematic analysis of interviews with senior IT leaders. Results reveal that adoption is governed by Phase-Dependent Drivers: national mandates and Strategy trigger the initial decision, while Collaboration and Coordination facilitate implementation. We further identify Cumulative Institutional Impacts, in which operational gains in Infrastructure and Efficiency precede later improvements in Teaching and Institutional Performance. Theoretically, the study reconceptualises governance structure as the determinant of factor timing; methodologically, it shows how explanatory mixed-methods resolve non-significant survey paths; practically, it offers university leaders a staged template for managing digital change. As a cross-sectional, single-country study, it invites longitudinal and comparative validation.

**Index Terms**—Digital transformation, Institutional impacts, Phase-Dependent Drivers, Cumulative Impacts, Mixed-Methods Research, Saudi public universities, TAM, TOE, Vision 2030, Institutional Performance, Thematic analysis.

## I. INTRODUCTION

DIGITAL transformation (DX) is now a strategic issue in higher education. It is not only about new tools and technologies, but also about changes in how universities plan, deliver services, and organise work [1, 2, 3]. In Saudi Arabia, this shift is tied to Vision 2030 and the National Transformation Program [4, 5], which aim to modernise public services and strengthen the knowledge economy [6, 7]. Saudi public universities (SPUs) are central to that agenda, yet many still face challenges in readiness, governance, and strategic alignment despite major investment in digital infrastructure [8, 9].

Much of the literature on DX adoption in higher education draws on models developed in Western or private-sector set-

tings [10, 11]. Although useful, such work does not fully reflect the conditions in public universities operating under centralised governance and national policy direction. In the Saudi case, adoption decisions are shaped simultaneously by user perceptions, institutional hierarchy, and state regulation. This mix has received limited attention in the relevant literature.

The DXA-SPU framework was developed for this decision-making setting by combining the Technology Acceptance Model (TAM) and the Technology-Organisation-Environment (TOE) framework [11, 12]. Our earlier work tested this framework with survey data from 447 participants [13]. While 12 of the 14 hypotheses were supported, quantitative methods are inherently limited to identifying statistical associations. It cannot explain the institutional processes or the logic behind adoption decisions. Our earlier work also did not explore whether this non-significance reflected the faculty-dominated nature of its sample.

This paper addresses these gaps by introducing the concepts of Phase-Dependent Drivers and Cumulative Impacts. We demonstrate that the importance of adoption factors is not static throughout the project lifecycle. Instead, drivers such as collaboration shift from being non-significant during the initial decision phase to becoming critical during implementation.

This paper differs from our earlier work [13] in three concrete ways. First, it contributes new primary data: whereas [13] analysed survey responses from 447 participants (67.3% academic faculty), this study draws on original interviews with 11 senior IT leaders who direct institution-level adoption, a population absent from the prior study. Second, it advances new theoretical constructs, Phase-Dependent Drivers and Cumulative Institutional Impacts, that were neither proposed nor testable in the cross-sectional quantitative design. Third, it makes an explanatory contribution that the prior paper could not explicitly make at the statistical analysis stage. It resolves the two statistically non-significant survey paths (Collaboration and Coordination, Competitive Pressure) by showing they are not weak factors but factors that act at a later

project phase. In short, [13] established which factors correlate with adoption; this paper explains how and why they operate, and when each becomes influential.

The current study presents the qualitative phase of the research design. We interviewed top-level decision-makers to explain the mechanisms behind our previous statistical results. This study provides a dynamic framework for digital transformation in centralised governance settings.

Section II reviews related work and the framework. Section III outlines the method. Section IV presents the findings, recapping the quantitative results before the qualitative analysis. Section V discusses both strands together, and Section VI concludes.

## II. LITERATURE REVIEW

Research on digital transformation in higher education has grown rapidly, but most studies still rely on surveys, structural models, or secondary data to identify factors influencing adoption [10]. Far fewer studies examine how adoption decisions are made inside universities. Where qualitative work exists, it often focuses on a single tool, such as an e-learning platform or learning management system, rather than a university-wide DX [14, 15]. As a result, existing research provides limited insight into the decision processes and institutional choices that underlie the statistical findings reported in quantitative studies.

Governance structure is central to understanding adoption in this context. Institutional perspectives on technology adoption have long recognised that in public sector organisations, governance arrangements shape how technologies are adopted and enacted in practice. Fountain's technology enactment framework argues that institutional arrangements, including cognitive, cultural, and legal structures, determine how objective technologies become enacted differently across organisations [16]. Dunleavy's Digital Era Governance approach similarly argues that in public institutions, organisational structure and process change drive technology adoption, not the other way around [17]. TAM [12] and TOE [11], two widely applied adoption frameworks, were developed and tested in settings where organisations have broad independence in technology decisions. In that literature, collaboration and competitive pressure are often treated as direct drivers of adoption [18, 19]. Saudi public universities operate differently: Vision 2030 and the National Transformation Program shape strategy, funding, and performance expectations [6]. In this context, adoption decisions are influenced by compliance and top-down direction as much as by internal agreement or peer competition.

Studies in Saudi Arabia have examined attitudes to DX [8], requirements for transformation [9], organisational routines [20], data governance [21], the effect of COVID-19 on digital adoption [22], and the role of digital learning in sustainable development [23]. These studies are valuable, but most focus on one part of the problem at a time. They do not provide an integrated account of how organisational, technological, and environmental factors interact in DX adoption, nor do they

consider the impacts of DX. Our paper provides qualitative evidence on how these factors operate in practice and why some expected relationships do not hold in the Saudi governance setting.

Survey-based studies identify which factors correlate, but provide less insight into how decisions are made and why a factor that is salient in practice may not emerge as a direct statistical predictor [24]. For that reason, qualitative evidence is especially valuable when a study seeks to explain non-significant paths in an adoption model rather than confirm expected ones [24].

In this study of SPUs, where national programmes, ministerial direction, and institutional hierarchy shape major investment decisions, adoption cannot be understood solely in terms of individual perceptions. The decision to launch a digital initiative may be made at the top, while the coordination, local adjustments, and user acceptance occur later. Our study addresses this gap by showing how the same factor can have different effects at different stages of the adoption process.

Therefore, this study positions the qualitative phase as an analytical bridge rather than a supplement to the survey. As argued by [24], this approach uses qualitative data to explain and elaborate initial quantitative patterns, interpreting factors that are highly visible in practice yet emerged as weak statistical predictors.

### A. The DXA-SPU Framework

The DXA-SPU framework combines TAM and TOE to examine DX adoption at the institutional level in SPUs [13]. TAM contributes to perceived ease of use and perceived usefulness [12]. TOE captures organisational, technological, and environmental conditions that shape adoption in public institutions. The framework includes 10 factors related to the intention to adopt DX (H1-H10) and 4 institutional outcome constructs (H11-H14). Full framework development is reported in [13].

The organisational context includes Effective Strategy (ES), Collaboration and Coordination (CC), and Funding and Resources (FR). The technological context covers Complexity (CX), Security (SC), Ease of Use (EU), and Usefulness (UF). The environmental context captures Regulatory Environment (RE), Competitive Pressure (CP), and Digital Upskilling (DU). The outcome side includes Institutional Performance (IP), Administrative Efficiency (AE), Technical Infrastructure (TI), and Teaching and Learning Effectiveness (TLE).

## III. STUDY METHODOLOGY

This study employed a sequential explanatory mixed-methods design to bridge the gap between statistical outcomes and institutional reality [24]. The primary rationale for this design was to provide a deeper explanation of the "how" and "why" behind the results of the initial quantitative phase ( $n=447$ ) [13]. Specifically, while the survey confirmed the majority of the hypothesised relationships, it could not explain why Collaboration and Coordination and Competitive Pressure were

non-significant in the Saudi context. By following the survey with qualitative interviews, this study explores whether these factors are truly secondary or if their influence is Phase-Dependent, shifting as a project moves from a top-down mandate to operational implementation.

#### Participant Selection and Sampling Logic

Purposive sampling was used to recruit 11 senior IT leaders across various Saudi Public Universities between June and September 2025. Purposive sampling is appropriate here because the study targets a specifically informed population, those who direct institution-level adoption, rather than a statistically representative one; this elite-informant logic is standard for the explanatory phase of a mixed-methods design. These participants were selected because they occupy a critical governance role, translating national strategic mandates into practical technology adoption. This selection also addresses a demographic limitation in the previous study phase, which relied heavily on academic faculty (67.3% of respondents). Unlike faculty, senior IT leaders are directly involved in system rollouts and manage the technical complexities that define the institutional adoption process. Each participant was required to have led a major digital transformation project within the last three years, ensuring the data reflects current institutional realities under the Saudi Vision 2030 framework.

#### Data Collection and Ethical Procedures

The qualitative data were gathered through semi-structured interviews lasting 45 to 60 minutes. These sessions were conducted either in person or via Zoom, recorded, and transcribed verbatim for analysis. The 39-question interview guide was mapped directly to the DXA-SPU framework constructs, such as Perceived Usefulness, Security, and Complexity, distributing roughly two to three open questions across each of the fifteen constructs rather than over-burdening any single theme. This alignment allowed the qualitative evidence to be correlated directly with the survey variables while still permitting an open-ended exploration of unexpected results.

Ethical approval for this study was obtained from the University of Sussex Sciences and Technology Cross-Schools Research Ethics Committee (ref ER/SA2142/3). All participants provided informed consent and were de-identified using codes P1–P11 to ensure institutional confidentiality.

#### Thematic Analysis and Trustworthiness

To ensure a rigorous analytical process, the interview data were processed through a multi-stage thematic analysis in NVivo [25]. The coding strategy was both deductive and inductive: an *a priori* codebook ensured alignment with the DXA-SPU constructs, while open coding allowed for the emergence of "temporal sequence" and "cumulative impact" patterns. To enhance interpretative consistency, a peer-checking process was implemented where a second researcher reviewed a subset of the transcripts. Thematic saturation was assessed iteratively throughout the process. The later interviews confirmed and elaborated on patterns already established in the earlier ones rather than introducing new thematic

categories, indicating that saturation had been reached within the sample.

## IV. FINDINGS

This section summarises the quantitative results [13]. Table I reproduces the structural model results from that prior work; these values are not re-estimated here. They are presented solely to make the sequential design transparent and to identify the patterns, particularly the non-significant paths for Collaboration and Coordination (H2) and Competitive Pressure (H9), that the qualitative phase was designed to explain.

Eight out of ten hypotheses related to DX Adoption were supported. Perceived Usefulness had the largest effect ( $\beta = .272$ ,  $p < 0.001$ ), followed by Digital Upskilling ( $\beta = .239$ ,  $p < 0.001$ ), Funding and Resources ( $\beta = .184$ ,  $p = 0.002$ ), Complexity ( $\beta = .181$ ,  $p < 0.001$ ), Security ( $\beta = .131$ ,  $p = 0.003$ ), Effective Strategy ( $\beta = .121$ ,  $p = 0.001$ ), Ease of Use ( $\beta = .103$ ,  $p = 0.047$ ), and Regulatory Environment ( $\beta = .091$ ,  $p = 0.009$ ). Two hypotheses were not supported: Collaboration and Coordination (H2:  $\beta = -0.009$ ,  $p = .773$ ) and Competitive Pressure (H9:  $\beta = .100$ ,  $p = 0.068$ ). All four Impact hypotheses (H11 to H14) were strongly supported, with Administrative Efficiency showing the largest coefficient ( $\beta = .753$ ,  $p < 0.001$ ), followed by Technical Infrastructure ( $\beta = .683$ ,  $p < 0.001$ ), Teaching and Learning Effectiveness ( $\beta = .639$ ,  $p < 0.001$ ), and Institutional Performance ( $\beta = .629$ ,  $p < 0.001$ ). Table I summarises these results.

TABLE I.  
SUMMARY OF HYPOTHESIS TESTING RESULTS

| H   | Path                | $\beta$ | S.E. | C.R.   | P    | Result        |
|-----|---------------------|---------|------|--------|------|---------------|
| H1  | DX $\leftarrow$ ES  | .121    | .037 | 3.256  | .001 | Supported     |
| H2  | DX $\leftarrow$ CC  | -.009   | .031 | -.288  | .773 | Not Supported |
| H3  | DX $\leftarrow$ FR  | .184    | .059 | 3.099  | .002 | Supported     |
| H4  | DX $\leftarrow$ CX  | .181    | .054 | 3.364  | ***  | Supported     |
| H5  | DX $\leftarrow$ SC  | .131    | .044 | 2.995  | .003 | Supported     |
| H6  | DX $\leftarrow$ EU  | .103    | .052 | 1.983  | .047 | Supported     |
| H7  | DX $\leftarrow$ UF  | .272    | .049 | 5.539  | ***  | Supported     |
| H8  | DX $\leftarrow$ RE  | .091    | .035 | 2.613  | .009 | Supported     |
| H9  | DX $\leftarrow$ CP  | .100    | .055 | 1.822  | .068 | Not Supported |
| H10 | DX $\leftarrow$ DU  | .239    | .052 | 4.601  | ***  | Supported     |
| H11 | IP $\leftarrow$ DX  | .629    | .043 | 14.491 | ***  | Supported     |
| H12 | AE $\leftarrow$ DX  | .753    | .047 | 16.004 | ***  | Supported     |
| H13 | TI $\leftarrow$ DX  | .683    | .046 | 14.990 | ***  | Supported     |
| H14 | TLE $\leftarrow$ DX | .639    | .049 | 13.165 | ***  | Supported     |

Note: \*\*\* = significance at the 0.001 level. Source: [13].

The non-significance of Collaboration and Coordination and Competitive Pressure is notable, as our interviewees identified both as essential drivers in practice, a visibility also reported in [20, 26]. Therefore, the qualitative analysis addresses this gap by examining decision-making processes, implementation conditions, and participants' interpretations of DX's institutional impacts.

This section presents the findings from the thematic analysis of interviews with 11 senior IT leaders across Saudi public universities. The analysis was framework-guided, using the DXA-SPU constructs to organise and interpret the interview evidence. At the same time, broader patterns emerged from participants' interpretations of how digital transformation is decided, implemented, and experienced in their institutions.

Thematic analysis of the interviews produced four themes: (1) Institutional Alignment and Support, (2) Trusted and Effortless Experience, (3) External Mandates and Pressures, and (4) Operational and Educational Gains.

#### A. Institutional Alignment and Support

Interviewees described digital transformation as gaining momentum when it was tied to institutional priorities and national goals. For participants, Strategy functioned as the primary mechanism for translating national objectives into specific projects. For instance, P5 noted that Strategy supported data-driven decision-making and a digital culture, while P10 viewed it as a path to leadership in e-learning. In these accounts, Strategy was more than a formal statement; it was the tool used to turn national goals into institutional priorities: *"Our digital transformation strategy is not just about technology; it is deeply connected with the national goals of Saudi Vision 2030. We are focused on using technology to create tangible benefits for our society."* (P1)

When asked why Collaboration and Coordination did not emerge as a significant driver in the earlier survey, P1 explained: *"This perception is likely because the core decision to start a digital transformation project is a strategic, top-down decision from our top leadership. This is especially true when it is connected to a national plan like Saudi Vision 2030."* (P1)

The practical value of Collaboration and Coordination is most significant after a project has moved from approval to execution. This distinction between strategic adoption and operational implementation is reinforced by how universities manage Collaboration and Coordination, as stated by P1. While coordination through steering committees and routine feedback channels, such as Blackboard and internal service platforms, is evident in daily service improvements, it rarely influences the initial decision to adopt a major digital initiative. For instance, P11 described a campus-wide scheduling project, in which disagreements were managed through centralised tracking and early stakeholder engagement rather than broad consultation. When internal disagreements over priorities occur, institutions rely on such central project management to ensure execution without reopening the adoption decision. Consequently, Collaboration and Coordination function as an implementation-stage facilitator rather than a strategic trigger, explaining the non-significance of H2.

Funding was treated as essential, but mostly as a question of pace and scale. P3 explained how endowment companies add to government funding, including a programme that offers up to 50,000 riyals for medium-sized research projects.

P4 described cost-control strategies, such as outsourcing non-critical services. Interviewees did not present funding as the sole factor determining whether DX would be adopted. They described it as the factor that shaped what could be done, how fast, and at what scale.

#### B. Trusted and Effortless Experience

The technological side of DX was described in practical terms. Interviewees focused on trust, ease of use, and clear value. They also returned repeatedly to one problem: new systems must work with old ones.

Legacy systems integration was the most common implementation problem. P1 said, *"The primary difficulty is integrating new, modern tools with our existing legacy platforms."* P5 linked poor compatibility to delays, errors, and lower confidence in new systems. P2 added that upgrades require careful planning so that integration and security are managed together. Universities responded with pilot testing, phased rollouts, and close work with vendors. P11 gave a concrete example: an AI-based attendance system failed to sync with biometric devices, so the university restored manual tracking while the vendor developed a patch. This episode showed the importance of pre-deployment testing and backup procedures.

Security was treated as a threshold issue that participants (P1, P9) directly linked to institutional trust and operational reliability. Beyond technical protocols, such as pre-approval security checks and incident response procedures, the interviews revealed that Security is intrinsically connected to Complexity. This is demonstrated by the practice of pausing deployments or reverting to stable versions to prevent service disruption when integration fails. Consequently, in a centralised environment, Security is not merely an isolated technical requirement but a continuous process of managing the Complexity inherent in integrated systems.

Ease of use was discussed less as convenience and more as inclusion. P3 stressed user-friendly design and guides for systems such as Blackboard and a university communication platform, adding, *"Ease of use is crucial."* P10 said the goal was to make systems simple enough for everyone to benefit. When systems are hard to use, less confident staff often return to manual work, which limits wider adoption.

Usefulness was the clearest reason for support. P5 said, *"I think they are extremely useful. Our research shows a direct and positive impact. For efficiency, data-driven decision-making has the highest correlation with continuous improvement."* Participants linked usefulness to things they had seen directly, such as faster administration and better teaching decisions. This underpins Perceived Usefulness (H7) as the strongest predictor of adoption intention.

#### C. External Mandates and Pressures

Environmental conditions also mattered, but not in the same way. Interviewees spoke about regulation, competition, and skills, each working at a different point in the process.

The Regulatory Environment was described less as a checklist and more as a basis for institutional confidence and

legitimacy. Alignment with Saudi Vision 2030, the Digital Government Authority, and national cybersecurity standards (P4, P9) allows senior leaders to justify investment and defend decisions internally. In this sense, regulation does not constrain action; instead, it reduces uncertainty by providing a formal basis for institutional and political alignment. This explains why Regulatory Environment emerged as a significant driver.

In contrast, Competitive Pressure was described as a secondary influence rather than a primary trigger for adoption. While participants like P7 noted it as a factor in specific AI and data science investments, most institutions monitor peers primarily through rankings and networks to shape their positioning once a project is underway. While the two forces are present simultaneously, they operate at different points: Regulatory Environment triggers the initial strategic decision, whereas Competitive Pressure influences how universities advance their endeavours.

While Digital Upskilling emerged as a strong predictor in the survey, the interviews show it is a manageable impediment rather than a constant barrier. Participants (P2, P6) described how universities mitigate this gap through key user testing, targeted workshops, and phased platform introductions. This clarifies the quantitative significance of Digital Upskilling: the primary challenge was not an inherent lack of capacity but rather weak support structures that hindered the pace of adoption. When these support systems are visible and integrated into routine practices, adoption accelerates.

#### D. Operational and Educational Gains

When discussing systems already in use, interviewees spoke less about adoption and more about what they believed had changed. Two patterns emerged from the interviews. First, participants often described outcomes as becoming easier to observe once systems had stabilised and routine use had increased. Second, they indicated that gains in one area could support gains in others.

Institutional Performance was framed as the cumulative, symbolic outcome of Digital Transformation, shifting the focus from internal efficiency to institutional reputation. Participants (P1, P5) indicated that while automation and infrastructure investment drive service quality, the broader value of Institutional Performance lies in building trust and signalling capability to national stakeholders. By reducing service interruptions and enhancing transparency, universities demonstrate that they are modern, responsive, and well-governed. This underpins H11: Institutional Performance aggregates administrative and technical gains into a broader signal of institutional standing.

Administrative Efficiency (H12) was identified as the most immediate and visible outcome of digital transformation. Participants provided concrete evidence of improved systems that resolved long-standing operational bottlenecks. One notable example provided by P11 illustrates this impact: *"One notable example is the implementation of a unified student*

*portal, which centralises academic records, course registration, and fee payments. This reduced student complaints by 40% and minimised administrative bottlenecks."* (P11)

Beyond the student portal, participants (P5, P6) noted that cloud platforms and infrastructure upgrades significantly raised service quality in libraries and administrative units. In the interviews, Technical Infrastructure emerged as the foundational layer on which the other three impacts rest, rather than as one impact among equals; P5 noted that network and server investment "creates the foundation for all of our digital transformation projects to succeed." Administrative Efficiency is the first gain to become visible; automation affects daily operations immediately, but it rests on the infrastructure foundation rather than preceding it.

Technical Infrastructure (H13) was described as both a starting point and a reinforcing outcome of Digital Transformation. Participants (P5, P8) noted a cyclical relationship in which adoption programmes justify investment in networks, which in turn lower the cost and difficulty of subsequent innovation. For example, the transition to AI-based tools and virtual classrooms on Blackboard Ultra (P8) was only possible because of prior investments in bandwidth and server nodes. This suggests that Technical Infrastructure is a cumulative asset that builds institutional capacity for future digital shifts.

Teaching and Learning Effectiveness (H14) represents a broad academic shift that integrates digital tools into the core curriculum. Participants (P6-8, P11) highlighted that the value of Digital Transformation in this area extends beyond simply moving content online. It involves providing students with flexible access to materials through virtual labs and aligning programmes more closely with labour market expectations in fields like AI and data science. In these accounts, AI-based plagiarism tools and blended delivery models (P11) strengthen academic integrity and student engagement.

The qualitative evidence suggests that these four outcome areas do not become visible at the same speed. Technical Infrastructure forms the foundational layer. Administrative Efficiency is typically the first gain to become *visible* because automation affects daily operations immediately. Teaching and Learning Effectiveness takes longer as it depends on changes in staff practice and student uptake. Finally, Institutional Performance serves as the cumulative result of these prior successes. While the survey data showed support for H11 through H14 simultaneously, the real-world benefits are interpreted as accumulating progressively over time.

In summary, these qualitative findings provide a multi-layered view of Digital Transformation that extends beyond the statistical associations reported in [13]. The interview data capture the transition from strategic triggers, such as national mandates, to the practical realities of infrastructure and cumulative impact. By providing the institutional logic behind the quantitative patterns, these operational experiences establish the evidentiary basis for the Phase-Dependent DXA-SPU framework discussed in the following section.

## V. DISCUSSION

To provide a comprehensive examination of digital transformation in Saudi public universities, this section integrates the structural model results from [13] with qualitative insights. The analysis employs a phase-dependent perspective to clarify how the influence of adoption drivers shifts across the institutional lifecycle. Moving beyond isolated statistical patterns, this synthesis explains the underlying organisational mechanisms and their longitudinal institutional impacts. The updated DXA-SPU framework, reflecting these findings, is presented in Fig. 1.

### A. Organisational Context

In the centralised Saudi governance model, Strategy and Funding serve as the primary organisational drivers of adoption. The strong statistical support for Strategy (H1) ( $\beta = .121$ ,  $p = .001$ ) confirms that a clear roadmap is essential for building institutional commitment. Qualitative evidence clarifies that Strategy is not merely a symbolic policy but a mechanism for translating national goals, such as Saudi Vision 2030, into sequenced institutional projects. This aligns with [27], which argues that strategic alignment is a prerequisite for bridging the gap between high-level policy and practical modernisation. Similarly, Funding (H3) ( $\beta = .184$ ,  $p = .002$ ) acts as the critical "safety net" for technical risk-taking. With the Saudi higher education market projected to reach \$317 million by 2025 [27], financial readiness remains the most significant precursor to large-scale action.

In contrast, the rejection of Collaboration and Coordination (H2) ( $\beta = -0.009$ ,  $p = .773$ ) highlights the "top-down" nature of this environment. While inter-departmental cooperation is highly valued during implementation, it does not drive the initial decision to adopt. As noted by [28], hierarchical legacies in Saudi institutions often create friction that delays collective efforts, meaning momentum is generated by mandate rather than horizontal consensus. These results indicate that Digital

Transformation adoption in Saudi public universities is governed by hierarchical triggers, specifically Strategy and Funding, while Collaboration and Coordination functions as an implementation-stage facilitator rather than a strategic trigger. This confirms that the institutional governance structure dictates the timing of factor influence, with top-down mandates preceding bottom-up coordination, providing a necessary correction to Western models that often assume collaborative drivers are present from the beginning.

### B. Technological Context

Technological adoption depends on a delicate balance between the perceived value of a tool and the technical barriers to its entry. Perceived Usefulness (H7) ( $\beta = .272$ ,  $p < 0.001$ ) emerged as the strongest predictor in the model, as users view digital tools as fundamental utilities for institutional readiness in an increasingly AI-literate global landscape [29]. However, this utility is contingent upon the mitigation of technical risks, specifically Complexity (H4) ( $\beta = .181$ ,  $p < 0.001$ ) and Security (H5) ( $\beta = .131$ ,  $p = .003$ ). Participants noted that integrating legacy systems with cloud platforms remains a primary hurdle, a perspective supported by [30], which identifies the simplification of traditional constraints as a requirement for institutional agility.

Security has transitioned from a technical requirement into a "threshold issue" for institutional trust. Alignment with national personal data protection laws is now critical to the user trust required for university-wide e-governance adoption [31]. While Perceived Ease of Use (H6) ( $\beta = .103$ ,  $p = .047$ ) was a significant predictor, its modest impact indicates it is an "enabler" rather than a primary driver. As argued in [30], while user-friendly design and training are essential for driving digital performance, users are willing to navigate greater complexity if the system's functional value and security are clearly established.

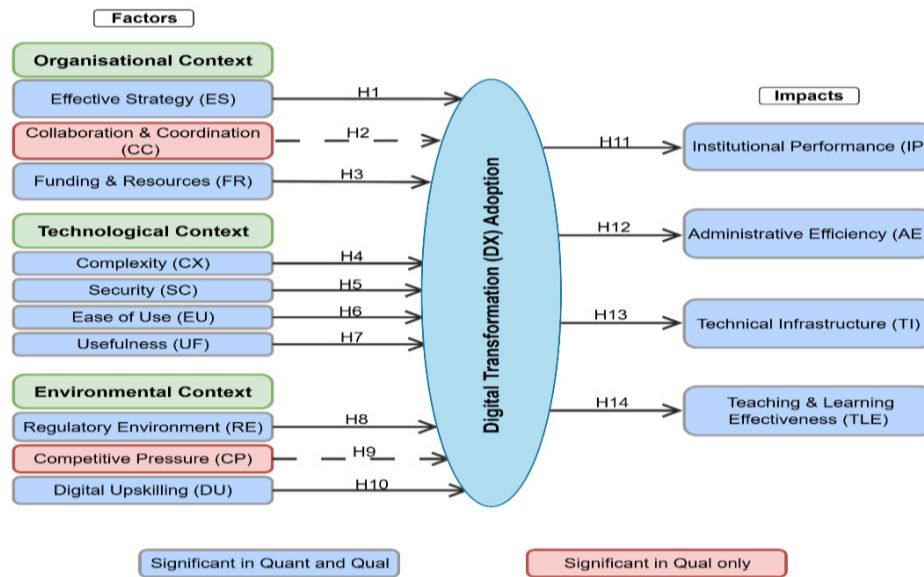


Fig. 1. The Validated and Updated DXA-SPU Framework

These findings suggest that technological adoption depends on a balance between perceived value (UF) and the mitigation of technical risk (CX, SC). Strategic commitment is highest when operational benefits are clear and technical barriers to entry are minimised.

### C. Environmental Context

The Regulatory Environment (H8) ( $\beta = .091, p = .009$ ) provides the legal "backbone" of the framework, with the National Transformation Program (NTP) serving as the enabling force that legitimises digital initiatives. Recent 2025 legislation and support policies specifically designed to stimulate innovation within public institutions [32] provide the confidence universities need to pursue modernisation. This mandate-driven logic clarifies why Competitive Pressure (H9) ( $\beta = .100, p = 0.068$ ) was not statistically significant. In a centralised system, market rivalry primarily influences institutional positioning and branding after the infrastructure is established, rather than triggering the core decision to adopt [26].

Furthermore, the strong support for the Digital Upskilling (H10) ( $\beta = .239, p < 0.001$ ) indicates that adoption is fundamentally motivated by the desire to align educational outputs with labour market needs. The strategic priorities identified in [33] emphasise that Saudi universities are under increasing pressure to align digital transformation with student employability and the human capability requirements of Vision 2030. These findings reveal that environmental drivers are not merely external pressures but are strategic imperatives tied to national socioeconomic goals.

### D. Institutional Outcome Impacts

The findings are compatible with a provisional cumulative sequence rather than a set of parallel benefits. Because the design is cross-sectional, this ordering rests on the convergence of the qualitative accounts and the pattern of effect magnitudes and is advanced as an interpretation rather than a tested causal sequence. Foundational operational successes create the conditions for subsequent academic and reputational shifts.

#### Foundational Gains: Technical Infrastructure and Administrative Efficiency

The chain begins with Technical Infrastructure (H13:  $\beta = .683, p < 0.001$ ), the foundation on which the other impacts depend; Administrative Efficiency (H12:  $\beta = .753, p < 0.001$ ) is the first and largest gain built on that foundation. Although Administrative Efficiency carries the higher path coefficient, Technical Infrastructure remains foundational—a foundation is judged by what it enables, not by the size of its own association with adoption. These results suggest that the primary "quick wins" of adoption are located in the university's back-office operations.

By automating routine tasks and reducing manual errors, digital tools have considerably improved administrative performance across Saudi institutions [34]. This operational streamlining is supported by rapid modernisation of hardware and software, as reflected at the national level, where Saudi

Arabia's digital experience maturity reached 86.71% in 2025 [35]. The high impact coefficients for these factors reveal that while a project may be triggered by national policy, its first visible institutional footprint rests on creating a robust, AI-ready technical foundation that simplifies the bureaucratic experience for staff and students.

#### Intermediate Gains: Educational Transformation

Once the infrastructure has stabilised, the impact of Digital Transformation shifts towards Teaching and Learning Effectiveness (H14) ( $\beta = .639, p < 0.001$ ). The qualitative evidence clarifies that academic shifts are not immediate; they depend on the prior successful adoption of infrastructure and a coherent digital strategy to enable personalised, flexible learning pathways. As noted in [29], the 2025 AI curriculum rollouts and digital educational shifts in Saudi Arabia are now beginning to transform classroom quality, but only in environments where the "technical friction" has already been minimised. This indicates that Teaching and Learning Effectiveness is a mediated outcome; it requires a period of institutional "settling" in which staff and students move beyond basic system use to meaningful digital engagement.

#### Long-Term Outcome: Cumulative Institutional Performance

The final stage of this sequence is Institutional Performance (H11:  $\beta = .629, p < 0.001$ ). While highly significant, its slightly lower coefficient compared to Administrative Efficiency suggests that performance gains are a delayed, cumulative effect. This aligns with the analysis of institutional maturity at King Khalid University, which posits that performance quality is the final result of successfully stabilising the underlying technical and administrative ecosystem [36]. In the Saudi context, Institutional Performance is not merely an internal metric but a symbolic signal of institutional standing. It combines gains in operational efficiency and pedagogical excellence to build the reputation and trust needed to meet the rising expectations of national stakeholders.

This provisional cumulative sequence reveals that digital transformation is a sustained commitment, with outcome measures responding to adoption at different rates. The chain begins with Technical Infrastructure (capacity), which enables Administrative Efficiency (automation); these operational gains together facilitate Teaching and Learning Effectiveness (educational quality). These successes culminate in marked improvement in overall Institutional Performance, which ultimately contributes to the high-level socioeconomic goals of Saudi Vision 2030. This sequential logic provides university leadership with a realistic framework for managing expectations, suggesting that the most profound academic and reputational benefits will only emerge after the foundational operational gains have been closed.

### E. Contributions and Implications

Before detailing these contributions, we state explicitly how they extend [13]. The earlier paper validated the DXA-SPU framework statistically but treated all adoption factors and all institutional impacts as simultaneous, undifferentiated effects. The present study replaces that static picture with a

dynamic conceptual vision, extended by phase-dependent drivers on the input side and cumulative impacts on the outcome side, grounded in evidence that the survey could not provide.

This study makes three connected contributions. Theoretically, it advances a stage-based account of adoption in centralised public universities. Whereas quantitative applications of TAM and TOE typically model organisational, technological, and environmental factors as simultaneous direct predictors [11, 12], our findings indicate that Strategy, Regulatory Alignment, and Perceived Usefulness shape the decision to adopt, while Collaboration and Coordination and Competitive Pressure operate primarily at the implementation and positioning stages [3, 13]. This distinction helps explain why factors highly visible in practice may not emerge as direct statistical predictors in cross-sectional data [20, 26]. The framework's outcome side is equally stage-dependent: rather than emerging together, institutional impacts follow a provisional cumulative sequence: Technical Infrastructure forms the foundation, Administrative Efficiency is the first and most visible gain built on it, Teaching and Learning Effectiveness follows, and Institutional Performance accumulates last. These compounding impacts extend to graduate readiness and the broader national knowledge economy [6, 7].

Methodologically, the study demonstrates that a sequential explanatory design can serve an analytical rather than merely descriptive function [24]: the non-significance of H2 and H9 was only resolvable through the qualitative strand, which revealed the centralised governance logic behind them. Interviewing senior decision-makers, rather than end users, is therefore especially suited to explaining the governance processes that shape institutional adoption.

Practically, the findings offer university leaders a staged template. National policy direction creates the legitimacy needed to initiate transformation; leaders should then make system value visible, treat security as an enabling condition rather than a post-adoption concern, support staff through phased rollouts, and plan resourcing across the full adoption cycle [13]. Because benefits accrue progressively, from the Technical Infrastructure foundation through Administrative Efficiency and Teaching and Learning Effectiveness to Institutional Performance, investment must be sustained across successive cycles rather than expected to deliver uniformly or immediately. These contributions invite extension to comparable governance contexts, such as Gulf Cooperation Council university systems, and to decentralised or private settings, to test whether the phase-dependent drivers and cumulative impacts identified here generalise [3, 13].

## VI. CONCLUSION, LIMITATIONS AND FUTURE WORK

This study provides a contextually grounded examination of Digital Transformation in Saudi public universities by integrating quantitative validation [13] with qualitative institutional evidence. By bridging the gap between the statistical results of the DXA-SPU framework ( $n=447$ ) and the operational experiences of senior IT leaders ( $n=11$ ), the research

clarifies that adoption in this setting is a Phase-Dependent process. The findings demonstrate that while national mandates and institutional Strategy trigger the initial decision to adopt, factors such as Collaboration and Coordination and Competitive Pressure emerge as vital facilitators during the implementation and positioning stages. Furthermore, the study establishes that the resulting benefits accumulate in a specific temporal sequence, from the Technical Infrastructure foundation through Administrative Efficiency and Teaching and Learning Effectiveness to Institutional Performance. This longitudinal perspective confirms that digital change is a sustained commitment that ultimately strengthens graduate labour market readiness and contributes to Saudi Arabia's national knowledge economy. Whereas our previous paper [13] reported that 12 of 14 hypothesised relationships held, the present study explains why the two exceptions occurred and the order in which the four institutional benefits materialise, thereby moving the framework from a validated static model to a dynamic, phase-aware conceptual vision.

### Limitations and Future Directions

Despite these insights, several limitations provide opportunities for further research. The primary constraint is the cross-sectional design, which captured institutional perceptions at specific points in time between 2024 and 2025. Because Digital Transformation is an ongoing process, a temporal snapshot cannot fully capture the "intention-behaviour gap," in which strong adoption intentions do not always translate into actual usage [37, 38]. Future research should employ longitudinal designs to track institutions over several years, providing objective evidence of how adoption decisions evolve into long-term implementation. Furthermore, while the qualitative phase targeted strategic decision-makers to address the initial survey's faculty-heavy composition, students' and administrative personnel's perspectives remain underrepresented. Future studies should adopt stratified, multi-stakeholder sampling to capture the diverse experiences of the entire university community.

The geographical and institutional scope of this study also suggests areas for expansion. The findings are specific to the centralised governance of Saudi public universities during a unique period of national transformation under Saudi Vision 2030. Consequently, the non-significance of factors such as Competitive Pressure may not hold for private universities or decentralised international systems. Comparative studies across the Gulf Cooperation Council (GCC) region or between public and private sectors would clarify which aspects of the DXA-SPU framework are universal and which are context-specific. Finally, as the current outcome measures are perceptual, future investigations should link survey data with objective institutional records, such as administrative processing times or research outputs, to strengthen the evidence base for the framework's impact on institutional excellence.

## REFERENCES

- [1] Tang, D. 2021. "WHAT is digital transformation?" EDPACS 64:9–13, <https://doi.org/10.1080/07366981.2020.1847813>.

- [2] Kane, G.C., D. Palmer, A.N. Phillips, D. Kiron, N. Buckley. 2015. "Strategy, not technology, drives digital transformation"
- [3] Vial, G. 2019. "Understanding digital transformation: a review and a research agenda" *The Journal of Strategic Information Systems* 28:118–144, <https://doi.org/10.1016/j.jsis.2019.01.003>.
- [4] Kingdom of Saudi Arabia. 2018. "Saudi Vision 2030." <https://www.vision2030.gov.sa/en>. Online.
- [5] National Transformation Program. 2024. "Beyond ambition: national transformation program annual report 2024"
- [6] Alghamdi, A. 2022. "Digital transformation within Saudi education system: 2020 and beyond" *The Educational Review, USA* 6:419–425, <https://doi.org/10.26855/er.2022.08.014>.
- [7] Singh, A., H.P. Singh, F. Alam, V. Agrawal. 2022. "Role of education, training, and e-learning in sustainable employment generation and social empowerment in Saudi Arabia" *Sustainability* 14:8822, <https://doi.org/10.3390/su14148822>.
- [8] Alhubaishy, A., A. Aljuhani. 2021. "The challenges of instructors' and students' attitudes in digital transformation: a case study of Saudi universities" *Education and Information Technologies* 26:4647–4662, <https://doi.org/10.1007/s10639-021-10491-6>.
- [9] Khawaji, T.M. 2023. "Digital transformation requirements at Saudi universities from faculty members' perspectives" *International Journal of Computer Science & Network Security* 23:8–20, [http://paper.ijcns.org/07\\_book/202311/20231102.pdf](http://paper.ijcns.org/07_book/202311/20231102.pdf).
- [10] Benavides, L., J. Tamayo Arias, M. Arango Serna, J. Branch Bedoya, D. Burgos. 2020. "Digital transformation in higher education institutions: a systematic literature review" *Sensors* 20:3291, <https://doi.org/10.3390/s20113291>.
- [11] Tornatzky, L.G., M. Fleischer, A.K. Chakrabarti. 1990. *The processes of technological innovation*. Lexington, MA: Lexington Books.
- [12] Davis, F.D. 1989. "Perceived usefulness, perceived ease of use, and user acceptance of information technology" *MIS Quarterly* 13:319–340, <https://doi.org/10.2307/249008>.
- [13] Alshehri, S.Z., N. Beloff, M. White. 2025. "Digital transformation in Saudi public universities: a novel framework for adoption drivers and impact analysis" in 2025 20th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 87–98, <https://doi.org/10.15439/2025F8917>.
- [14] Bygstad, B., E. Øvrelid, S. Ludvigsen, M. Dæhlen. 2022. "From dual digitalization to digital learning space: exploring the digital transformation of higher education" *Computers & Education* 182:104463, <https://doi.org/10.1016/j.compedu.2022.104463>.
- [15] Garcia-Morales, V.J., A. Garrido-Moreno, R. Martín-Rojas. 2021. "The transformation of higher education after the COVID disruption: emerging challenges in an online learning scenario" *Frontiers in Psychology* 12, <https://doi.org/10.3389/fpsyg.2021.616059>.
- [16] Fountain, J.E. 2001. *Building the virtual state: information technology and institutional change*. Washington, DC: Brookings Institution Press.
- [17] Dunleavy, P. 2005. "New public management is dead—long live digital-era governance" *Journal of Public Administration Research and Theory* 16:467–494, <https://doi.org/10.1093/jopart/mui057>.
- [18] Verhoef, P.C., T. Broekhuizen, Y. Bart, A. Bhattacharya, J.Q. Dong, N. Fabian, M. Haenlein. 2021. "Digital transformation: a multidisciplinary reflection and research agenda" *Journal of Business Research* 122:889–901.
- [19] Abolhassan, F. 2017. *The drivers of digital transformation*. Cham: Springer International Publishing, <https://doi.org/10.1007/978-3-319-31824-0>.
- [20] Almatrodi, I., D. Skoumpopoulou. 2023. "Organizational routines and digital transformation: an analysis of how organizational routines impact digital transformation transition in a Saudi university" *Systems* 11:239, <https://doi.org/10.3390/systems11050239>.
- [21] Omar, A., A. Almaghthawi. 2020. "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* 11, <https://doi.org/10.14569/IJACSA.2020.0110873>.
- [22] Abdulrahim, H., F. Mabrouk. 2020. "COVID-19 and the digital transformation of Saudi higher education." *Asian Journal of Distance Education* 15:291–306.
- [23] Alotaibi, N.S. 2022. "The significance of digital learning for sustainable development in the post-covid19 world in Saudi Arabia's higher education institutions" *Sustainability* 14:16219, <https://doi.org/10.3390/su142316219>.
- [24] Creswell, J.W., V.L.P. Clark. 2017. *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.
- [25] Braun, V., V. Clarke. 2006. "Using thematic analysis in psychology" *Qualitative Research in Psychology* 3:77–101, <https://doi.org/10.1191/1478088706qp0630a>.
- [26] Al-Khudairi, Z.I. 2026. "Digital transformation and its influence on the marketing strategies of Saudi universities [التحول الرقمي وأثره على استراتيجيات تسويق الجامعات السعودية]" *مجلة العلوم التربوية و الدراسات الإنسانية* 470–53:439, <https://doi.org/10.55074/e9fkat12>.
- [27] IMARC Group. 2025. "Saudi Arabia higher education market: industry trends, share, size, growth, opportunity and forecast 2025-2034."
- [28] Al-Umairi, S.S.M., W.A.A. Muhdar. 2025. "Mechanisms of digital transformation to achieve competitive advantage in Saudi universities: an analytical study [آليات التحول الرقمي لتحقيق الميزة التنافسية في الجامعات: دراسة تحليلية]" *Journals of Qualitative Education* 38, <https://doi.org/10.21608/raes.2025.421112>.
- [29] Ayoub, H. 2025. "Education as the engine of Saudi Vision 2030" *The EDiT Journal, TI - Education as the engine of Saudi Vision 2030: scaling digital transformation*.
- [30] Saeed, M., R. Taher. 2026. "The effectiveness of digital transformation in enhancing institutional performance in higher education institutions" *Arado Business Journal* 1:91–110, <https://doi.org/10.64190/abj.1.2.2026.14>.
- [31] National Transformation Program. 2021. "Delivery plan 2021 - 2025"
- [32] Digital Government Authority. 2024. "Digital government legislation: regulations stimulating digital innovation"
- [33] Alotaibi, E.K. 2026. "Enhancing employability through digital transformation: a strategic approach to strengthening the competitiveness of tourism and hospitality education in Saudi Arabia" *Frontiers in Education* 11, <https://doi.org/10.3389/feduc.2026.1788967>.
- [34] Al-Anzi, S.H.R., M. Toukabri. 2025. "The impact of digital modernity on administrative performance in Saudi universities" *Journal of Management World* 2025:268–277, <https://doi.org/10.53935/jomw.v2024i4.661>.
- [35] Digital Government Authority. 2025. "Digital experience maturity index cycle 2025"
- [36] Mostafa, E.A.A.G., A.B.S. Alhaj, H.E.O. Bashier. 2025. "The impact of digital transformation on TGM of educational institutions in Saudi Arabia (case study: King Khalid University)" *International Journal of Accounting and Economics Studies* 12:208–212, <https://doi.org/10.14419/y4qwyf98>.
- [37] Sheeran, P. 2002. "Intention-behavior relations: a conceptual and empirical review" *European Review of Social Psychology* 12:1–36, <https://doi.org/10.1080/14792772143000003>.
- [38] Turner, M., B. Kitchenham, P. Brereton, S. Charters, D. Budgen. 2010. "Does the technology acceptance model predict actual use? A systematic literature review" *Information and Software Technology* 52:463–477, <https://doi.org/10.1016/j.infsof.2009.11.005>.



# A multithreaded Java implementation of Monte Carlo simulation for Value-at-Risk estimation

Paweł Borowiecki

0009-0008-6825-4038

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 continuous development of computers has made software that performs calculations in various fields of science, including economics, significantly more efficient.

This paper presents two parallel implementations developed in Java and a performance analysis of Monte Carlo simulations for Value-at-Risk estimation, using geometric Brownian motions and a log-normal distribution of returns. The first implementation variant utilizes data parallelism and parallel loops. The second implementation variant, in addition to data parallelism and parallel loops, also utilizes task parallelism and the fork-join model.

Software tests were conducted on three computers equipped with Intel processors. The main area of research was the dependence of simulation execution time and speedup on the number of threads and data size. The program's performance was accelerated by more than 42 times.

## I. INTRODUCTION

ADVANCES in computer technology in recent years have significantly contributed to increasing the efficiency of computations in various fields of science, including economics. This article presents two variants of a multithreaded implementation of Monte Carlo simulation for Value-at-Risk estimation. Value-at-Risk is a key element in financial analysis, as it is one of the factors that allows financial institutions to estimate market risk [1]. One method for estimating Value-at-Risk is Monte Carlo simulation. Obtaining an accurate result using this method involves lengthy computation time [2], so it is worthwhile to employ parallel programming to reduce this time.

To date, numerous studies have been conducted on how to accelerate the Monte Carlo simulation process for Value-at-Risk estimation using Intel AVX [3] and high-performance GPU accelerators [4], as well as the CUDA [5] and Intel MIC architectures [6]. Parallel programming in C/C++ was used to estimate Value-at-Risk using the Delta-Gamma Monte Carlo method [7].

This article aimed to develop an implementation using parallel programming to accelerate the overall simulation process. Java was chosen as the programming language, as it provides numerous tools for handling multithreading. The first implementation variant utilizes data parallelism and parallel loops. The second implementation variant, in addition to data parallelism and parallel loops, also utilizes task parallelism and the fork-join model. Furthermore, few references in the

literature that combine the topic of Value-at-Risk with programming use the Java language.

The software tests were conducted on three server-class computers equipped with Intel processors, each of which has a different architecture and number of cores.

The remainder of the article is structured as follows. Section II provides an introduction to the economic and mathematical theory, which familiarizes the reader with the definition of Value-at-Risk and the basic mathematical formulas on which the implemented simulation is based. Section III describes the details of both implementations, highlighting mechanisms from the `java.util.concurrent` package. In addition, the functioning strategies of both implementations are presented, highlighting the types of parallelism and parallel programming patterns used. Section IV contains a detailed description of the test environment, as well as the results of the software testing, focusing on how time and speedup depend on the number of threads and the data size.

## II. MONTE CARLO SIMULATION IN VALUE-AT-RISK ESTIMATION

Value-at-Risk is a statistical measure of the risk of a decline in market value, based on the current market value, which depends on a specific time horizon and tolerance level [8]. One method for estimating Value-at-Risk is Monte Carlo simulation, which involves repeatedly simulating a random process for given financial instruments in an effort to account for various situations that may arise in the financial market [8]. This allows for the calculation of return rates from predefined probability distributions. This approach enables the reflection of the distribution of the entire investment portfolio, which allows for the estimation of Value-at-Risk [1].

One of the fundamental pillars of Monte Carlo simulation is the use of historical data — past observed asset prices. Furthermore, this method requires the use of an appropriate stochastic process. A correlation matrix is also used, created by multiplying the covariance matrix of the assets, which has previously undergone Cholesky decomposition, by a vector of generated standard normal variables. The process of calculating Value-at-Risk in this way boils down to the following steps:

- 1) Selection of the stochastic process and parameters.

- 2) Generation of pseudo-random numbers used to calculate individual asset prices.
- 3) Calculation of asset values over a given time horizon.
- 4) Calculating potential losses and sorting them.
- 5) Determining Value-at-Risk as a quantile [8].

In this study, geometric Brownian motions combined with a log-normal distribution of returns was used to simulate price paths, and the future values of the assets in the portfolio were calculated using the following formula:

$$S(t) = S(0) \exp \left( \left[ \mu - \frac{1}{2}\sigma^2 \right] t + \sigma W(t) \right), \quad (1)$$

where:

- $S(0)$  — initial value of the asset,
- $\mu$  — drift,
- $\sigma$  — volatility,
- $W$  — standard Brownian Motion.

The above formula can be generalized to the following form:

$$S(t_{i+1}) = S(t_i) \exp \left( \left[ \mu - \frac{1}{2}\sigma^2 \right] (t_{i+1} - t_i) + \sigma \sqrt{t_{i+1} - t_i} Z_{i+1} \right), \quad i = 0, 1, \dots, n-1, \quad (2)$$

where:

- $S(t_{i+1})$  — asset price at time  $t_{i+1}$ ,
- $Z_{i+1}$  — independent standard normal variable [9].

The marginal distribution  $S(t)$  is an exponential normal distribution, known as a log-normal distribution, which allows returns to be calculated using the formula:

$$R_i = \ln \left( \frac{S(t_j)}{S(t_{j-1})} \right) \quad (3)$$

where:

- $R_i$  — return on the  $i$ -th asset at time  $t_{j-1}$ ,
- $S(t_j)$  — asset price at time  $t_j$ .

Potential losses are calculated by subtracting the expected value of the asset from its initial price. These losses are then sorted, and Value-at-Risk is calculated based on this.

### III. IMPLEMENTATION

Two variants of the parallel implementation of the previously described simulation were created. Both implementations use parallel loops. Additionally, the second variant also uses the fork-join model. The algorithm takes as input an array of asset prices in the investment portfolio, an array of asset weights, a confidence level, the number of days for which the simulation is to be executed, the number of price path simulations for each asset, and the number of threads on which the program is to run.

The primary data type used in both implementations is `BigDecimal`, which provides high decimal precision and control over number rounding — a feature that is extremely important from an economic standpoint, as it allows for highly accurate results.

The class responsible for generating pseudorandom numbers when creating a vector of independent normal variables is the `SplittableRandom`, whose instances are passed to threads by calling the `split` method on the parent object, which is not subsequently used to generate pseudorandom values, allowing threads to safely use the newly created objects [10]. Additionally, the Box-Muller transformation was used to obtain values from the normal distribution.

In both implementations, potential losses are sorted by calling the `parallelSort` method of the `Arrays` class. However, ensuring that the covariance matrix is positive definite and performing the Cholesky decomposition is executed sequentially what does not significantly affect the program's efficiency.

The first implementation primarily focuses on data parallelism. The most important class is `MonteCarloSimulationMultithreaded`, which contains the `runSimulation` method responsible for starting the simulation and calling the other methods tasked with calculating the individual parameters necessary to perform the price path simulation. These methods launch thread classes, most of which implement the `Runnable` interface, with the exception of the class involved in calculating the initial portfolio value, which implements the `Callable` interface. The use of this interface necessitated the use of the `ExecutorService` and `Future` interfaces. Each thread receives its own uniform data range on which it performs specific operations, although the last thread created may receive a larger data range if an even distribution is not possible. This approach prevents a race condition for resource access. The `CountDownLatch` class was used as the synchronization mechanism, which notifies the main thread when it is possible to resume work.

In the second parallel implementation, task parallelism was used in addition to data parallelism. The most important class in this implementation is the `MonteCarloSimulationMultithreaded2` class, which, as in the first implementation, has a `runSimulation` method. A `ForkJoinPool` thread pool is created with the number of threads allocated for the entire simulation. Additionally, the `CompletableFuture` class is used, which allows for the asynchronous execution of the following tasks:

- creating and decomposing the covariance matrix,
- calculating volatility,
- calculating the initial portfolio value.

If only two threads are allocated for the entire simulation, the decomposed matrix is created on one thread, while the operations responsible for calculating volatility and the initial portfolio value are performed on the other thread. The methods responsible for performing the tasks listed above do not create separate threads — the method execution process occurs sequentially.

The rest of the methods in the simulation work identically to the first implementation, with the exception of price path simulation, for which the `parallel` method of the `IntStream`

interface is used, allowing the previously created thread pool to fully manage the assignment of tasks to threads.

#### IV. PERFORMANCE ANALYSIS AND TEST ENVIRONMENT

The tests were conducted on three computers designated C1, C2, and C3, each equipped with a different server-class processor, with each subsequent model has better parameters than the previous one. Their detailed specifications are as follows:

- C1 (2x Intel Xeon Gold 5218R) — 20 cores / 40 threads, Cascade Lake Refresh architecture (14 nm);
- C2 (2x Intel Xeon Gold 6342) — 24 cores / 48 threads, Ice Lake-SP architecture (10 nm);
- C3 (2x Intel Xeon Platinum 8358) — 32 cores / 64 threads, Ice Lake-SP architecture (10 nm).

The testing process began by running the sequential implementation, followed by each of the parallel implementations, executed sequentially on a number of threads ranging from 1 to the maximum number of logical threads available. The simulation duration was the total execution time of the runSimulation method. Time measurement was performed using the `System.nanoTime()` method. The speedup is the ratio of the execution time of the sequential version of the program to the execution time of the parallel version. The program was primarily tested using ten different datasets (asset portfolios and weight matrices), each of which had a different size. The remaining input parameters, with the exception of the number of threads such as the confidence level were constant for all datasets.

The symbol  $x$  in the legend of the figures presented later in this article denotes the multiple of the data size relative to the smallest tested data size. This means that in the case of a size of  $10x$ , the data size is ten times larger than the smallest tested data size.

For all computers, the execution time of both implementations is inversely proportional to the number of threads used in the simulation, which is fully consistent with Amdahl's law, which states that the maximum theoretical speedup has already been achieved [11].

Figure 1 shows the relationship between acceleration and the number of threads for the first version of the parallel implementation on computer C1. A sharp increase in performance is noticeable up to 20 threads, followed by a certain plateau, likely caused by the fact that the processor had been using physical cores up to that point, and after exceeding 20 threads, it begins to utilize Hyper-Threading, which causes threads to compete for resources at the level of the physical processor core. From around 40 threads, the performance begins to rise slowly again, as the second processor is also utilized, though the increases are no longer as sharp as at the beginning, because the processor must access resources from RAM.

In Figure 2, which shows the performance of the second parallel implementation on computer C1, a sharp increase in performance up to about 20 threads can be observed, followed by a slight plateau at around 40 threads. This may

The relationship between the speedup and the number of threads for the first parallel implementation on computer C1

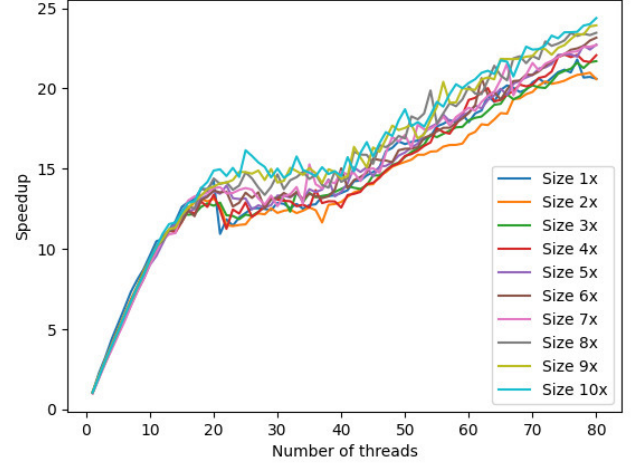


Fig. 1. The relationship between the speedup and the number of threads for the first parallel implementation on computer C1

be caused by competition for thread resources on the same cores, which could be due to work-stealing algorithm in the `ForkJoinPool` thread pool. Next, there is another sharp increase in performance due to the use of a second processor, followed by a plateau at a high number of threads, for the same reasons as in the case of the first plateau.

The relationship between the speedup and the number of threads for the second parallel implementation on computer C1

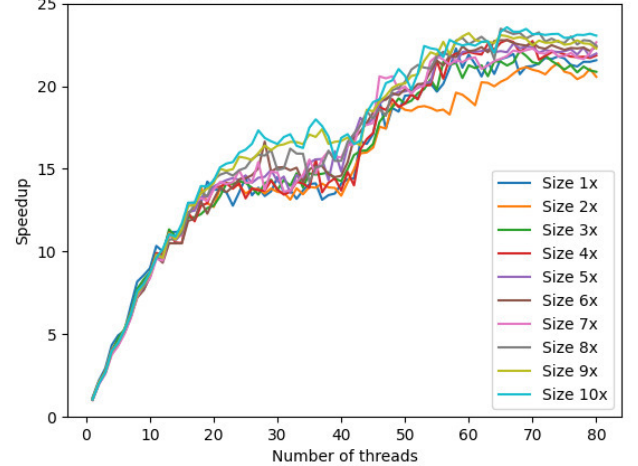


Fig. 2. The relationship between the speedup and the number of threads for the second parallel implementation on computer C1

In Figures 3 and 4, which show the relationship between performance and the number of threads for the first parallel implementation on computers C2 and C3, a rapid increase in performance can be observed until the number of threads equals the number of processor cores, followed by a period of very large fluctuations resulting from the start of more intensive use of the second processor and overheads associated with

the NUMA architecture, such as access to remote memory and increased cache coherence traffic. This is followed by a slight decline in performance and a steady increase that is no longer as rapid, which is likely due to reaching memory bandwidth limits and increased synchronization costs between threads.

processor and the low load on the second processor. However, as the number of threads increases and the second processor's workload grows, further increases in acceleration can be observed. Noteworthy is the drastic drop in performance on the C2 computer for just over 60 threads with the smallest dataset, likely caused by excessive overhead associated with thread management by the thread pool.

The relationship between the speedup and the number of threads for the first parallel implementation on computer C2

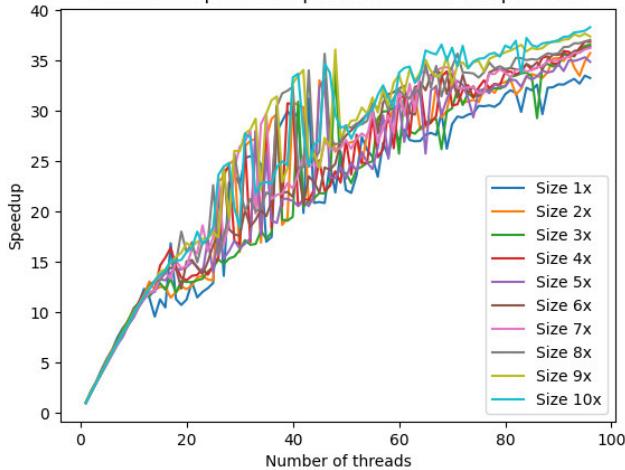


Fig. 3. The relationship between the speedup and the number of threads for the first parallel implementation on computer C2

The relationship between the speedup and the number of threads for the second parallel implementation on computer C2

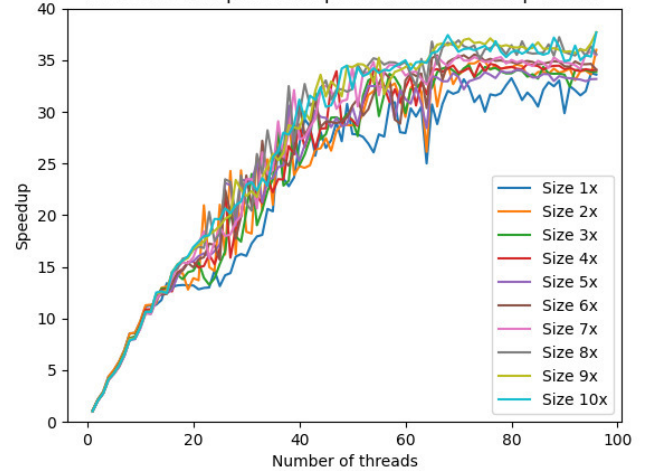


Fig. 5. The relationship between the speedup and the number of threads for the second parallel implementation on computer C2

The relationship between the speedup and the number of threads for the first parallel implementation on computer C3

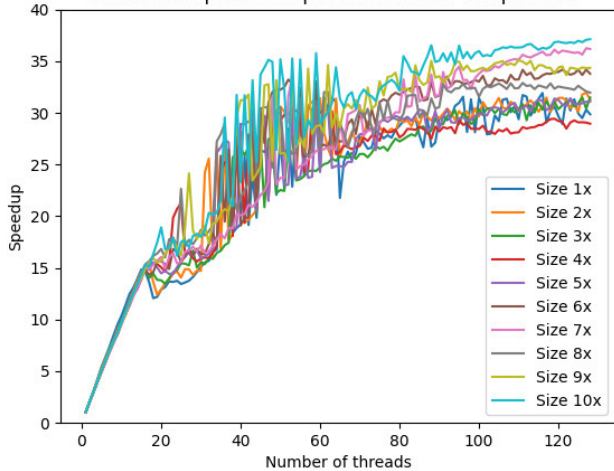


Fig. 4. The relationship between the speedup and the number of threads for the first parallel implementation on computer C3

The relationship between the speedup and the number of threads for the second parallel implementation on computer C3

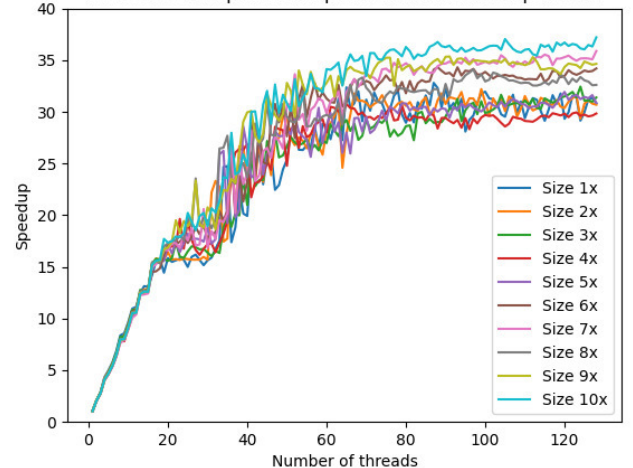


Fig. 6. The relationship between the speedup and the number of threads for the second parallel implementation on computer C3

Figures 5 and 6 show the relationship between performance scaling and the number of threads for the second implementation on computers C2 and C3. A steep increase in performance is observed as long as the number of threads does not exceed the number of physical cores in the processor. Once the number of cores is exceeded, the rate of increase slows down and fluctuations appear, related to the high load on the first

Summarizing the results from all computers, the shortest runtime for the maximum number of threads for data sizes ranging from 1x to 10x was recorded for the first implementation on computer C2, although for small datasets, the results of both implementations on computers C2 and C3 are

very similar, as shown in Figure 7. In contrast, the longest runtime was measured for both implementations running on computer C1.

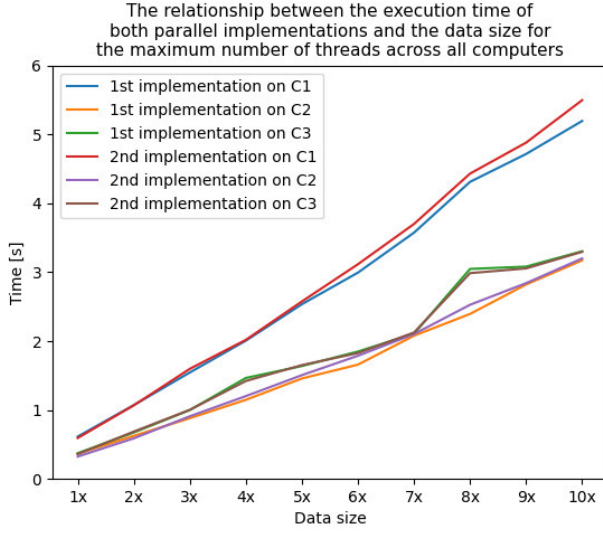


Fig. 7. The relationship between the execution time of both parallel implementations and the data size for the maximum number of threads across all computers

The greatest speedup for data sizes ranging from 1x to 10x was observed for the first version of the implementation on computer C2, with the exception of the smallest dataset, for which the greatest speedup was observed for the second version of the implementation on computer C2, as shown in Figure 8.

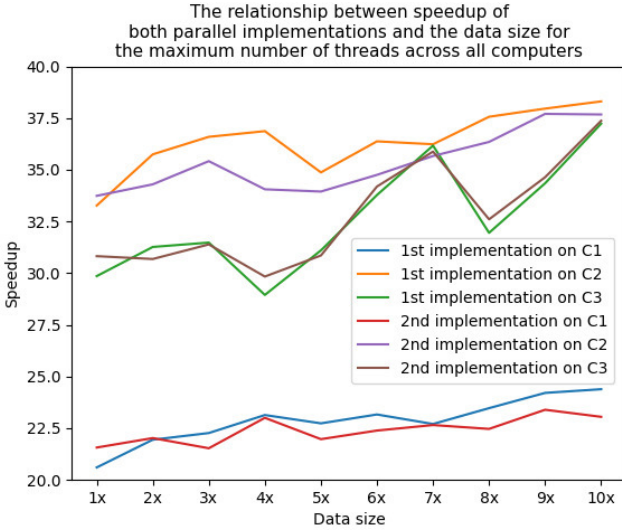


Fig. 8. The relationship between speedup of both parallel implementations and the data size for the maximum number of threads across all computers

The speedups measured for both implementations across data sizes ranging from 1x to 10x, when run on the C3 computer, show a similar trend, and for the 7x data size,

they match the programs running on computer C2. It should be noted that computer C3, which has the largest number of cores, did not achieve the highest speedup for this data range in every case. Therefore, both implementations were additionally tested for data sizes of 15x, 20x, 25x, and 30x. The results are presented in Tables I and II. For a data size of 15x, the highest speedup was recorded for the second implementation variant, running on computer C2. However, for data sizes of 20x, 25x, and 30x, the greatest speedups were recorded for both implementations running on computer C3, and the speedup for the second implementation for 30x data exceeded 42x, indicating that for this computer, the input data size must be sufficiently large.

TABLE I  
SPEEDUP OF THE FIRST IMPLEMENTATION FOR DATA SIZES OF 15x, 20x, 25x AND 30x ON ALL COMPUTERS

| Speedup of the first implementation for data sizes of 15x, 20x, 25x and 30x on all computers |         |         |         |
|----------------------------------------------------------------------------------------------|---------|---------|---------|
|                                                                                              | C1      | C2      | C3      |
| 15x                                                                                          | 19.7235 | 29.6798 | 27.5808 |
| 20x                                                                                          | 20.6108 | 31.0633 | 32.5892 |
| 25x                                                                                          | 25.2650 | 35.7620 | 37.5054 |
| 30x                                                                                          | 25.8888 | 38.0322 | 40.9800 |

TABLE II  
SPEEDUP OF THE SECOND IMPLEMENTATION FOR DATA SIZES OF 15x, 20x, 25x AND 30x ON ALL COMPUTERS

| Speedup of the second implementation for data sizes of 15x, 20x, 25x and 30x on all computers |         |         |         |
|-----------------------------------------------------------------------------------------------|---------|---------|---------|
|                                                                                               | C1      | C2      | C3      |
| 15x                                                                                           | 20.5501 | 32.5164 | 32.3631 |
| 20x                                                                                           | 22.9676 | 34.4352 | 38.1631 |
| 25x                                                                                           | 25.8211 | 37.3377 | 40.5544 |
| 30x                                                                                           | 25.7501 | 38.3150 | 42.7427 |

The smallest speedups were recorded for both implementations running on computer C1. For all data sizes, the first implementation was faster, with the exception of the smallest data size.

## V. CONCLUSION

This paper demonstrated how parallelizing the implementation of Monte Carlo simulations for Value-at-Risk estimation, using geometric Brownian motions and a log-normal return distribution, contributed to faster computation times. The performance of both implementations depended on the types of parallelism and parallelism patterns employed. Tests were conducted on three server-class computers with varying numbers of cores and architectures, focusing on the dependence

of execution time and speedup on the number of threads and the size of the input data.

The tests confirmed the superiority of multithreaded implementations over the sequential version, as the speedup ranged from just over  $19\times$  to more than  $42\times$ .

The greatest performance gains for the  $1x$ – $10x$  datasets were observed on computer C2 for the first implementation variant, with the exception of the smallest dataset, for which the second implementation variant, run on the same computer, was the fastest. In contrast, for larger datasets, the greatest performance gains were recorded for the implementations run on computer C3. The programs running on computer C1 performed the slowest, particularly the second implementation variant, with the exception of the result for the smallest dataset.

Research has confirmed that the effectiveness of parallel implementations depends on the number of processor cores, as well as on how the operating system and the Java virtual machine manage threads. The architecture of the computer system, as well as the type and size of the input data, are also key factors.

Future research on the parallel implementation of Monte Carlo simulations for Value-at-Risk estimation could focus on testing the software not only on various CPUs but also on GPUs. It would be worthwhile to explore other programming languages and compare which implementation yields the greatest performance boost.

## REFERENCES

- [1] Alexander J. McNeil, Rüdiger Frey, Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition*. Princeton University Press, 2015.
- [2] John C. Hull. *Options, Futures and other derivatives*. Pearson Education, 2015.
- [3] Liyanage, DNSS and Fernando, GVMPA and Arachchi, DDMM and Karunathilaka, RDDT and Perera, Amal Shegan. Utilizing Intel advanced vector extensions for Monte Carlo simulation based value at risk computation. Elsevier, 2017. <http://dx.doi.org/10.1016/j.procs.2017.05.156>
- [4] Wen-Mei W. Hwu. *Gpu Computing Gems Jade Edition*. Morgan Kaufmann Publishers, 2011. <http://dx.doi.org/10.1016/C2010-0-68654-8>
- [5] Wei Wu. *Acceleration of Monte Carlo Value at Risk Estimation Using Graphics Processing Unit (GPU)*. City University of New York Academic Works, 2010.
- [6] Tiansheng Wen, Rui Mao, Cheng Tan. Parallel Monte Carlo Simulation of VaR Calculation Based on Intel MIC Architecture. Institute of Electrical and Electronic Engineers, 2020. <http://dx.doi.org/10.1109/ICBDIE50010.2020.00116>
- [7] Nan Zhang, Ka Lok Man, Eng Gee Lim. Parallel Computation of Value at Risk using the Delta-Gamma Monte Carlo Approach. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2014.
- [8] Philippe Jorion. *Value At Risk The new Benchmark for Managing Financial Risk*. The McGraw-Hill Companies, 2007.
- [9] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003. <http://dx.doi.org/10.1007/978-0-387-21617-1>
- [10] Oracle Corporation. *Java Platform, Standard Edition & Java Development Kit Version 22 API Specification*. Oracle America Inc, 2024.
- [11] Jamie Allen Roland Kuhn, Brian Hanafey. *Reactive Design Patterns*. Manning Publications Co, 2017.

# INCLUDE-AI and G.I.U.L.I.A.: A Multi-Level European and Territorial Framework for Inclusive, Trustworthy, and Gender-Aware Artificial Intelligence in Education

Matteo Ciaschi  
0009-0009-5119-3563

National Research Council (CNR), Rome, Italy  
matteo.ciaschi@cnr.it

Daniel Dan  
0000-0002-7251-7899

School of Applied Data Science, Modul University Vienna, Austria  
daniel.dan@modul.ac.at

**Abstract**—Artificial Intelligence (AI) is increasingly having a strong impact on education, spanning personalization, automation, and data-driven decision-making. Although promising and disruptive, its implementation also entails structural risks, including algorithmic bias, unequal access to digital resources, and a lack of governance frameworks. OECD (2023) reports that over 60% of schools in high-income countries use AI-driven educational tools, while a UNESCO (2023) survey of 450 institutions identifies the absence of national regulations as a primary reason for the lack of institutional guidance. This results in uneven and fragmented adoption of AI tools and policies.

This paper provides a dual contribution. First, it introduces INCLUDE-AI (INCLUsive Artificial Intelligence for Democratic and Accessible Education), a Europe-oriented multi-level governance framework designed to operationalize inclusive and trustworthy AI across educational systems. Second, it presents G.I.U.L.I.A. (Gender-Inclusive Use of Learning Intelligence & AI), a territorial implementation model developed in the Bologna educational ecosystem as a context-sensitive extension of the broader framework. The study adopts a cross-national and participatory design perspective involving Italy and Austria, combining ethnographic observation with applied socio-technical analysis. To demonstrate how European-level strategies can be translated into localized, gender-aware educational practices, the paper proposes a socio-technical architecture, a governance taxonomy, and a set of measurable indicators. Evidence from an exploratory mixed-methods pilot involving 312 students and 78 educators indicates a persistent gender-based gap in self-assessed AI literacy and conditional educator trust, suggesting that inclusion treated as a design principle rather than a corrective measure is a precondition for equitable AI adoption in education.

**Index Terms**—Artificial Intelligence in Education, Inclusive Education, AI Ethics, Digital Governance, Learning Analytics, Socio-Technical Systems, Educational Innovation, Complex Systems Management

## I. INTRODUCTION

ATTEMPTS to introduce Artificial Intelligence into education began in 1960 with PLATO (Programmed Logic for Automatic Teaching Operations), but the major breakthrough came with ChatGPT in 2022. The integration of Natural Language Processing is among today's major transformations

in knowledge distribution, with formal and informal learning now using adaptive environments, predictive analytics, and automated assessment. This fast diffusion of generative AI poses significant ethical and regulatory challenges, including data management, algorithmic bias, and systemic inequalities [1]. On the opportunity side, personalized learning has been shown to produce statistically significant gains in student performance, particularly in mathematics [2].

However, notwithstanding the technological progress, critical evidence challenges this promise, pointing to systemic risks. Because algorithmic systems use data at their core, implement assumptions, and are applied in institutional contexts, they are not unbiased. Commercial facial analysis systems based on machine learning can exhibit substantial disparities in accuracy across demographic groups [3]. This points out to biases that are automatically embedded at the core of AI systems through data. In educational settings, these risks are further amplified by structural inequalities in access to digital infrastructures. The World Bank (2021) [4] estimates that a majority of students in many countries lack access to digital devices or internet connectivity. This results in a strong, uneven participation and a broadened lack of opportunities for them.

At the same time, governance has not kept pace with technological development. The EU has responded with major initiatives—the AI Act (2024), which classifies educational applications as high-risk given their impact on fundamental rights, and the Digital Education Action Plan (2021–2027) [5], which sets out a vision for inclusive, accessible digital education. Yet the translation of these frameworks into operational practice remains uneven and fragmented, particularly at the regional level: schools frequently lack the technical infrastructure, competencies, and governance structures needed to engage critically with AI, while the opacity of many systems further reduces educators' trust and capacity to evaluate algorithmic recommendations [6].

This paper contributes to the emerging field of Artificial Intelligence in education in three main ways. First, it devel-

ops INCLUDE-AI, a multi-level governance framework that integrates infrastructure, data governance, algorithmic design, and pedagogy within a unified socio-technical perspective. Second, it introduces G.I.U.L.I.A. as a territorially grounded implementation model, offering an applied, context-sensitive extension of European-level strategies, with a specific focus on gender inclusion. Third, the paper advances the operationalization of inclusion-by-design by proposing a set of measurable dimensions and indicators to evaluate fairness, accessibility, and participation in AI-enabled educational systems.

Unlike existing frameworks, this study explicitly operationalizes inclusion through measurable indicators and integrates gender as a structural dimension of AI governance.

The remainder of the paper is organized as follows. Section II reviews related work and positions INCLUDE-AI against existing frameworks. Section III presents the INCLUDE-AI architecture, governance taxonomy, and indicators. Section IV introduces G.I.U.L.I.A. and maps it onto INCLUDE-AI. Section V describes the mixed-methods design and empirical findings. Sections VI and VII discuss implications and limitations, and Section VIII concludes.

## II. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

The literature on Artificial Intelligence in education increasingly recognizes both its potential and its limitations. While AI enables new forms of personalization, allowing systems to adapt to individual learning paths, several studies highlight that these systems are often built on data structures that reflect existing social and institutional biases. Students from disadvantaged backgrounds use less structured and less complete data due to factors such as unstable internet access, lower digital literacy, and limited engagement with digital learning platforms [7]. A recent meta-systematic review highlights that the field of AI in education remains dominated by performance-oriented studies, while ethical and governance dimensions are still underdeveloped [8].

From an ethical perspective, the debate has increasingly focused on the principles of transparency, fairness, and accountability. A comprehensive framework for ethical AI has emphasized the importance of human-centered design and responsible innovation [9]. However, ethical principles alone are insufficient if they are not embedded within the technical and organizational structures of AI systems [10]. This observation is particularly relevant in educational contexts, where the interaction between technology and pedagogy is deeply intertwined.

AI literacy is widely seen as a critical enabler of meaningful engagement with AI systems [11], encompassing not only technical skills but also the capacity to understand, critically evaluate, and ethically reflect on algorithmic systems. Crucially, this competence is unevenly distributed, with disparities associated with gender, socio-economic background, and geographic context [12]. Where students and educators cannot critically assess the outputs of adaptive systems, existing

inequalities are reproduced rather than reduced, making inclusive and context-sensitive AI-literacy programs a precondition for equitable integration of AI in education [13].

In parallel, research on digital infrastructures highlights the importance of robust, interoperable systems to support AI applications. In education, interoperable digital infrastructures and shared platforms enable data to flow across systems and support large-scale storage, processing, and analysis [14]. Yet, the relationship between infrastructure and pedagogy remains underexplored, with most studies treating these domains separately. This fragmentation limits the ability to develop integrated solutions that address the multifaceted nature of AI in education. For example, recent FedCSIS contributions present capable AI-driven educational tools—an LLM-based feedback generator for an intelligent tutoring system [15] and an AI-supported generator of programming exercises [16]—yet such tools are typically designed and evaluated as isolated artifacts, without an explicit data-governance, fairness, or inclusion layer wrapped around them. A device-provisioning decision taken independently of teacher training and data governance can likewise leave a well-designed adaptive tool unused where connectivity or competence is lacking. INCLUDE-AI is positioned precisely at this gap: rather than proposing another tool or another set of abstract principles, it specifies how infrastructure, data, algorithms, and pedagogy must be governed jointly, and pairs each dimension with a measurable indicator—an emphasis that the compared frameworks [6], [9], [13] address only partially (Table I).

The theoretical foundation of this work builds upon the concept of socio-technical systems, which underlines the interdependence between technological components and social practices. Within this perspective, educational institutions are systems in which AI technologies, data governance mechanisms, pedagogical practices, and institutional values are intertwined. Taken together, these perspectives suggest that AI in education cannot be adequately addressed through isolated technical or ethical approaches. It requires an integrated socio-technical framework in which governance, infrastructure, and pedagogy are treated as interdependent dimensions.

Table I provides a dimension-based comparison between INCLUDE-AI and the most relevant existing frameworks in this area. Specifically, the table contrasts INCLUDE-AI with Floridi et al.'s general ethical AI framework [9], Holmes et al.'s education-specific treatment of AI ethics [6], and UNESCO's AI Competency Framework for Teachers [13]. This comparison is not intended to equate these frameworks methodologically, but rather to examine whether they explicitly address core dimensions relevant to equitable AI integration in education, namely infrastructure, data governance, algorithmic fairness, pedagogy, gender inclusion, and measurable indicators.

## III. THE INCLUDE-AI FRAMEWORK

The central contribution of this paper is a multilevel framework that operationalizes inclusive and ethical AI within

TABLE I  
COMPARISON OF INCLUDE-AI WITH EXISTING FRAMEWORKS

| Dimension             | INCLUDE-AI | Floridi | Holmes  | UNESCO  |
|-----------------------|------------|---------|---------|---------|
| Infrastructure        | ✓          | –       | –       | –       |
| Data                  | ✓          | ✓       | ✓       | Partial |
| Governance            | ✓          | ✓       | ✓       | Partial |
| Algorithmic           | ✓          | –       | ✓       | ✓       |
| Fairness              | ✓          | –       | Partial | Partial |
| Pedagogy              | ✓          | –       | –       | –       |
| Gender Inclusion      | ✓          | –       | –       | –       |
| Measurable Indicators | ✓          | –       | –       | –       |

educational ecosystems, moving beyond abstract ethical discussion toward a design-oriented approach to AI governance at the European level. Its premise is that inclusion and trustworthiness cannot be achieved through isolated interventions but require systemic integration across multiple layers of the educational ecosystem.

At the infrastructural level, the framework emphasizes the importance of robust, scalable digital environments that support data-intensive AI applications. This includes cloud-based architectures, interoperable data systems, and secure network infrastructures aligned with European strategies on digital sovereignty. Infrastructure is not treated as a neutral enabler, but as a determinant of access and participation, directly influencing the distribution of opportunities within educational systems.

The data governance dimension addresses issues related to data quality, bias mitigation, and regulatory compliance. Particular attention is given to the alignment with European legal frameworks, including the General Data Protection Regulation and the emerging AI Act. The framework aims to adopt data models that ensure transparency, accountability, and ethical principles.

At the algorithmic level, INCLUDE-AI supports the development of explainable and fairness-aware models. The opacity of many AI systems has been identified as a major barrier to trust and adoption, particularly among educators, for whom understandability and explainability are central to confidence in AI-enabled educational tools [17]. By promoting explainability and auditability, the framework seeks to enhance users' capacity to critically engage with algorithmic outputs.

The pedagogical level integrates principles of inclusive education into the design and implementation of AI systems. This involves adapting technologies to diverse learning needs and rethinking educators' roles as active agents in building AI-enhanced learning environments. This level also addresses the risk of cognitive offloading. In such cases, students may delegate critical thinking and analytical reasoning to AI systems, thereby undermining the development of their cognitive skills. The framework advocates pedagogical approaches that position AI as a tool to augment, not replace, human cognitive engagement, ensuring that learners retain agency over their learning processes.

The four levels of the INCLUDE-AI framework are not

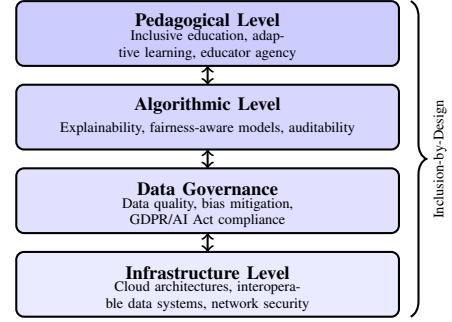


Fig. 1. INCLUDE-AI Framework: Four-layer socio-technical architecture with bidirectional interdependencies.

conceived as independent layers, but as interdependent dimensions that interact dynamically. For example, the quality and representativeness of training data (data governance) directly affect the fairness of algorithmic outputs (at the algorithmic level). This, in turn, shapes the pedagogical experiences of diverse learner populations (pedagogical level). Similarly, the availability and robustness of digital infrastructure (at the infrastructure level) constrain the range of AI applications that can be deployed in a given educational context. Figure 1 illustrates the four-layer architecture. The vertical ordering should be read as a socio-technical stratification (from material substrate to human practice) rather than as a strict service-dependency stack: the bidirectional arrows denote mutual constraint between adjacent strata, not a one-directional, top-down dependency. The model is thus better understood as a constrained interaction graph rendered in layered form for readability, and the layers are not assumed to be independently substitutable.

The governance taxonomy maps core dimensions across levels, actors, and mechanisms. This structure facilitates the transition from normative principles to actionable governance practices, supporting measurable, context-sensitive implementation. Table II presents this taxonomy.

A key innovation of INCLUDE-AI is the development of measurable indicators. This allows for the empirical assessment of inclusion, addressing a major limitation of existing frameworks that remain primarily normative. Table III presents the core inclusion-by-design indicators that guide the framework's operationalization.

#### IV. G.I.U.L.I.A.: TERRITORIAL IMPLEMENTATION AND GENDER-INCLUSIVE GOVERNANCE

While INCLUDE-AI provides a European-level architecture, its effective implementation requires translation into context-specific practices. The G.I.U.L.I.A. framework represents this translation, embodying a territorial<sup>1</sup> and applied dimension developed within the Bologna (Emilia-Romagna Region) educational ecosystem.

<sup>1</sup>We use *territorial* in the sense of European cohesion and regional policy, i.e. a sub-national (regional/local) level of implementation; the term is used interchangeably with *regional* throughout.

TABLE II  
INCLUDE-AI GOVERNANCE TAXONOMY

| Dimension       | Level       | Actors                       | Mechanisms                            | Indicators                       |
|-----------------|-------------|------------------------------|---------------------------------------|----------------------------------|
| Infrastructure  | Macro/Meso  | Governments, Institutions    | Cloud systems, interoperability       | Access rate, system availability |
| Data Governance | Macro/Meso  | Regulators, Schools          | GDPR compliance, data stewardship     | Data transparency, bias audits   |
| Algorithmic     | Meso/Micro  | Developers, EdTech providers | Explainability, fairness-aware models | Perceived fairness, auditability |
| Pedagogical     | Micro/Meso  | Teachers, Students           | Co-design, inclusive curricula        | Engagement, inclusion index      |
| Governance      | Cross-level | Policymakers, Stakeholders   | Multi-stakeholder accountability      | Trust level, participation rate  |

TABLE III  
INCLUDE-AI: INCLUSION-BY-DESIGN INDICATORS

| Dimension     | Indicator                                 | Measurement Approach                                |
|---------------|-------------------------------------------|-----------------------------------------------------|
| Access        | Availability of AI-enabled learning tools | Infrastructure audit; student survey                |
| Fairness      | Perceived algorithmic fairness            | Likert-scale educator/student survey                |
| Literacy      | Gender-disaggregated AI literacy levels   | Self-assessment questionnaire                       |
| Transparency  | Transparency perception among educators   | Semi-structured interviews; survey                  |
| Participation | Stakeholder involvement in AI governance  | Co-design workshop attendance; qualitative analysis |

The conceptual contribution of this project lies in its explicit focus on gender inclusion as a structural component of AI governance. Existing approaches often treat gender disparities as secondary effects of technological systems; by contrast, G.I.U.L.I.A. frames gender as a central dimension influencing access to resources, participation in decision-making, and interaction with AI systems.

In operational terms, “gender-aware” refers to four concrete mechanisms enacted in the pilot rather than to a general orientation. First, *gender-disaggregated baseline measurement*: AI-literacy and confidence were measured separately for male and female students (Table VIb) and used as a design input, so that observed gaps shaped subsequent activities rather than merely being reported. Second, *balanced participation*: co-design workshops applied a participation target ensuring that female students and teachers were represented in proportion to, or above, their share in each setting, and facilitation explicitly elicited their input. Third, *stereotype review of materials*: worked examples, prompts, and role models used in the AI-supported activities were screened to avoid gendered stereotypes (e.g. default male framing of technical roles) and revised accordingly. Fourth, *gender-sensitive monitoring*: the inclusion indicators in Table III retain gender as a disaggregation axis, so equity effects are tracked over time rather than assumed. These mechanisms are deliberately lightweight and procedural, which is what makes them transferable to institutions with

TABLE IV  
MAPPING G.I.U.L.I.A. LEVELS OF ACTION ONTO INCLUDE-AI LAYERS

| G.I.U.L.I.A. level | INCLUDE-AI layer(s) enacted                               | Concrete mechanism in pilot                                                              |
|--------------------|-----------------------------------------------------------|------------------------------------------------------------------------------------------|
| Individual         | Pedagogical, Algorithmic (end-user view)                  | Critical AI-literacy modules; reading and contesting algorithmic outputs                 |
| Pedagogical        | Pedagogical, Data Governance                              | Inclusive curricula; consent-aware data use; bias review of teaching materials           |
| Organizational     | Infrastructure, Data Governance, Governance (cross-level) | Procurement and interoperability decisions; accountability roles; transparency protocols |

limited resources.

The framework adopts a multi-level perspective. At the individual level, it promotes the development of critical AI literacy among educators and students, enabling them to understand and question algorithmic processes. At the pedagogical level, it supports the design of inclusive learning environments that account for diverse cognitive and social dynamics. At the organizational level, it introduces governance mechanisms to enhance transparency, accountability, and participation in AI adoption decision-making.

It is important to clarify the relationship between the two frameworks, since their organizing axes differ by design. INCLUDE-AI is structured by socio-technical *layers* (infrastructure, data governance, algorithmic, pedagogical), i.e. *what* must be governed; G.I.U.L.I.A. is structured by *levels of action* (individual, pedagogical, organizational), i.e. *who* acts and at which scale. G.I.U.L.I.A.’s levels are therefore not an alternative taxonomy but the operational projection of INCLUDE-AI’s layers onto the actors that enact them in a given territory: each level of action takes responsibility for one or more layers and translates the corresponding requirements into concrete practices. Table IV makes this mapping explicit. This separation of concerns is what allows a single European-level architecture to be implemented by heterogeneous local actors without prescribing a uniform technical stack.

Bologna offers fertile ground for this experimentation: a longstanding tradition of educational innovation, major research institutions, and a dense network of schools and civic associations support multi-stakeholder collaboration. The pilot activities involved secondary schools and vocational training institutions across the Bologna and Ravenna metropolitan

areas, spanning urban and semi-urban settings to capture variation in institutional capacity and digital readiness.

The pilot implementation followed a phased approach. In the first phase, baseline assessments were conducted to map existing levels of AI literacy, digital infrastructure availability, and institutional governance practices across participating schools. In the second phase, co-design workshops brought together educators, students, school administrators, and technology providers to collaboratively develop inclusive pedagogical strategies and governance protocols for AI adoption. In the third phase, the proposed strategies were implemented in classroom settings and monitored through surveys, interviews, and ethnographic observation. This iterative cycle of assessment, co-design, and implementation embodies the research-action methodology underpinning the project, ensuring the framework remains grounded in real-world constraints and responsive to stakeholder feedback.

As summarized in the crosswalk of Table IV, the model translates INCLUDE-AI’s European-level governance into the Bologna ecosystem through three levels of action (individual, pedagogical, organizational), with gender inclusion as a cross-cutting dimension spanning all three.

## V. METHODOLOGY

The research adopts a mixed-method and participatory approach, combining qualitative and quantitative techniques within a comparative framework. The choice of Italy and Austria as case studies reflects their differing levels of digital maturity and policy development, thus providing a meaningful basis for cross-contextual analysis.

Data collection integrates multiple sources and operates across three analytical levels—macro, meso, and micro—ensuring both robustness and real-world (contextual) validity.

### A. Macro-Level Analysis

At the macro-level, consolidated international datasets contextualize AI in education globally. OECD data (2021; 2024) [14], [18] indicate that 60–65% of AI-based educational systems rely on behavioral and interaction data, with adaptive technologies improving outcomes by up to 30% in well-resourced settings. UNESCO reports (2023–2025) [19], [20] reveal a governance gap: fewer than 10% of countries have comprehensive AI-in-education regulation and fewer than 35% have formal governance guidelines, while some 2.6 billion people remain offline. World Bank data (2021) [4] confirm that around 40% of students lack reliable access to digital learning tools, with markedly lower rates in low-income contexts.

### B. Meso-Level Analysis

At the meso-level, the research develops a comparative analysis between Italy and Austria, integrating data from the European Commission’s Digital Education Action Plan (2022–2025) [5] and national statistics. Table V presents a comparative assessment of digital readiness between the two countries.

TABLE V  
COMPARATIVE DIGITAL READINESS: ITALY VS. AUSTRIA

| Indicator                          | Austria     | Italy  |
|------------------------------------|-------------|--------|
| Schools with digital platforms (%) | 85          | 68     |
| Teacher digital training (%)       | 64          | 42     |
| Female STEM participation (%)      | 34          | 27     |
| AI experimentation level           | Medium-High | Medium |

Here, the *AI experimentation level* is an ordinal summary of three observable signals: the presence of a national AI-in-education strategy or guidelines, the share of schools running structured AI pilots (as opposed to ad-hoc individual use), and the availability of dedicated teacher-training provision. A *medium* level denotes emerging, largely bottom-up experimentation without a consolidated national strategy, whereas *medium-high* denotes a coordinated national strategy with institutionalized pilots. By this measure both countries exhibit medium to medium-high levels of AI experimentation; the gaps in digital platforms (85% vs. 68%) and teacher digital training (64% vs. 42%) reflect Austria’s more structured national investment in educational technology and teacher competence.

Gender disparities in STEM participation persist in both contexts (34% in Austria, 27% in Italy). This is relevant to AI education because lower female STEM participation correlates with reduced exposure to computational thinking, which affects self-assessed AI literacy and confidence—providing empirical grounding for the gender-aware design of G.I.U.L.I.A.

### C. Micro-Level Analysis and Empirical Findings

At the micro-level, the study incorporates primary data collected through the G.I.U.L.I.A. pilot implementation within the Bologna educational ecosystem. The empirical component includes surveys administered to students (N=312) and teachers (N=78), ethnographic observation in educational settings, semi-structured interviews, and co-design workshops involving local stakeholders.

The qualitative component included 4 co-design workshops involving educators, students, and institutional stakeholders (N=46 participants total), as well as 12 semi-structured interviews with teachers, school leaders, and policy actors. The co-design workshops followed a participatory design protocol structured in three phases: (i) problem framing, (ii) collaborative ideation, and (iii) validation of governance and pedagogical solutions. Each workshop lasted approximately two hours and used scenario- and persona-based prompts (e.g. an adaptive tutoring tool producing an opaque recommendation) to surface concrete trust and fairness concerns; their outputs fed directly into draft governance protocols and the inclusion indicators. The student and teacher surveys used closed Likert-scale items adapted from established AI-literacy and technology-acceptance instruments, complemented by demographic items enabling the gender-disaggregated analysis. Semi-structured interviews were conducted using a protocol focusing on trust, perceived risks, usability, and gover-

nance expectations related to AI systems. All qualitative data were analyzed using thematic coding, following an inductive-deductive approach.

The quantitative findings are summarized in Table VI.

Gender-disaggregated analysis reveals significant differences in self-perceived AI competence. While 26% of male students report high confidence in using AI systems, only 14% of female students do so. Conversely, a higher proportion of female students (42%) report low confidence compared to their male counterparts (26%). Within this exploratory sample, these findings indicate a gender-based confidence gap, consistent with existing literature on digital self-efficacy and STEM participation. As the data are self-reported and the sample is local, the gap should be read as a measured association rather than as a generalizable effect.

Thematic analysis of the qualitative data identified three dominant dimensions that reinforce and contextualize the quantitative findings:

*Trust.* Educators consistently expressed that trust in AI systems is contingent on understanding how decisions are produced. As one secondary school teacher stated: “I would use AI more if I understood how decisions are made.” This conditional trust aligns with the quantitative finding that 61% of educators report trust dependent on improved transparency.

*Agency.* Participants emphasized the importance of maintaining human control and critical engagement with AI systems. A workshop participant noted: “Students should not just use AI, but understand and question it.” This theme underscores the need for AI literacy as a prerequisite for meaningful participation.

*Opacity.* The perceived lack of transparency in AI systems emerged as a recurring concern across both interviews and workshops. As one teacher expressed: “AI feels like a black box—we don’t know what happens behind the interface.” This perception is directly connected to the 63% of educators who identified a lack of transparency as a primary risk.

These themes underscore the centrality of transparency and user empowerment, and directly informed the design of the INCLUDE-AI governance taxonomy and inclusion indicators; the participatory, co-design dimension also strengthens the study’s real-world (contextual) validity.

Overall, the integrated dataset reveals three structural tensions that underpin the rationale of this study: (i) a performance–equity gap, whereby the benefits of AI are unevenly distributed; (ii) an adoption–governance gap, reflecting the misalignment between technological diffusion and regulatory frameworks; and (iii) an access–capability gap, between infrastructure availability and actual competencies and trust. Exploratory in scale, these insights nonetheless directly informed the design of the dual INCLUDE-AI and G.I.U.L.I.A. frameworks as multi-level, inclusion-oriented governance models.

## VI. DISCUSSION

To keep claims of different kinds distinct, we read the results in three registers, as suggested in the review process: empirical

findings from the pilot, which are exploratory and local; the conceptual proposals embodied in the two frameworks; and policy recommendations, which follow analytically from the analysis and remain to be tested. Read in this way, the exploratory findings are consistent with—rather than proof of—the argument that integrating governance, inclusion, and socio-technical design enables more transparent and context-aware adoption of AI in education. At the conceptual level, the analysis highlights a recurring limitation of current approaches to AI in education: the tendency to treat ethical considerations as external constraints rather than integral components of system design.

Theoretically, the work extends socio-technical systems theory by treating inclusion and gender as structural rather than incidental dimensions; practically, the Bologna case suggests that regional ecosystems can serve as laboratories for validating and refining European strategies, underlining the value of place-based innovation in responsible AI.

The pilot data give each of the three tensions a concrete face. The performance–equity gap appears in the contrast between students’ positive perceptions of AI (74%) and the gender disparities in self-assessed literacy: recognition of AI’s benefits is not matched by an even capacity to engage with it critically. The adoption–governance gap shows in conditional educator trust (61%), indicating that current deployment does not yet meet stakeholders’ governance expectations. The access–capability gap surfaces in the Italy–Austria comparison, where differences in infrastructure and teacher training track different levels of AI readiness.

These tensions are not unique to the Italian or Austrian context but reflect broader challenges in the European educational landscape. The INCLUDE-AI framework, by addressing these tensions through an integrated socio-technical approach, provides a model that is potentially transferable to other national and regional contexts, subject to appropriate contextual adaptation.

A legitimate objection is that many schools and institutions lack the technical infrastructure, staff competencies, and governance structures needed to adopt AI tools at all, which could place a framework of this kind out of reach. The design responds to this in three ways. First, the framework is meant to be entered incrementally rather than wholesale: an institution with limited resources can begin at the data-governance and pedagogical layers—which are largely organizational and procedural—before committing to infrastructure investment. Second, the governance taxonomy (Table II) doubles as a readiness self-assessment, letting an institution locate its current capacity per dimension and prioritize the smallest viable next step rather than a full programme. Third, by assigning the infrastructure and interoperability burden primarily to the macro/meso levels (regional authorities, shared platforms), the model deliberately shifts cost away from individual under-resourced schools toward pooled provision. The gender-aware mechanisms described in Section IV were chosen on the same principle: they are procedural and low-cost rather than technology-dependent. We do not claim this removes the ca-

TABLE VI

G.I.U.L.I.A. PILOT SURVEY RESULTS: (A) STUDENT PERCEPTIONS AND SELF-ASSESSED AI LITERACY (N=312); (B) SELF-ASSESSED AI LITERACY BY GENDER (N=312; MALE=168, FEMALE=144); (C) EDUCATOR TRUST AND PERCEIVED RISKS (N=78).

| Dimension          | Category | %  |                        |        |          | Dimension       | Category             | %  |
|--------------------|----------|----|------------------------|--------|----------|-----------------|----------------------|----|
| Overall perception | Positive | 74 | AI literacy            | Male % | Female % | Trust level     | High                 | 22 |
|                    | Neutral  | 21 |                        |        |          |                 | Conditional          | 61 |
|                    | Negative | 5  |                        |        |          |                 | Low                  | 17 |
| AI literacy (self) | High     | 18 |                        |        |          | Perceived risks | Data privacy         | 71 |
|                    | Medium   | 46 | High                   | 26     | 14       |                 | Lack of transparency | 63 |
|                    | Low      | 36 | Medium                 | 48     | 44       |                 | Algorithmic bias     | 48 |
| (a) Students       |          |    | Low                    | 26     | 42       | (c) Educators   |                      |    |
|                    |          |    | (b) Literacy by gender |        |          |                 |                      |    |

capacity gap—closing it ultimately requires sustained investment and policy support—but the framework is structured so that partial, resource-constrained adoption is still meaningful.

Within the EDUC-AI-TION perspective, conversational AI systems (chatbots and generative assistants) are a critical application domain, raising specific challenges of (i) epistemic trust in AI-generated responses, (ii) bias amplification, as language models may reinforce stereotypes, and (iii) student dependency and cognitive offloading. By integrating explainability, co-design, and inclusion indicators, INCLUDE-AI offers a governance structure for deploying such tools transparently and responsibly.

#### A. Implications for Practitioners

The findings carry concrete implications for three groups of actors. For educators, the results underscore AI literacy not merely as a technical skill but as a critical competence; teacher-training programs should integrate modules on algorithmic transparency, data governance, and ethical AI use, so that educators can act as informed mediators between AI systems and learners. The finding that only 22% of teachers report high trust suggests that current professional development does not yet address the governance and transparency concerns that shape adoption decisions. For school administrators, the governance taxonomy (Table II) functions as a readiness tool: by mapping dimensions, actors, and mechanisms, leaders can locate gaps in institutional capacity, while the inclusion indicators (Table III) support monitoring of equity effects over time.

For EdTech developers, the emphasis on explainability and fairness-aware design should inform tools that expose transparent decision processes and accommodate diverse learners. The recurrent theme of opacity indicates that current educational AI systems fall short of users' transparency expectations, implying both an ethical and a market imperative for more interpretable, auditable tools.

#### VII. LIMITATIONS AND FUTURE WORK

This study has several limitations. First, the empirical component is limited in scale: the G.I.U.L.I.A. pilot is confined to the Bologna/Ravenna context, which favours real-world (contextual) validity and depth but constrains generalizability, so transferability to other—especially non-European—contexts needs further validation. Second, the research-action design

prioritizes iterative development over large-scale quantitative testing, and the absence of longitudinal data limits claims about long-term impact. Third, the operationalization of inclusion remains partially exploratory: the indicators require psychometric refinement and cross-institutional comparison. Fourth, explainability and auditability are moving targets as AI systems evolve, and the alignment between European policy, local practice, and infrastructure is an ongoing institutional process beyond the scope of this study.

These limitations define a research agenda. The priorities are: (i) large-scale, cross-country validation beyond a single region (e.g. Nordic, Southern-European, and non-European contexts); (ii) longitudinal designs tracking AI literacy, trust, and pedagogical practice over several years to test whether the governance mechanisms produce sustained equity gains; (iii) refinement of the inclusion indicators (Table III) into validated composite indices, and extension of the four gender-aware mechanisms—as a reusable template—to other equity axes such as socio-economic status, migrant or cultural background, and disability, ultimately through intersectional rather than single-axis analysis; and (iv) deeper study of human-AI interaction (trust, agency, decision-making), where AI shapes cognitive and emotional dimensions of learning in ways not yet well understood [21], alongside practical compliance tools for aligning institutional practice with the European AI Act.

#### VIII. CONCLUSION

This paper reframes Artificial Intelligence in education not as a purely technological innovation, but as a socio-technical transformation that reshapes access, participation, and governance within educational ecosystems. Because AI systems are embedded in structures of power, data, and institutional design, they call for governance models that are correspondingly multi-level and inclusive.

The dual framework proposed—INCLUDE-AI at the European level and G.I.U.L.I.A. at the regional level—addresses a persistent gap between normative principles and implementation by integrating infrastructure, data governance, algorithmic transparency, and pedagogy into a single architecture and pairing each dimension with a measurable indicator. The exploratory evidence indicates that AI can enhance learning outcomes while also amplifying inequalities unless inclusion is embedded by design rather than added as a corrective measure;

inclusion-by-design is thus the paper's central proposition. The G.I.U.L.I.A. pilot further illustrates the value of regional ecosystems as settings in which European strategies can be translated, tested, and refined, while the explicit treatment of gender as a structural dimension moves the discussion beyond generic notions of fairness toward context-sensitive, measurable inclusion.

The work aligns with emerging research positioning AI in education as a force shaping not only performance but also cognition, agency, and emotional engagement [21]. From a policy standpoint, the governance taxonomy (Table II) offers institutional leaders a structured mapping to align practice with evolving European regulation, including the AI Act; from a theoretical standpoint, the paper embeds inclusion, governance, and gender into socio-technical systems theory. The same architecture extends to conversational AI tools—chatbots and generative assistants—whose governance challenges (epistemic trust, bias amplification, and cognitive offloading) are increasingly salient within the EDUC-AI-TION community.

Ultimately, the trajectory of AI in education will depend less on technological sophistication alone than on the capacity to design systems that are transparent, accountable, and equitable—an outcome that, as recent work stresses, requires multi-stakeholder alignment among educators, students, and institutions [22] and integrated approaches combining ethical, technical, and institutional dimensions [23]. Rather than a finished paradigm, INCLUDE-AI and G.I.U.L.I.A. are offered as an operational starting point for the empirical validation and policy-oriented implementation of inclusive AI in education.

#### ACKNOWLEDGMENTS

This work was encouraged by Fondazione Del Monte di Bologna e Ravenna. The authors also acknowledge the contribution of local educational institutions and stakeholders involved in the G.I.U.L.I.A. pilot activities. Claude Sonnet (Anthropic) was used as an editorial assistant to enhance readability and ensure the absence of typographical errors in the final manuscript. All intellectual content, analysis, and argumentation remain the sole responsibility of the authors.

#### REFERENCES

- [1] I. M. García-López and L. Trujillo-Lián, "Ethical and regulatory challenges of generative AI in education: A systematic review," *Frontiers in Education*, vol. 10, 2025, doi: 10.3389/educ.2025.1565938.
- [2] J. F. Pane, E. D. Steiner, M. D. Baird, and L. S. Hamilton, "Informing progress: Insights on personalized learning implementation and effects," RAND Corporation, Tech. Rep. RR-2042-BMGF, 2017, doi: 10.7249/RR2042.
- [3] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability and Transparency (FAT\*)*, ser. Proceedings of Machine Learning Research, vol. 81, 2018, pp. 77–91.
- [4] World Bank, "Remote learning during COVID-19: Lessons from today, principles for tomorrow," World Bank Group, Tech. Rep., 2021.
- [5] European Commission, *Digital Education Action Plan (2021–2027)*. European Union, 2020, cOM(2020) 624 final.
- [6] W. Holmes, K. Porayska-Pomsta, and S. Nemorin, "The ethics of AI in education," in *Handbook of Artificial Intelligence in Education*, 1st ed., B. du Boulay, A. Mitrovic, and K. Yacef, Eds. Cheltenham, UK: Edward Elgar Publishing, 2023, pp. 571–604, doi: 10.4337/9781800375413.00038.
- [7] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education," in *The Ethics of Artificial Intelligence in Education*, W. Holmes and K. Porayska-Pomsta, Eds. Routledge, 2022, doi: 10.4324/9780429329067-10.
- [8] M. Bond, H. Khosravi, M. De Laat *et al.*, "A meta-systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 4, 2024, doi: 10.1186/s41239-023-00436-z.
- [9] L. Floridi, J. Cows, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.
- [10] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, 2016, doi: 10.1177/2053951716679679.
- [11] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing AI literacy: An exploratory review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100041, 2021, doi: 10.1016/j.caeai.2021.100041.
- [12] K. J. Davin, S. Kissau, H. Haudeck, B. Ade-Thurrow, D. Gómez-Pereira, P. T. Oon, and A. Dornburg, "Global and educational disparities in AI integration: A study of L2 teacher training and usage patterns," *System*, vol. 136, p. 103901, 2026, doi: 10.1016/j.system.2025.103901.
- [13] UNESCO, *AI Competency Framework for Teachers*. Paris: UNESCO, 2024.
- [14] OECD, *Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*. OECD Publishing, 2023, doi: 10.1787/c74f03de-en.
- [15] E. Rudolph, H. Seer, C. Mothes, and J. Albrecht, "Automated feedback generation in an intelligent tutoring system for counselor education," in *Proc. 19th Conf. Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 39, 2024, pp. 501–512, doi: 10.15439/2024F1649.
- [16] T. C. Freitas, M. J. V. Pereira, A. C. Neto, and P. R. Henriques, "Goliath, a programming exercises generator supported by AI," in *Proc. 19th Conf. Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 39, 2024, pp. 331–342, doi: 10.15439/2024F8479.
- [17] Y. Feldman-Maggor, M. Cukurova, C. Kent, and G. Alexandron, "The impact of explainable AI on teachers' trust and acceptance of AI EdTech recommendations: The power of domain-specific explanations," *International Journal of Artificial Intelligence in Education*, 2025, doi: 10.1007/s40593-025-00486-6.
- [18] OECD, *OECD Digital Education Outlook 2021*. OECD Publishing, 2021, doi: 10.1787/589b283f-en.
- [19] UNESCO, *Guidance for Generative AI in Education and Research*. UNESCO Publishing, 2023.
- [20] —, *AI and the Future of Education: Disruptions, Dilemmas and Directions*. UNESCO Publishing, 2025.
- [21] L. Favero, J. A. Pérez-Ortiz, T. Käser, and N. Oliver, "AI in education beyond learning outcomes: Cognition, agency, emotion, and ethics," *arXiv preprint arXiv:2602.04598*, 2026, doi: 10.48550/arXiv.2602.04598.
- [22] A. J. Karran, P. Charland, J.-T. Martineau, A. Ortiz de Guinea Lopez de Arana, A.-M. Lesage, S. Sénécal, and P.-M. Léger, "Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education," *arXiv preprint arXiv:2402.15027*, 2024, doi: 10.48550/arXiv.2402.15027.
- [23] P. Mahajan, "What is ethical: AIHED driving humans or human-driven AIHED? A conceptual framework enabling the ethos of AI-driven higher education," *arXiv preprint arXiv:2503.04751*, 2025, doi: 10.48550/arXiv.2503.04751.

# Predicting Centrality Measures in Complex Networks Utilizing Markov Chains

Anchal Gera\*, Jens Dörpinghaus<sup>†‡§¶</sup>, Robert Rockenfeller\*

\* Mathematical Institute, University of Koblenz, Germany

<sup>†</sup> Department of Computer Science, University of Koblenz, Germany

<sup>‡</sup> Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany

<sup>§</sup> Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden

<sup>¶</sup> University of Pretoria, South Africa

Email: anchalgera@uni-koblenz.de, jens.doerpinghaus@bibb.de, rrockenfeller@uni-koblenz.de

**Abstract**—Centrality measures are of critical importance for the identification of influential nodes in networks. However, classical metrics such as degree and betweenness are static and thus fail to capture temporal dynamics. The present paper introduces a probabilistic framework for modeling centrality as a stochastic process using Markov chains. By discretizing centrality values into states, transition probability matrices are learned from temporal data, enabling future centrality prediction. First-order and higher-order models were evaluated, with the results indicating that higher-order models improve accuracy, particularly for betweenness. The results of the study indicate that centrality is inherently dynamic and probabilistic. The proposed framework offers an interpretable method to forecast node importance and distinguish stable from volatile influencers in evolving networks.

## I. INTRODUCTION

NETWORKS are foundational components of complex systems in nature, technology, and society, influencing the dissemination of information, the propagation of diseases, and the distribution of resources. Network science is a field of study that utilizes centrality measures to analyze such systems, with these measures quantifying the importance of nodes within a network. It is evident that classical metrics, incorporating degree, betweenness, closeness, and eigenvector centrality, have yielded valuable insights into static network structures. However, in actual practice, networks are seldom static, for example social networks [1], [2], [3], [4], biomedical networks [5], [6], [7] or education networks [8]. Phenomena evolve, adapt and fluctuate over time, thus rendering traditional centrality measures insufficient for capturing the temporal and probabilistic nature of influence and interaction [9].

Observations of natural and engineered systems suggest that complex structures often evolve according to underlying patterns governed by physical laws or mathematical principles, often formulated in the form of differential equations [10]. Concurrently, a substantial proportion of these systems can be naturally represented as dynamic networks, where interactions between components undergo alteration over time. Examples of such phenomena include ant colonies forming trails from initially random movements, leaf venation emerging through local growth rules, and social systems organizing through repeated interactions [11], [12]. When observed through the

lens of a network, such systems demonstrate a tendency towards self-organization, albeit constrained by parameters that operate across specific temporal and spatial scales. Thus, the central question guiding this study is whether temporal measurements of network structure, governed by simple probabilistic rules, can provide a comprehensive explanation of interaction patterns and enable the prediction of future network evolution.

A significant gap in extant research is the limited ability to predict how network structures and node importance evolve over time. While a substantial body of research has focused on describing network snapshots, fewer studies have addressed the prediction of centrality dynamics in evolving networks, see [9] and the literature review. A key issue is the absence of flexible frameworks that integrate probabilistic modeling with centrality analysis to quantify influence and volatility in dynamic settings.

The present study addresses this gap by investigating centrality from a probabilistic perspective. Recent advances in the field of network analysis have seen the integration of Markov chains, and it is this foundation that the present study builds upon. The investigation of stochastic transition models as a means of capturing temporal dependencies in network evolution is the primary focus of this study.

## II. LITERATURE REVIEW

The study of complex networks has evolved into a central paradigm for understanding systems across disciplines, including sociology, biology, and computer science. Foundational concepts and terminology in network science are comprehensively introduced in the work of [13], which emphasizes networks as a universal framework for representing interacting entities. Historically, the field developed through interdisciplinary contributions, ranging from early sociological theories to formal mathematical modeling and computational analysis [14].

A core problem in network science is the identification of important or influential nodes. This has led to the development of centrality measures, which quantify node importance from different structural perspectives. Classical centrality measures

include degree, betweenness, closeness, and eigenvector centrality, each capturing a distinct notion of influence [15]. Degree centrality, one of the earliest and simplest measures, reflects the number of direct connections of a node and was formalized by [16]. Betweenness centrality captures the extent to which a node lies on shortest paths between other nodes, thereby measuring its role as an intermediary [17]. Closeness centrality quantifies how close a node is to all others in terms of shortest path distances, while eigenvector centrality assigns importance based on connections to other highly connected nodes [15].

While these measures have been widely applied, they are fundamentally limited to static network representations. Real-world systems, however, are inherently dynamic, with network structures evolving over time. Recent work highlights that applying static centrality measures to evolving networks can lead to significant inaccuracies and misinterpretations [9]. This has motivated the development of temporal network analysis, where networks are represented as sequences of time-dependent snapshots.

Early approaches to dynamic networks extended static centrality measures by computing them independently at each time step. Although straightforward, this approach neglects temporal dependencies and fails to capture the continuity of node importance. More advanced methods incorporate time explicitly into centrality definitions. For instance, dynamic PageRank models treat centrality as a time-dependent dynamical system driven by evolving transition matrices [18]. These approaches enable faster detection of emerging influential nodes by emphasizing recent structural changes.

Parallel to temporal extensions, probabilistic approaches have gained importance in network analysis. Probabilistic graphical models, such as Bayesian networks, provide a principled framework for representing uncertainty and conditional dependencies between nodes [19]. These models allow for robust inference in the presence of noisy or incomplete data and enable the identification of statistically significant network structures.

Dynamic Probabilistic Networks (DPNs) extend these ideas to temporal settings by modeling network evolution as a stochastic process under the Markov assumption [20]. In this framework, the state of the network at time  $t + 1$  depends only on its state at time  $t$ , allowing for tractable learning and inference. Such models have been successfully applied in domains such as biological network analysis and social interaction modeling.

Markov chains provide a particularly powerful tool for modeling stochastic processes on networks. In this setting, transitions between nodes are governed by a probability matrix, and long-term behavior is characterized by a stationary distribution. This perspective forms the basis of PageRank and related centrality measures. Research has shown that in certain network structures, particularly scale-free networks, some nodes exhibit remarkable stability in their rankings despite structural perturbations [21].

Recent work has extended Markov-based models to cap-

ture temporal dynamics and improve predictive performance. Temporal PageRank incorporates time-dependent transition matrices, enabling the modeling of evolving influence patterns [18]. Furthermore, higher-order Markov models have been proposed to account for temporal dependencies beyond the memoryless assumption, demonstrating improved accuracy in predicting centrality trajectories in dynamic networks [22].

Notwithstanding these advances, substantial gaps persist within the relevant literature. The majority of existing approaches concentrate on describing centrality at specific time points rather than predicting its evolution. Furthermore, there is a lack of integration between probabilistic modeling and centrality analysis within a unified framework. Finally, centrality is frequently regarded as a deterministic quantity, despite its inherently stochastic nature in dynamic environments.

The present paper builds upon these lines of research by introducing a probabilistic framework for centrality analysis in evolving networks. The objective of this study is to integrate the principles of Markov chain modeling with the utilization of discretized centrality trajectories. This approach is intended to facilitate the capture of both the temporal dynamics and the associated uncertainty of node importance. Furthermore, a flexible weighted centrality measure is proposed, allowing customization based on application-specific priorities. This enables the identification of both stable and volatile nodes. This approach contributes to a deeper understanding of how influence emerges and evolves in complex, dynamic systems.

### III. METHODOLOGICAL BACKGROUND

The incorporation of probability theory into complex network analysis has been shown to have a significant impact on our capacity to model uncertainty, characterize random structural evolution, and quantify dynamic interdependence among nodes on a large scale, see [19], [20]. Whilst classical graph theory is typically predicated on deterministic assumptions, probabilistic frameworks explicitly acknowledge the stochastic nature of real-world networks. This paradigm shift enables researchers to move beyond static representations and address three central challenges:

- 1) Modeling incomplete or noisy observational data through uncertainty quantification,
- 2) Capturing temporal restructuring mechanisms in evolving networks,
- 3) Understanding how local stochastic interaction rules give rise to emergent global patterns.

At the core of probabilistic network modeling lies the conceptualization of networks as probabilistic graphical models. Bayesian networks provide a principled framework for representing conditional dependencies between nodes, while avoiding overfitting through likelihood-based model selection [19]. In the context of a network comprising random variables  $X_1, \dots, X_n$ , the joint probability distribution can be factorized as follows:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pa}(X_i)),$$

where  $\text{Pa}(X_i)$  denotes the parent set of node  $X_i$ . This factorization induces sparse dependency structures that exclude redundant correlations and improve interpretability [19].

Probabilistic formulations are particularly powerful for dynamic networks through *Dynamic Probabilistic Networks* (DPNs) [20]. These extend the Markov assumption to the temporal domain:

$$P(X[t+1] \mid X[0], \dots, X[t]) = P(X[t+1] \mid X[t]),$$

where temporal transitions are governed by a time-invariant dependency structure. Learning such models from data involves maximizing a Bayesian score

$$\text{Score}(G) = \log P(D \mid G) + \log P(G),$$

where  $D$  represents observed temporal snapshots and  $P(G)$  encodes prior structural knowledge [20]. This approach enables the joint inference of static dependencies and dynamic transition rules, with applications ranging from biological pathway modeling to social contagion analysis.

Centrality measures undergo a fundamental reinterpretation under probabilistic network models. Consider Katz centrality,

$$C_i = \sum_{k=1}^{\infty} \sum_{j=1}^n \alpha^k (A^k)_{ji},$$

where  $\alpha$  is a damping factor controlling the contribution of longer paths [23]. In random networks,  $C_i$  becomes a random variable whose distribution depends on the degree distribution  $p(k)$ . For sparse, locally tree-like networks, cavity methods yield the approximation

$$P_C(c) \approx \sum_{k=0}^{\infty} p(k) \delta\left(c - \frac{1}{1 - \alpha k \mathbb{E}[C]}\right),$$

revealing a multimodal structure in which each mode corresponds to a distinct degree class [23]. This result explains empirically observed correlations between degree and centrality, as nodes of different degrees occupy distinct probabilistic centrality regimes.

Probabilistic network theory therefore provides a powerful analytical lens, enabling robust uncertainty quantification, dynamic modeling, and improved interpretability across domains ranging from protein interaction networks to resilient infrastructure design. This theoretical framework provides us with one more idea that is "Influence is uncertain and evolves probabilistically".

#### A. Markov Chains and Transition Matrices in Network Centrality

Markov chain theory offers a mathematically rigorous framework for modeling stochastic transitions and predicting node importance in evolving networks. In this setting, network navigation is represented as a memoryless random walk governed by a transition matrix  $P$ , where  $P_{uv}$  denotes the probability of transitioning from node  $u$  to node  $v$  in one

step [24]. For a network with adjacency matrix  $A$ , a common construction is

$$P_{uv} = \frac{A_{uv}}{k_u},$$

where  $k_u$  is the out-degree of node  $u$ , ensuring row-stochasticity  $\sum_v P_{uv} = 1$  [24], [21].

The stationary distribution  $\pi$ , satisfying  $\pi = \pi P$ , forms the theoretical basis of PageRank centrality. Introducing a teleportation parameter  $\alpha$  yields

$$\pi = (1 - \alpha)v + \alpha\pi P,$$

where  $v$  is the teleportation distribution [18]. In scale-free networks, this formulation gives rise to *super-stable* nodes whose rankings remain robust under topological perturbations due to their disproportionate influence on  $\pi$  [21]. A complementary global connectivity metric is the Kemeny constant [25],

$$K = \sum_{j=2}^n \frac{1}{1 - \lambda_j},$$

where  $\lambda_j$  are the non-principal eigenvalues of  $P$ . Lower values of  $K$  correspond to faster mixing and more efficient information diffusion [24].

In dynamic networks, the transition matrix becomes time-dependent, leading to non-homogeneous Markov chains [18]. Temporal PageRank models this evolution via the dynamical system

$$x'(t) = (1 - \alpha)v(t) - (I - \alpha P(t))x(t),$$

where  $x(t)$  represents instantaneous centrality. Empirical evaluation on Wikipedia traffic data shows that this approach detects emerging trends substantially faster than static PageRank by emphasizing recent transitions [18].

Prediction of future centrality values can be achieved using higher-order Markov chains that condition transitions on multi-step histories [22]. For a centrality time series  $C_t$ , a  $K$ -order model estimates

$$P(C_{t+1} \mid C_t, \dots, C_{t-K}) = \prod_{k=0}^K P(C_{t+1} \mid C_{t-k}),$$

with parameters learned via maximum likelihood estimation. Applied to mobile social networks, this approach achieves high prediction accuracy by capturing periodic interaction patterns, see [22].

Recent work integrates Markov transitions with graph neural networks by learning transition matrices  $P_\theta$  parameterized by node features [26]. The Transition Probability Matrix (TPM) method computes node representations as

$$\phi(v) = \text{MLP}\left(\sum_{w \in N(v)} P_\theta(v, w) h_w\right),$$

where  $h_w$  denotes node embeddings. This hybrid approach demonstrates improved performance over classical Graph Convolutional Networks (GCNs) in cross-network

classification tasks and Here, MLP denotes a multi-layer perceptron that applies a nonlinear transformation to the Markov-weighted neighborhood aggregation, enabling expressive node representations beyond linear diffusion [26].

Finally, dynamic-sensitive centrality analysis shows that temporal correlation between consecutive snapshots governs prediction error bounds [27]:

$$\epsilon \leq \frac{1-\rho}{\sqrt{T}} \sum_{t=1}^T \|P_t - P_{t-1}\|_F,$$

where  $\rho$  is the temporal correlation coefficient and  $T$  is the observation window. High correlation ensures robustness even under substantial missing data, highlighting the suitability of Markov-based approaches for real-world temporal networks such as epidemic contact graphs and IoT systems.

#### B. Upper and lower Bounds for Centrality Measures

Several upper and lower bounds for centrality measures have been proposed in literature, see [28], [29]. Here, the authors defined  $\mathfrak{G}^p = \{G_1, \dots, G_\ell\}$  as a series of graphs and  $p \in \mathbb{R}$  with  $0 \leq p \leq 1$  and

$$|(V(G_i) \cup V(G_{i+1})) \setminus (V(G_i) \cap V(G_{i+1}))| \leq p|V(G_i)|,$$

$$|(E(G_i) \cup E(G_{i+1})) \setminus (E(G_i) \cap E(G_{i+1}))| \leq p|E(G_i)|,$$

for  $i \in \{1, \dots, \ell-1\}$ . As  $\mathfrak{G}^p$  now is a series of graphs with a fixed set of differences and changes from one to the other, the authors proposed approximations of the changes over time, or the error in the centrality measures that can occur due to these changes. Unless otherwise noted, we will consider  $\mathfrak{G}^p = \{G_1, \dots, G_\ell\}$ .

**Theorem III.1.** *Let  $i \in \{1, \dots, \ell-1\}$  so that  $v \in V(G_i)$  and  $v \in V(G_{i+1})$ . Then it holds that*

$$dc^{i+1}(v) \geq \frac{d^i(v) - p|V(G_i)|}{|V(G_i)| - 1 + p|V(G_i)|},$$

$$dc^{i+1}(v) \leq \frac{d^i(v) + p|V(G_i)|}{|V(G_i)| - 1 - p|V(G_i)|}.$$

For betweenness centrality, we define

$$\sigma = |N(G_i)|p,$$

$$\epsilon = \min\{D(G_i)^2, 2|V(G_i)|p\},$$

where  $D(G)$  is the diameter of  $G$ . [28] provided two lemmata to obtain a bound for  $bc^{i+1}(v)$  for  $v \in V$ .

**Lemma III.2.** *Let  $i \in \{1, \dots, \ell-1\}$  so that  $v \in V(G_i)$  and  $v \in V(G_{i+1})$ . Then,*

$$P_v(k, j) \frac{1}{\sigma} \leq P_v^{i+1}(k, j)$$

*holds.*

**Lemma III.3.** *Let  $i \in \{1, \dots, \ell-1\}$  so that  $v \in V(G_i)$  and  $v \in V(G_{i+1})$ . Then,*

$$P_v^{i+1}(k, j) \leq \begin{cases} P_v(k, j)\epsilon & P_v(k, j) > 0 \\ D(G_i)^2\epsilon & P_v(k, j) = 0 \end{cases}$$

*holds.*

**Theorem III.4.** *Let  $i \in \{1, \dots, \ell-1\}$  so that  $v \in V(G_i)$  and  $v \in V(G_{i+1})$ . Then,*

$$bc^i(v)\epsilon \leq bc^{i+1}(v) \leq bc^i(v) \frac{1}{\sigma}$$

*holds.*

For details and proofs we refer to [28], [29]. Clearly, this notation is equivalent to  $\{G_t = (V_t, E_t)\}_{t=1}^T$  as a sequence of network snapshots over  $T$  discrete time steps, see Section IV-B.

## IV. METHOD

### A. Generation of Networks

All network simulations are implemented in Python using the *NetworkX* library, which provides data structures and algorithms for constructing, manipulating, and analyzing complex networks [13]. In order to systematically study the effects of topology on centrality dynamics, a range of canonical network models must first be generated. These models will serve as controlled environments for the analyses to be conducted.

Random networks are constructed using the Erdős-Rényi model  $G(n, p)$ . In this model, each of the nodes,  $n$  in number, is connected independently with probability  $p$ . This process yields homogeneous degree distributions characterized by minimal clustering and the absence of intrinsic hubs. Consequently, a neutral baseline is established, providing a reference point for the subsequent comparison of more structured topologies. Scale-free networks are generated via the Barabási-Albert preferential attachment mechanism, whereby new nodes attach to existing nodes with probability proportional to their degree. This phenomenon gives rise to heterogeneous topologies comprising a small number of high-degree hubs, which mirrors numerous real-world systems, including social and citation networks. Small-world networks are generated using the Watts-Strogatz model, which begins with a regular lattice and randomly rewires edges with probability  $p_{\text{rewire}}$ . This process produces high local clustering while maintaining short average path lengths. Such networks have been shown to capture structural features that are relevant to social and biological systems, where influence can rapidly propagate within and across communities.

To model temporal evolution, we construct dynamic networks as sequences of snapshots, denoted by  $G_t = (V_t, E_t)$ . Each simulation commences with a modestly connected seed graph, which is generally initiated using an Erdős-Rényi topology. At each discrete time step, a new node is added and connected to a predefined number of existing nodes, sampled from a distribution. Concurrently, a limited number of edges can undergo removal or reconfiguration to emulate network

TABLE I: Experiment 1 parameters for synthetic network generation

| Parameter                           | Value                                    |
|-------------------------------------|------------------------------------------|
| Initial nodes ( $n_{\text{init}}$ ) | 5                                        |
| Initial graph model                 | Erdős–Rényi ( $p = 0.5$ )                |
| Total snapshots ( $T$ )             | 150                                      |
| Attachments per new node ( $m$ )    | Uniform(2, 5)                            |
| Detachments per step                | Uniform(1, 3)                            |
| Discretization bins                 | 5                                        |
| Bin edges (degree centrality)       | [0.00, 0.01, 0.02, 0.03, 0.04, 0.05]     |
| State labels                        | {Very Low, Low, Medium, High, Very High} |
| Random seed                         | 42                                       |

churn and adaptation. This procedure produces evolving network structures suitable for analyzing centrality trajectories and for probabilistic modeling of influence. Extensions to multilayer networks, where multiple types of edges coexist over the same node set, are also supported to capture multiplex interactions.

For all simulations, parameters such as the initial node count, attachment range, and total time steps are recorded and fixed to ensure reproducibility. This framework provides a flexible and mathematically precise foundation for subsequent analysis of centrality and influence in evolving networks. See Table I for chosen parameters.

#### B. Dynamic and Adaptive Network Simulation

To study influence and centrality in evolving networks, we extend static network models into dynamic, time-varying structures. Let  $\{G_t = (V_t, E_t)\}_{t=1}^T$  denote a sequence of network snapshots over  $T$  discrete time steps. Each  $G_t$  represents the state of the network at time  $t$ , including the current set of nodes  $V_t$  and edges  $E_t$ .

Simulations begin with a small connected seed graph  $G_1$ , typically generated as an Erdős–Rényi random network to ensure baseline connectivity and reproducibility. At each subsequent time step, network evolution is driven by both growth and adaptation mechanisms. Growth is modeled by the addition of new nodes, where each new node  $v_{\text{new}}$  connects to a set of  $m$  existing nodes sampled according to a specified attachment criterion, which may be uniform, degree-based, or guided by composite measures. This process captures preferential attachment and mimics the emergence of hubs and local leaders over time.

Network adaptation is incorporated through edge addition, deletion, or rewiring. For each node, a small number of edges may be removed or redirected, and new connections may form with nodes in its local neighborhood. Formally, if  $E_t$  is the edge set at time  $t$ , the updated edge set at  $t + 1$  is

$$E_{t+1} = (E_t \setminus E_{\text{del}}) \cup E_{\text{add}},$$

where  $E_{\text{del}}$  and  $E_{\text{add}}$  denote edges removed and added, respectively. The rewiring and attachment probabilities can be uniform or weighted by node centrality, trust accumulation, or proximity to influential hubs, allowing the network to adapt in ways that reflect social, biological, or technological dynamics.

At each time step, a snapshot  $G_t$  is recorded, enabling the computation of node-level statistics such as degree, betweenness, closeness, and eigenvector centrality. This framework

generates networks that are both temporally coherent and structurally adaptive, forming the foundation for subsequent modeling of centrality trajectories and Markov chain-based influence prediction. By explicitly controlling growth and adaptation parameters, we can systematically investigate how local and global network dynamics affect the evolution of node importance over time.

*a) Selection of Centrality Measures:* The present study begins its analysis of node importance and prediction using a set of four fundamental and widely studied centrality measures: degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. These measures are employed as baseline indicators of node importance, providing interpretable and complementary perspectives on network structure. The utilization of classical centrality measures facilitates a straightforward comparison and serves as a foundational framework for assessing the added value of more complex approaches.

- 1) Degree centrality measures the relative number of connections of a node and captures its immediate local influence. In our implementation, degree centrality is computed using the `nx.degree_centrality()` function from the NetworkX library in Python, which normalizes the node degree by the maximum possible number of connections in the network [15].
- 2) Betweenness centrality quantifies the extent to which a node acts as an intermediary by lying on shortest paths between other node pairs. Nodes with high betweenness centrality often play a crucial role in controlling information flow. We compute this measure using the `nx.betweenness_centrality()` function [17].
- 3) Closeness centrality reflects how efficiently a node can reach all other nodes in the network, based on the average length of shortest paths originating from it. This metric captures the notion of accessibility and speed of information spread and is computed using `nx.closeness_centrality()` [15].
- 4) Eigenvector centrality extends the concept of degree centrality by accounting not only for the number of connections a node has, but also for the importance of its neighbors. Nodes connected to other highly central nodes receive higher scores, capturing a notion of global influence within the network. In our analysis, eigenvector centrality is computed using the `nx.eigenvector_centrality()` function [30].

Collectively, these four centrality measures furnish a balanced and well-established toolkit for analyzing node importance. The computational availability, interpretability, and theoretical significance of these models renders them suitable for initial prediction tasks. In the following sections, we build upon these classical measures by introducing dynamic, probabilistic, and temporal centrality frameworks that are better suited for evolving networks and influence prediction.

### C. Using Markov Chains to Predict Centrality

In this section, the focus is on the predictive modeling of node centrality trajectories within dynamic networks. The central idea is to leverage Markov chains as a probabilistic framework to forecast the future state of a node's importance, based on its historical behavior. To this end, an end-to-end experimental pipeline was developed, integrating dynamic network generation, centrality computation, discretizations of continuous measures, transition probability estimation, and predictive evaluation.

*a) Discretization of Centrality Trajectories:* Since centrality measures are continuous-valued, direct application of Markov chain models is not feasible. To address this, we discretize centrality trajectories into a finite set of qualitative states (e.g. *Very Low*, *Low*, *Medium*, *High*, *Very High*). This transformation enables categorical analysis and allows the evolution of a node's importance to be interpreted as transitions between symbolic states.

Formally, given a sequence of centrality values  $\{c_t\}_{t=1}^T$  for a node over  $T$  time steps, we partition the range of values into bins  $B = \{b_1, b_2, \dots, b_k\}$ . Each observed centrality value is assigned to the corresponding bin, producing a sequence of discrete states  $S = \{s_1, s_2, \dots, s_T\}$  where  $s_t \in \{1, \dots, k\}$ . This sequence serves as the input to our Markov chain model.

To model temporal evolution using Markov chains, we transform continuous centrality values into a small set of discrete states (labels). The default configuration uses 5 bins:

Labels = {Very Low, Low, Medium, High, Very High},

and default bin edges [0.00, 0.01, 0.02, 0.03, 0.04, 0.05]. The discretization function maps a numeric centrality  $x$  to the bin index  $j$  satisfying  $b_j \leq x < b_{j+1}$  (last bin inclusive). Formally, for a centrality sequence  $\{x_t\}$  of node  $i$ , we define the state sequence

$$s_i = (s_i^1, s_i^2, \dots, s_i^T), \quad s_i^t \in \{\ell_1, \dots, \ell_5\}.$$

Alternative binning strategies are supported:

- Uniform bins on theoretical bounds: compute lower and upper bounds for centralities and set edges with `linspace(min, max, k + 1)`.

*b) Markov Chain Modeling and Prediction:*

*c) First-order Markov model.:* From the discretized state sequences  $\{s_i\}$ , we compute first-order transition counts

$$C_{ab} = \#\{(t, i) : s_i^t = a, s_i^{t+1} = b\},$$

and then transition probabilities

$$T_{ab} = \Pr(s^{t+1} = b \mid s^t = a) = \frac{C_{ab}}{\sum_{b'} C_{ab'}}. \quad (1)$$

To predict the next state given current state  $a$ , we either sample from the categorical distribution defined by row  $T_a$ . (stochastic prediction), or choose  $\arg \max_b T_{ab}$  (deterministic prediction).

*d) Markov chain construction.:* Using the discretized sequences, we construct a first-order Markov chain for each centrality measure. The transition probability matrix  $P$  is estimated empirically as

$$P_{ij} = \frac{N_{ij}}{\sum_j N_{ij}},$$

where  $N_{ij}$  denotes the number of observed transitions from state  $i$  to state  $j$  across all nodes and time steps. Each row of  $P$  thus represents a probability distribution over the possible next states, conditioned on the current state. This model encodes the stochastic dynamics of centrality evolution.

*e) Centrality Computation and Storage:* For each generated or ingested snapshot  $G_t$ , we compute the following node-level statistics using NetworkX functions:

$$\text{deg}_t(i), \quad \text{btw}_t(i), \quad \text{clo}_t(i), \quad \text{eig}_t(i).$$

Results are stored as lists of dictionaries (one dictionary per snapshot mapping node to score), e.g. `degree_matrix[t][i] = deg_t(i)`. This sparse mapping reduces memory compared to full matrices when nodes are created over time.

*f) Centrality Bounds and Adaptive Binning:* The code contains functions to estimate conservative upper and lower bounds for degree and betweenness centralities per node. Bounds can be used to:

- derive dataset-adaptive bin edges (e.g. `linspace(min_bound, max_bound, k + 1)`);
- detect out-of-range centrality values during simulation;
- reason about theoretical limits.

Computing bounds as discussed in Section III-B improves robustness of the discretization step for networks with very different sizes or growth dynamics.

*g) Extensibility and Higher-Order Models:* While our initial experiments focus on degree centrality and first-order Markov chains, the methodology is readily extensible in two directions:

- 1) additional centrality measures (betweenness, closeness, eigenvector) can be modeled within the same framework;
- 2) higher-order Markov models can be employed to capture memory effects.

*h) k-th order Markov model.:* To capture richer temporal dependencies, we build a  $k$ -th order Markov model by aggregating transition tuples

$$((s^{t-k+1}, \dots, s^t), s^{t+1}).$$

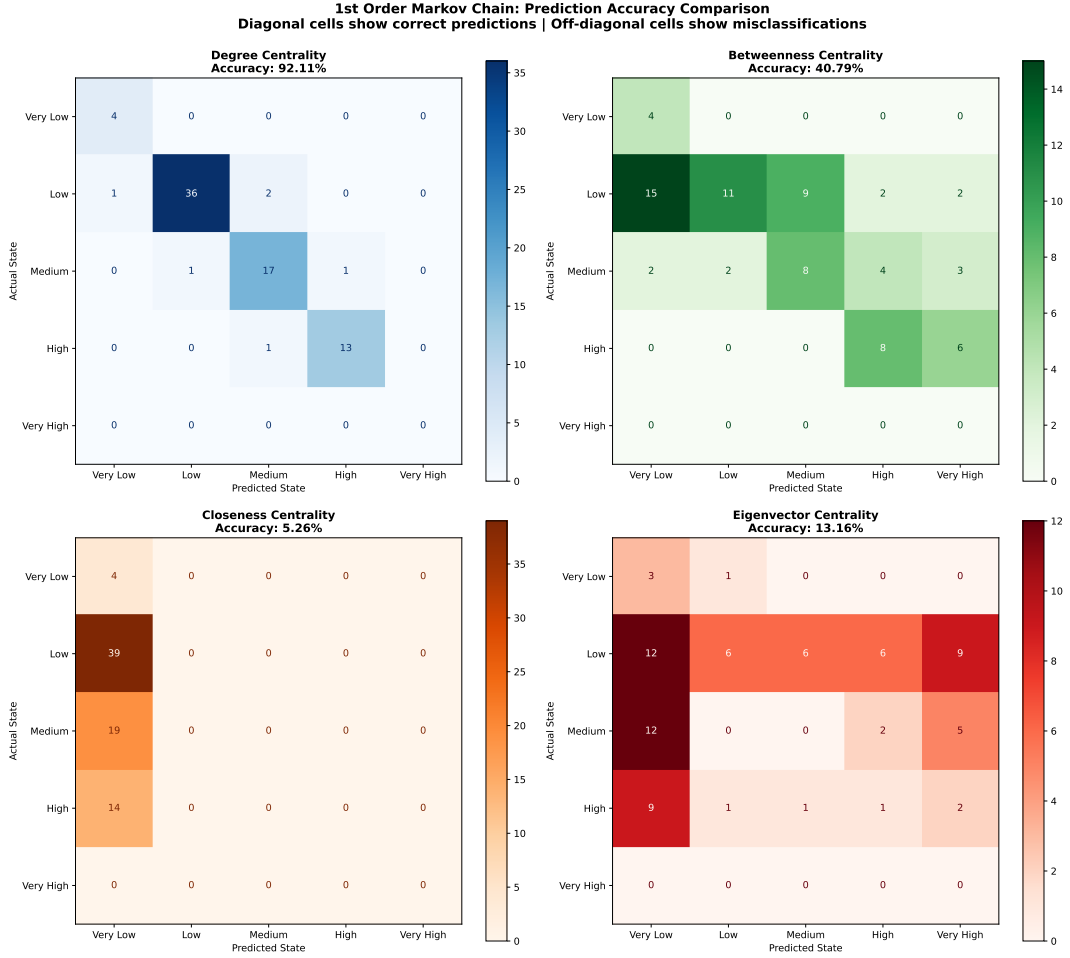


Fig. 1: Confusion matrices for first order Markov predictions of discretized centrality states.

TABLE II: Table showing the evaluation metrics

|                           |    | $F_1$  | Accuracy |
|---------------------------|----|--------|----------|
| First order Markov Model  | DC | 0.9589 | 0.9211   |
|                           | BC | 0.6739 | 0.5082   |
|                           | CC | 0.1    | 0.0526   |
|                           | EC | 0.2325 | 0.1315   |
| Second order Markov Model | DC | 0.9792 | 0.9594   |
|                           | BC | 0.9596 | 0.9225   |
|                           | CC | 0.9951 | 0.9904   |
|                           | EC | 0.9699 | 0.9416   |

## V. EXPERIMENTAL RESULTS

### A. First Order Markov Chain Prediction Performance

Prior to the interpretation of the results, it is essential to comprehend the confusion matrix. A confusion matrix is a tabular representation used to evaluate the performance of a classification model by comparing the predicted class labels with the true labels. Each row in the data corresponds to the actual class, while each column represents the predicted class. Predictions that are accurate are indicated by entries along the main diagonal of the matrix, whereas misclassifications are

indicated by off-diagonal entries. The confusion matrix is a data structure that provides a concise overview of the outcomes of a classification process. It facilitates the calculation of performance metrics, including accuracy, precision, recall, and the  $F_1$ -score. As recall was uniformly 1 across all classes and models, Table II reports only accuracy and the  $F_1$ -score.

Each centrality time series was discretized into five bins: Very Low, Low, Medium, High, and Very High. A first-order Markov chain was trained to predict the next state of each node. As illustrated in Figure 1, the resulting confusion matrices are presented.

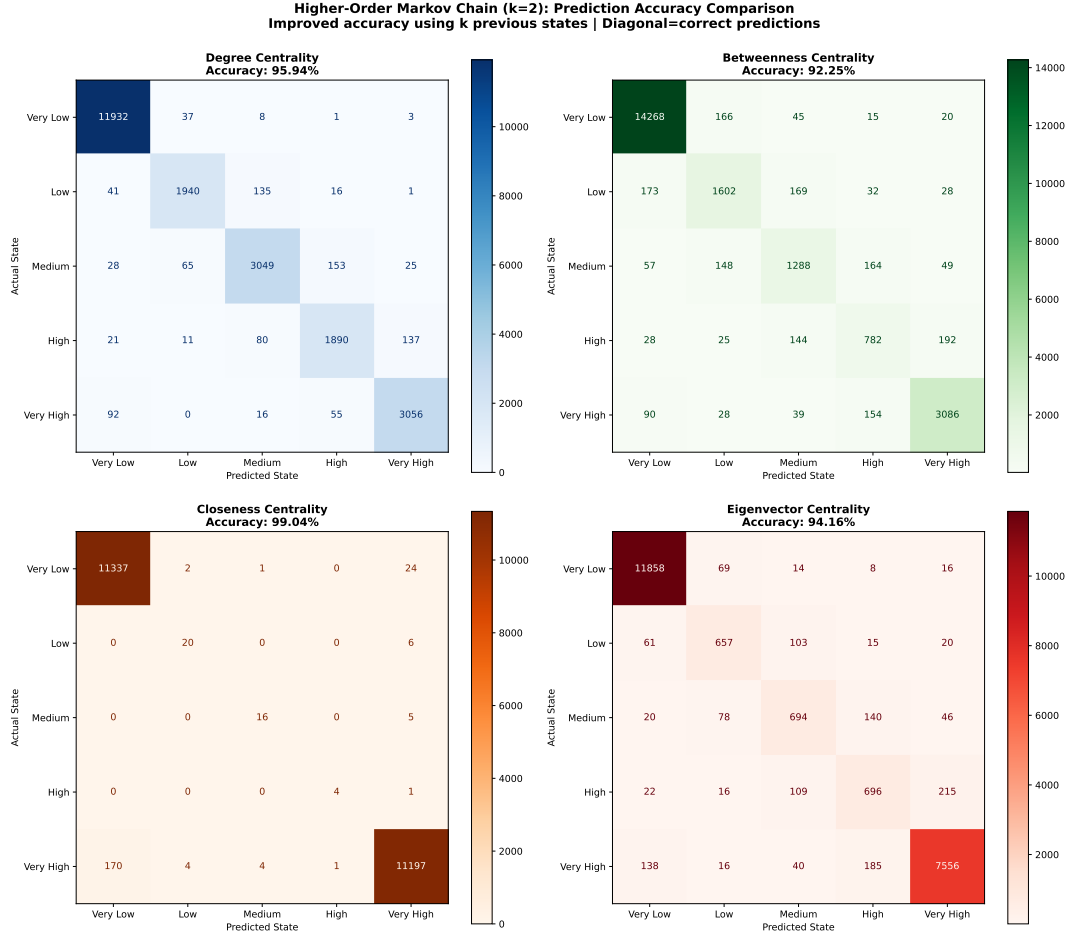


Fig. 2: Confusion matrices for second order Markov predictions with  $k = 2$ .

The confusion matrix reveals a strong diagonal dominance, particularly evident in the Low, Medium, and High categories, with respect to degree centrality. This finding suggests that the degree of evolution occurs in a relatively smooth and predictable manner. Misclassifications primarily occur between adjacent bins, such as Low and Very Low or Medium and Low, reflecting gradual changes in connectivity rather than abrupt transitions. As indicated by the findings, betweenness centrality demonstrates a conspicuously diminished diagonal structure, with a considerable proportion of high and medium states being erroneously classified as low. This phenomenon is indicative of the highly volatile nature of betweenness, where nodes can rapidly gain or lose brokerage roles due to the formation or deletion of shortest paths.

Closeness centrality demonstrates a pronounced bias toward the Very Low bin, suggesting that as the network expands, the majority of nodes experience an increase in distance from one another. The model accurately forecasts this collapse, yielding high apparent accuracy but low informational diversity. Eigenvector centrality demonstrates moderate predictability, with substantial ambiguity between Medium, High, and Very High states. This phenomenon underscores the heightened sensitiv-

ity of eigenvector centrality to subtle topological alterations in influential subgraphs.

### B. Nodewise Prediction at the Final Timestep

Figure 3 compares the actual and predicted degree centrality bins for all nodes at the final timestep.

The majority of nodes are accurately classified into Low or Medium states, which predominate the distribution at late times. Errors have been found to be concentrated among nodes with borderline values, where minute numerical disparities lead to transitions between adjacent bins. The Markov model has been demonstrated to effectively capture the global stratification of the network, effectively distinguishing between hubs and peripheral nodes, even in large-scale networks.

### C. Higher Order Markov Models

The second-order Markov model is predicated on the assumption that two previous states are incorporated to predict the next state. As illustrated in Figure 2 and Table II, the metrics provide a quantitative representation of the data.

Across all centralities, diagonal dominance is found to be significantly stronger than in the first order case. The

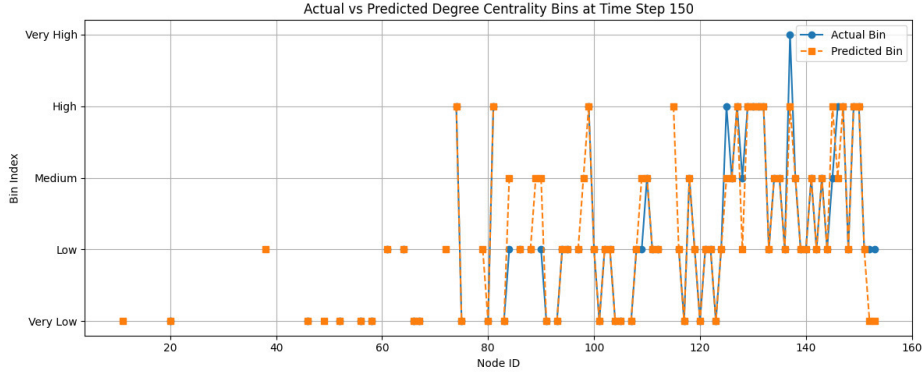


Fig. 3: Actual versus predicted degree centrality bins for all nodes at timestep 150.

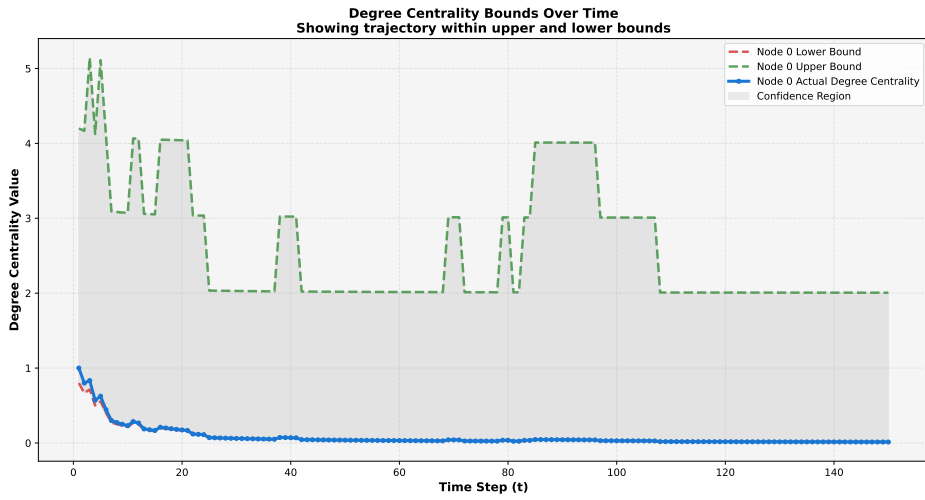


Fig. 4: Degree centrality of Node 0 with theoretical upper and lower bounds over time.

degree and closeness centrality indices achieve near-perfect classification, while the eigenvector and betweenness indices also improve substantially. This finding serves to substantiate the hypothesis that node importance is endowed with short-term temporal memory and that the incorporation of past states significantly enhances predictability.

#### D. Degree Centrality Bounds and Structural Stability

Figure 4 shows the degree centrality of Node 0, along with analytically derived upper and lower bounds.

The observed degree centrality remains strictly within the theoretical envelope throughout the simulation. In the nascent stages of network evolution, both the upper bound and the actual centrality are high, reflecting hub dominance. As the network expands, both bounds and actual values approach zero, thereby substantiating the notion that the decay in normalized degree is a structural consequence of network growth, rather than a modeling artifact.

#### E. Betweenness Centrality Bounds

Figure 5 displays betweenness centrality and its theoretical bounds at the final timestep.

All observed values lie far below the maximum theoretical bound, showing that although the network contains hubs, it does not collapse into a single dominating bridge. This reflects the creation of multiple alternative shortest paths as the network grows.

## VI. DISCUSSION AND OUTLOOK

The present work has introduced a probabilistic framework for the analysis and prediction of node centrality in evolving networks. In the context of classical network models, the centrality measures—including degree, betweenness, closeness, and eigenvector centrality—were computed at each temporal interval. These measures were subsequently transformed into discrete state trajectories. This discretization enabled the formulation of centrality evolution as a stochastic process, which was modeled using first- and higher-order Markov chains. The transition probability matrices were estimated through empirical means, and the predictive performance was evaluated through the use of classification-based metrics and confusion matrices. The findings indicate that centrality dynamics manifest structured temporal dependencies, and that incorporating

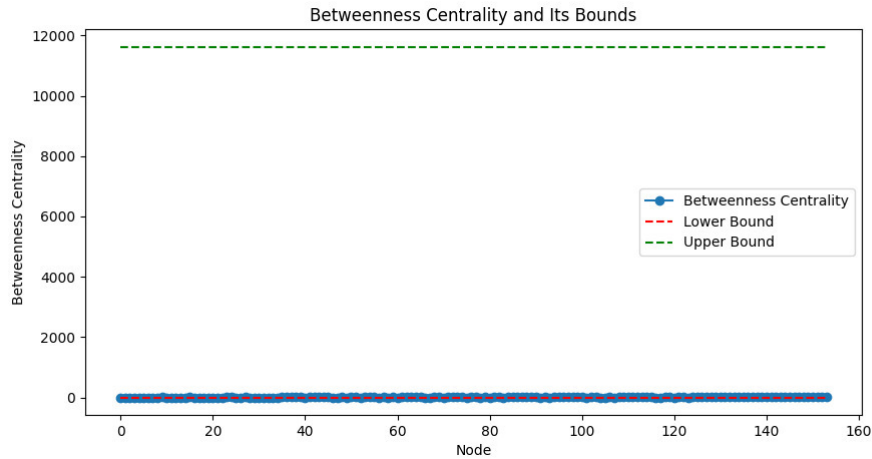


Fig. 5: Betweenness centrality with lower and upper bounds at the final timestep.

short-term memory through higher-order models significantly enhances predictive accuracy across all considered metrics.

A principal strength of the proposed approach is its interpretability. By mapping continuous centrality values to a finite set of states, the framework provides an intuitive description of how node importance evolves, while still retaining a probabilistic foundation. The employment of Markov chains enables the lucid modeling of transition dynamics and expedites the direct interpretation of persistence, volatility, and drift in node influence. The experimental design, predicated on the controlled synthetic network generation, ensures reproducibility and enables systematic investigation of how different topological mechanisms affect centrality evolution. The observed consistency between empirical findings and theoretical predictions, including the limited variability of centrality measures and the gradual progression of degree centrality, further reinforces the approach's validity.

However, it is imperative to acknowledge several limitations inherent to this approach. The discretization of centrality values, while necessary for Markov modeling, introduces a loss of information and may obscure fine-grained variations, particularly near bin boundaries. The model's reliance on first- and second-order Markov assumptions imposes limitations on its capacity to accommodate long-range temporal dependencies, which are likely to be present in complex real-world systems. Furthermore, the simulations are grounded in stylized network models, such as Erdős-Rényi and preferential attachment mechanisms. While these models have been thoroughly validated, they may not fully capture the heterogeneity and domain-specific constraints inherent to empirical networks. Another limitation pertains to the assumption of stationarity in the transition probabilities. However, real networks may exhibit non-stationary dynamics due to external shocks or regime changes. Finally, while the empirical results indicate strong predictive performance, a comprehensive theoretical

analysis of error bounds and convergence properties remains an open problem.

These limitations suggest several promising avenues for future research. One natural extension of this work is the development of non-Markovian or higher-memory models that can capture longer temporal dependencies. These models can be developed through variable-order Markov processes or sequence modeling techniques. Another significant avenue involves the integration of continuous-state models, which would circumvent discretization and enable centrality to be regarded as a fully continuous stochastic process. The incorporation of non-linear relationships between different centrality measures signifies a progression toward realism, as influence in numerous systems emanates from complex interactions rather than linear aggregation. Furthermore, the implementation of the proposed framework in real-world datasets, including social, biological, and technological networks, would offer significant insights into its practical utility and generalizability. Finally, the integration of probabilistic network modeling with contemporary machine learning methodologies, such as graph neural networks, has the potential to unify interpretability with high expressive capacity. This, in turn, promises to enhance our understanding of how influence emerges and evolves in complex dynamic systems.

In summary, the present study demonstrates that centrality is not a static property; rather, it is a dynamic and inherently stochastic quantity that could be effectively modeled and predicted within a probabilistic framework. The work combines network simulation, centrality analysis, and Markov chain modeling to provide a coherent approach to understanding and forecasting node importance. While further theoretical and empirical developments are required, the results highlight the potential of probabilistic methods to advance the study of dynamic networks and to uncover the temporal structure underlying complex systems.

## ACKNOWLEDGMENTS

During the preparation of this manuscript, the authors used DeepL to check the language and grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

## REFERENCES

- [1] V. Sekara, A. Stopczynski, and S. Lehmann, "Fundamental structures of dynamic social networks," *Proceedings of the national academy of sciences*, vol. 113, no. 36, pp. 9977–9982, 2016.
- [2] J. Dörpinghaus, "Die soziale Netzwerkanalyse: Neue Perspektiven für die Auslegung biblischer Texte?" *Biblich erneuerte Theologie*, vol. 5, pp. 75–96, 2021.
- [3] F. Grando, L. Z. Granville, and L. C. Lamb, "Machine learning in network centrality measures: Tutorial and outlook," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–32, 2018.
- [4] J. Dörpinghaus, "Digital theology: New perspectives on interdisciplinary research between the humanities and theology," *Interdisciplinary Journal of Research on Religion*, vol. 18, p. 1, 2022.
- [5] J. Dörpinghaus and M. Jacobs, "Semantic knowledge graph embeddings for biomedical research: Data integration using linked open data." in *SEMANTICS (Posters & Demos)*, 2019.
- [6] J. Dörpinghaus, S. Schaaf, and M. Jacobs, "Soft document clustering using a novel graph covering approach," *BioData mining*, vol. 11, no. 1, p. 11, 2018.
- [7] J. Dörpinghaus, J. Klein, J. Darms, S. Madan, and M. Jacobs, "Scaiview—a semantic search engine for biomedical research utilizing a microservice architecture." in *SEMANTICS (Posters & Demos)*, 2018.
- [8] J. Dörpinghaus, D. Samray, and R. Helmrich, "Challenges of automated identification of access to education and training in germany," *Information*, vol. 14, no. 10, p. 524, 2023.
- [9] M. P. Mangroliya, J. Dörpinghaus, and R. Rockenfeller, "Towards the analysis of errors in centrality measures in perpetuated networks," pp. 417–428. [Online]. Available: [https://annals-csis.org/Volume\\_39/drpf4841.html](https://annals-csis.org/Volume_39/drpf4841.html)
- [10] H. T. Tu, T. T. Phan, and K. P. Nguyen, "Modeling information diffusion in social networks with ordinary linear differential equations," *Information Sciences*, vol. 593, pp. 614–636, 2022.
- [11] B. Tadić and R. Melnik, "Self-organised critical dynamics as a key to fundamental features of complexity in physical, biological, and social networks," *Dynamics*, vol. 1, no. 2, pp. 181–197, 2021.
- [12] W.-N. Chen, D.-Z. Tan, Q. Yang, T. Gu, and J. Zhang, "Ant colony optimization for the control of pollutant spreading on social networks," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 4053–4065, 2019.
- [13] F. Menczer, S. Fortunato, and C. A. Davis, *A First Course in Network Science*, 1st ed. Cambridge University Press. [Online]. Available: <https://www.cambridge.org/highereducation/books/first-course-in-network-science/EE22722F27519D8BB1443C7225C57BAF#contents>
- [14] L. C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press ; BookSurge.
- [15] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, 2010.
- [16] L. C. Freeman, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.
- [17] —, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
- [18] D. F. Gleich and R. A. Rossi, "A Dynamical System for PageRank with Time-Dependent Teleportation," *Internet Mathematics*, vol. 10, no. 1–2, pp. 188–217, Apr. 2014.
- [19] C. E. Graafland, J. M. Gutiérrez, J. M. López, D. Pazó, and M. A. Rodríguez, "The probabilistic backbone of data-driven complex networks: An example in climate," *Scientific Reports*, vol. 10, no. 1, p. 11484, Jul. 2020.
- [20] N. Friedman, K. Murphy, and S. Russell, "Learning the Structure of Dynamic Probabilistic Networks."
- [21] G. Ghoshal and A.-L. Barabási, "Ranking stability and super-stable nodes in complex networks," vol. 2, no. 1, p. 394. [Online]. Available: <https://www.nature.com/articles/ncomms1396>
- [22] M. Ruan, X. Chen, and H. Zhou, "Centrality prediction based on K-order Markov chain in Mobile Social Networks," vol. 12, no. 6, pp. 1662–1672. [Online]. Available: <http://link.springer.com/10.1007/s12083-019-00746-y>
- [23] S. Bartolucci, F. Caccioli, F. Caravelli, and P. Vivo, "Distribution of centrality measures on undirected random networks via the cavity method," *Proceedings of the National Academy of Sciences*, vol. 121, no. 40, p. e2403682121, Oct. 2024.
- [24] *Markov-Chain based centralities and Space Syntax' Angular Analysis: an initial overview and application*, Krakow, Poland, 2023.
- [25] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. Princeton, NJ: D. Van Nostrand, 1960.
- [26] S. N. Mohammed and S. Gündüç, "TPM: Transition Probability Matrix – Graph Structural Feature based Embedding," Mar. 2023.
- [27] D.-W. Huang and Z.-G. Yu, "Dynamic-Sensitive centrality of nodes in temporal networks," vol. 7, no. 1, p. 41454. [Online]. Available: <https://www.nature.com/articles/srep41454>
- [28] J. Dörpinghaus, V. Weil, C. Düing, and M. W. Sommer, "Centrality measures in multi-layer knowledge graphs," *Annals of Computer Science and Information Systems*, vol. 32, 2022.
- [29] J. Dörpinghaus, V. Weil, and M. W. Sommer, "Towards modeling and analysis of longitudinal social networks: J. dörpinghaus et al." *Applied Network Science*, vol. 9, no. 1, p. 52, 2024.
- [30] S. P. Borgatti, "Centrality and network flow," *Social Networks*, vol. 27, pp. 55–71, 2005.



# Potential of Found Establishment Data Linkage for the Social Sciences

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**—Social science research on establishments classically relies on representative survey data from official registers. More recently, digital found data sources that can be analysed with AI, especially geolocatable sources such as online job advertisements and OpenStreetMap (OSM), have opened up new opportunities for establishment research. However, their representativeness and possible biases are often unclear. This paper discusses the question of how the representativeness of operational data from different geospatial sources can be compared. It proposes linking these sources and discusses the practical, methodological, and data protection challenges involved, including AI-supported matching at the individual establishment level. The paper also discusses how generative AI can be used to derive key variables, such as industry classifications, from available information, enabling comparisons with register-based benchmark data. The approach is illustrated by showing how linked geolocatable establishment data from two German projects could be used: OSM-based establishment information and online job advertisement data.

**Index Terms**— establishment research, business registers, online job advertisements, OpenStreetMap (OSM), AI-supported data linkage.

## I. INTRODUCTION: JOB ADVERTISEMENT DATA AND OSM AS DATA SOURCES FOR ESTABLISHMENT RESEARCH

**I**NFORMATION at the establishment level, as a meso level of analysis, is highly important for the social sciences, especially in the field of organisational research, labour and education research. The increasing shift of societal processes into the virtual world is making the potential of found data for social-scientific analyses of establishments increasingly relevant. These found data are particularly interesting because, although official register data are comprehensive, they provide only standard information for very basic research questions. Moreover, access to official register data and the possibilities for linking them to other data sources are often restricted. The question of how well data sources cover establishments is of both empirical and methodological importance in order to assess whether, and to what extent, the data allow

for representative conclusions. Two heterogeneous examples of found-data sources that have become more important in the organizational, labour and education research in recent years are data from online job advertisements and geodata, such as data from OpenStreetMap (OSM). Data from job advertisements [1] provide a source offering insights into recruitment processes in training firms. OSM, in turn, is an openly accessible found-data source that are predominantly entered by volunteers and that captures business locations, points of interest, and address information on a broad scale [2]. Since neither of these two sources was designed as a complete business register, their analytical value depends not only on sample size, but also on how selectively, completely, and systematically they represent different types of establishments. This paper therefore discusses, how both sources could be linked at the establishment level to assess their suitability for analysing firms and to identify coverage errors.

The proposed approach focuses on linking heterogeneous digital resources through fuzzy matching and record linkage [3], and on using web mining, text mining, and generative AI to derive key establishment characteristics such as industry, firm size, and training activity from names, addresses, job advertisements, legal notices, and linked websites. These derived variables could then be compared with benchmark distributions from register-based data in order to assess data quality, reliability, and bias.

Legal and ethical questions need to be clarified when dealing with establishment data for example because, in the event of re-identification, such data could create a competitive disadvantage for the establishments concerned. Important steps therefore include anonymisation, data protection, and the permissibility of linking OSM and job advertisement data, especially if the enriched data are later used for establishment surveys.

The approach is situated at the intersection of AI-based and computational approaches in the social sciences and economics. It shows how two heterogeneous digital found data sources in the field of labour market and qualification research can be integrated into a critically reflected analytical framework.

<sup>1</sup>ChatGPT was used as a translation aid to support the drafting of the English text.

## II. STATE OF RESEARCH AND THEORETICAL PERSPECTIVE

### A. Previous Research on the Selectivity and Coverage

To date, there has been little research on the representativeness (Selectivity and Coverage)—and thus the usefulness—of found data on establishments for the social sciences. Thus, it remains unclear, if online job advertisements can be used as a representative sampling frame for subsequent establishment surveys. The approach proposed here is illustrated, first, by means of a project in Germany that collects online job advertisements: Since 2011, the German Federal Institute for Vocational Education and Training (BIBB) has collected such advertisements from a range of online job boards in Germany collected by a private provider. The project is called KASI, short for Kombinierte Ausbildungsstellenanzeigenanalyse und Inserierendenbefragung—Combined Analysis of Apprenticeship Vacancy Advertisements and Survey of Advertisers [1]. As the project shows, job advertisements do not provide a representative picture of all establishments in Germany. Within this data they identified apprenticeship vacancies using a dedicated classification model. Between 2016 and 2024, the number of identified advertisements increased substantially, while the number of new entrants into vocational training remained relatively stable. In addition, there is pronounced regional variation in the relationship between online advertisements and apprenticeship vacancies reported to the Federal Employment Agency. This selectivity is also confirmed at the international level. The Joint Research Centre (JRC), the European Commission's in-house science and knowledge service, notes that online job advertisements are timely and information-rich, but tend to overrepresent higher-skilled profiles and more formalised, socially desirable skill requirements [4]. For this reason, it is recommended to combine online job advertisement data with survey and administrative sources to assess biases.

Research on OSM likewise makes clear that broad coverage should not be mistaken for completeness. In another project of BIBB Gerhards [2] compares OSM data with the official establishment register of the Federal Employment Agency (BA, short for Bundesagentur für Arbeit) and identifies about 825,000 OSM entries, compared with 2,100,000 establishments in the official BA statistics. Under the assumption that both sources refer to identical units, this would correspond to roughly 39 percent coverage. At the same time, the study shows that although the macro-level relationship between OSM and official BA data is correlated, micro-level validity remains uncertain. Earlier research on OSM in Germany also points to spatially and socially uneven completeness. For retail establishments in the German federal states of Baden-Wuerttemberg and Saxony, Brückner et al. report relatively high average completeness levels, but also a considerable range from 42 to 100 per cent [5]. In particular, large cities and regions with lower unemployment were found to be mapped more comprehensively.

### B. Theoretical Perspective on Selective Observability and Unit Incongruence

From a theoretical perspective, the biases of both found data sources can be conceptualised as a problem of selective observability. KASI does not observe establishments as such; rather, it captures a selected subsample: establishments that publicly posted an apprenticeship vacancy online (on established job platforms and on the official website of the Federal Employment Agency) during the observation period, provided an email address, and entered the sample through a selection process. And additionally, as a sampling strategy to avoid duplicate data per establishment, only one training occupation per company—typically the first posted advertisement—is included. OSM, likewise, does not include the full population of establishments (in Germany), but only those business-related units that become visible through mapper activity, data imports, or publicly identifiable points of interest [6]. In both cases, inclusion probabilities are therefore not equal or random, but are shaped by the specific mechanisms through which the data are generated. In the case of KASI, these mechanisms are driven by recruitment practices and digitalisation (strategy to present job postings online in job advertisements); in the case of OSM, they are driven by visibility (potentially establishments that attach importance to being findable via OSM), community activity, and platform-specific logics.

A second theoretical issue concerns the incongruence of observational units. In labour market statistics, the reference concept of an establishment refers to a place of work with employees liable to social insurance contributions; establishments must therefore be distinguished from companies, which may consist of multiple establishments. This distinction is crucial for both KASI and OSM, since in both data sources, individual data points may refer to companies rather than specific establishments. An apprenticeship advertisement may refer to a company as a whole or to several establishments of one company simultaneously, while an OSM entry may represent a branch, company headquarters, showroom, club venue, or another type of point of interest. The central source of bias is therefore not limited to undercoverage, but also includes unit error: entries may formally exist in the data without correctly corresponding to the target population of establishments.

## III. LEGAL AND ETHICAL CONSIDERATIONS IN LINKING AND REUSING ESTABLISHMENT DATA

For data on establishments, specific legal and ethical issues arise, particularly in relation to data linkage and the potential subsequent enrichment of these data with additional survey data. In principle, establishment data are often less sensitive than personal data. However, this does not mean that they can be linked, enriched, and reused without legal or ethical constraints. A decisive issue is whether the data remain at the level of organisations or whether they can be connected to identifiable natural persons. This distinction is particularly important when combining KASI and OpenStreetMap (OSM)

data and when considering whether web-scraped contact details may later be used for establishment surveys [7].

#### *A. Permissibility of Linking KASI and OSM*

By linking data from KASI and OSM, their representativeness can at least be partially assessed. The linkage of KASI and OSM at the individual establishment level is not prohibited as such, but it is only permissible under clearly defined conditions. The first key question is whether personal data are involved at all. The EU's General Data Protection Regulation (GDPR) protects natural persons, not legal entities [8]. Data relating only to firms in the organisational sense—such as company names, legal forms, or generic contact addresses—are therefore often less problematic. The situation changes where sole proprietors (e.g. “Name Inc.”), self-employed persons, named contact persons, personalised email addresses (e.g. name@inc.com), or direct phone numbers are involved. In such cases, the data can relate to identifiable individuals, and matching, enrichment, and subsequent contact may constitute forms of personal data processing.

Where personal data are involved, the crucial issues are the necessity of the processing for a genuine research purpose (this may constitute an exception permitting the processing of personal data), and the proportionality of the procedure. In research contexts, this generally means that only those variables should be linked that are actually required for the analytical objective, and that safeguards such as data minimization (collecting and linking only the data that are necessary), early anonymisation (removing or altering identifying information) where possible, or at least pseudonymization (replacing direct identifiers with codes) should be built into the design. The legal assessment therefore depends not simply on the existence of a scientific interest, but on whether the linkage is genuinely necessary and whether the interests of affected persons are adequately protected [9] [10].

The fact that (found) data are publicly accessible does not remove these obligations. Publicly visible contact details or other business-related information may not be a legally unrestricted resource. If information is scraped or otherwise collected from sources without being obtained directly from the data subject, researchers must generally consider transparency obligations, including informing affected persons about the source, purpose, and legal basis of the processing. Public availability therefore does not automatically legitimise large-scale collection, linkage, or reuse.

In addition to data protection law, other legal constraints may also apply. OSM data are comparatively straightforward in this regard, as they are explicitly made available for reuse under the Open Database License, which allows reuse subject to conditions such as attribution (the data source must be properly credited) and, where applicable, share-alike requirements (derived databases may have to be shared under the same or a compatible open license). It might remain unclear, at least initially, whether data on establishments contained in the two data sources are present without the establishments' consent and may subsequently be processed for

purposes entirely different from those for which they were originally made available. KASI raises additional issues because the underlying job advertisement corpus originates from a commercial provider. Even where processing within a research project is lawful, this does not automatically imply that raw data or identifiable components may be freely shared or republished. If further contact details are additionally scraped from company websites, researchers must also take into account website terms of service.

#### *B. Use of Scraped Establishment Addresses for Establishment Surveys*

For social science analyses, found data offer new potential alongside designed data from purpose-built surveys. One way to extend the analytical potential of found data is to enrich them with designed data. A possible approach is to use the contact information of establishments (within the framework of the legal and ethical limitations outlined above) — in this case obtained from OSM or KASI — to conduct dedicated establishment surveys. In the case of the KASI project, this is already an integral part and objective of the project. If linked or scraped data are to be used for subsequent establishment surveys, it is important to distinguish clearly between postal contact, email contact, and telephone contact.

Initial contact by post may be comparatively more feasible, but should still be subject to institutional data protection and legal review, especially when it is addressed to a legal entity or a generic business address and the communication is clearly framed as scientific research. In such cases, the research purpose may be more straightforward to assess in the context of institutional data protection and legal review, while the interference with personal privacy may be comparatively limited.

Email contact is more sensitive [11]. Unsolicited electronic communication is subject to stricter legal scrutiny (in the EU and Germany), particularly where no prior relationship or consent exists. Unsolicited contact by email remains legally delicate when it targets named individuals or personalised addresses. Telephone contact is the most sensitive channel [12]. Telephone recruitment based on scraped contact information is generally the most problematic option for establishment surveys.

#### *C. Ethical Guidelines*

Even where the legal framework does not prohibit data processing, established research ethics guidelines call for a broader assessment. The ethics guidelines of the Association of Internet Researchers (AoIR), a learned society dedicated to the advancement of the transdisciplinary field of Internet studies, stress that public accessibility should not be equated with ethical acceptability. What matters instead are the context in which the data were made available, the expectations of the persons concerned, the degree of identifiability, the possibility of harm, and the question of whether data use may alter the social situation of those concerned. For establishments in particular, the possibility of re-identification and linkage with data that are otherwise unavailable or difficult to

TABLE I.  
PROPOSED APPROACH: STEPS FOR LINKING AND EVALUATING BIASES IN FOUND DATA AT THE ESTABLISHMENT LEVEL

|   |                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|---|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | legal and ethical considerations in linking and reusing establishment data perspective: permissibility, use of scraped contact information, ethical guidelines                                                                                                                                                                                                                                                                                |
| 2 | Data cleaning: standardization of words, removal of duplicates, fuzzy matching (blocking, similarity measures, manual review of borderline cases)                                                                                                                                                                                                                                                                                             |
| 3 | entity resolution (matching of datasets): blocking or candidate generation, pairwise similarity features, human-in-the-loop pipeline (manually coded pairs, active learning, global assignment rule)                                                                                                                                                                                                                                          |
| 4 | generative AI for Classification (e.g. on Industry Codes, Establishment Size Groups, and Training Activity) as part of a hybrid pipeline: web research on company websites; substantive enrichment of linked establishments (compilation into a standardized dossier; LLM with restricted classification task);<br>simple cases: Deterministic rules and classical Natural Language Processing (NLP);<br>more complex or ambiguous cases: LLM |
| 5 | Assessing Coverage Bias: descriptive coverage matrix, multivariate models                                                                                                                                                                                                                                                                                                                                                                     |

obtain may create competitive risks. Correspondingly, the GESIS as a major German research institute guidance on digital behavioral data highlights privacy, data minimisation, the risk of inferring sensitive attributes, and the need to consider the entire data lifecycle from collection to publication [13].

#### IV. PROPOSED METHODOLOGICAL APPROACH

Table 1 illustrates the pipeline of the proposed approach for linking and evaluating biases in found data at the establishment level.

##### *A. Error Structure, Data Cleaning, and AI for Fuzzy Matching Between KASI and OSM Establishments*

A fair evaluation of the representativeness of KASI and OSM requires a distinction between two establishment-level target populations. The first is the full population of establishments with employees, which serves as the general establishment population. The second is the subset of establishments that advertised apprenticeship vacancies online during the observation period. On both levels the official establishment register of the Federal Employment Agency (BA) can serve as a reference for assessing representativeness. However, due to restrictions, matching at the individual establishment level is not possible, and only marginal distributions can be compared (macro-level comparison). Only at this second level can KASI be interpreted as a coverage instrument (to assess how well it covers the target population) as it captures the subpopulation of apprenticeship-providing establishments that recruit digitally.

Both data sources may contain incorrect inclusions (false positive) as well as missing establishments (false negative).

In KASI, such problems arise mainly from advertisements listed at the company rather than establishment level, multi-location postings, central recruiting mailboxes, or advertisements that cannot be assigned to a specific establishment. Missing cases include establishments that did not advertise online (first level) or have no/could not be linked to a contactable email address.

In OSM, problematic entries may include outdated nodes (individual geographic points), non-economic institutions,

brand-only entries, headquarters without a corresponding employee-bearing establishment, or units without employees subject to social insurance contributions. Gerhards [2] notes that the macro-level comparison with Federal Employment Agency data does not resolve whether OSM entries and agency establishments actually refer to the same units, or whether some recorded entries are not employee-bearing establishments at all. In OSM, missing cases include establishments that are unmapped, less visible, or primarily oriented toward business-to-business markets.

Data cleaning must therefore be treated as a separate methodological step. Names, legal-form suffixes, punctuation, umlauts, abbreviations, and address elements first need to be standardized. Duplicates must then be removed within each source, since both repeated job postings and repeated points of interest may occur. Only in a third step should multi-stage fuzzy matching be applied: first, exclusion (blocking) of cases from matching by postal code, municipality, federal state, or spatial distance (in many cases geocoded GPS coordinates are available for OSM and KASI Data); second, similarity measures for names and domains; and third, probabilistic linkage with manual review of borderline cases. External websites and legal notices are especially useful for distinguishing between corporate groups, brands, branches, and local establishments. A validated hand-coded sample is advisable for calibrating thresholds and making precision and recall transparent.

The most obvious and methodologically central application of classical AI (in this step not generative AI) in the proposed approach for linking KASI and OSM is entity resolution [14], that is, the probabilistic linkage of identical establishments across two data sources. OSM contains spatially referenced entries with names, address components, and sometimes websites or phone numbers. KASI contains establishments that advertised apprenticeship vacancies online between July 2024 and July 2025, including contact details and further information extracted from the advertisements. A simple exact match based on names or addresses would fail systematically, because one must expect name variants, branch labels, legal-form suffixes, misspellings, company-level rather than es-

establishment-level designations, and incomplete location information. For this reason, AI-assisted matching at the individual establishment level represents a key methodological advance.

A multi-stage approach seems well suited for this task and is proposed here. In the first step, strict blocking (exclusion from comparison) or candidate generation (probable matches) should be used so that not every KASI establishment is compared with every OSM entry. Appropriate blocks can be defined by postal code, municipality, district, or a geographic radius around geocoded addresses. In the second step, pairwise similarity features are constructed, including Jaro–Winkler or Levenshtein similarities for names, token-based similarities for street and locality, exact or partial matches for house number, website domain, telephone number, and email domain, as well as spatial distance measures. Domain-specific features are also useful, such as the detection and standardization of common German or international legal-form suffixes like GmbH, AG, e.K., eG, or gGmbH, and branch markers such as “Center,” “Market,” “Store,” “Branch,” or place-name extensions. These features can then be combined in a supervised matching model, for example through gradient boosting, random forests, or simple logistic regression.

In practice, the procedure should be designed as a human-in-the-loop pipeline (with active human oversight). First, a gold standard of manually coded pairs should be created, distinguishing confirmed matches, confirmed non-matches, and ambiguous cases. This serves as the basis for training and iteratively improving the classifier. Active learning (a machine learning approach where the model selectively asks for labels on the most informative data points to improve performance efficiently) is especially useful, since the model can present uncertain or contradictory cases for manual review [15]. After pairwise classification, a global assignment rule is required to prevent multiple OSM entries from being incorrectly assigned to a single KASI establishment, or vice versa. The procedure should be evaluated using classification evaluation metrics (precision, recall, F1 score), and calibrated match probabilities, supplemented by error analyses across industry, establishment size class, and federal state.

#### *B. Generative AI for Constructing Variables on Industry, Establishment Size, and Training Activity.*

Potential biases in the data should be examined using a multivariate approach. The aim is to compare the marginal distributions in official Federal Employment Agency (BA) data with those observed in the found data of OSM and KASI. To enable this comparison using marginal distributions along classifications of the BA, the information from KASI and OSM must be classified into classes like industry-codes or firm size levels. BA data contains information on industry, location (e. g. German federal states), and employment counts.

Such an analysis requires that key establishment characteristics be available in a consistent form or be reconstructable. For industry, the official WZ classification (short

for Wirtschaftszweig, branch, a German system for categorizing economic activities into standardized industry sectors) provides the reference framework.

For KASI and OSM, where direct classification from a reference source is unavailable, industry coding should proceed in stages: through web research on company websites and legal notices; and through text-based classification using job advertisement titles, OSM tags, and activity descriptions.

Federal state can be identified relatively robustly from geocoded addresses or OSM coordinates.

The second field of application concerns the substantive enrichment of linked establishments. This is where generative AI may be especially useful in the approach proposed here. Many analytically central characteristics are not available in KASI or OSM as clean, standardized variables but as texts. Industry, establishment size, and training activity therefore need to be reconstructed or approximated from heterogeneous traces: establishment names, legal form and address details, apprenticeship advertisement text, legal notices, career pages, “about us” pages, and other linked content on company websites. KASI is particularly informative in this respect, using additional information contained in advertisements and a dedicated classifier for apprenticeship postings [1]. Conversely, the OSM study highlights the potential of linking OSM with company websites, job advertisements, and other digital resources [2].

In this context, generative AI should not be used as an open-ended text tool (that generates unrestricted, free-form text without predefined constraints), but as a structured extraction and classification component. A two-stage procedure is methodologically advisable. First, all available text sources and metadata for an establishment should be compiled into a standardized “dossier” (set of variables): name, legal form, OSM tags, advertisement title, advertisement text, legal notice, self-description on the website, contact page, location list, and, where available, career pages. In the second step, a large language model (LLM) [16] should be given a restrictive extraction task with a fixed output schema, for example in JSON (a data format used to store and exchange structured data, such as model inputs, outputs and labels), including the presumed main industry, a WZ-related industry description, indicators of branch structure, indicators of employee count or size class, evidence of training activity, and finally a confidence score (numerical value indicating how certain the model is about its prediction or extracted information), and an “evidence span” field (a reference to the specific text segment(s) in the source data that support or justify the extracted information). This forces the model to derive variables from explicit textual evidence rather than merely guessing.

For example, for industry, terms such as “bakery,” “dental practice,” “logistics provider,” “metalworking,” “car dealership,” or “tax consultancy” are strong signals. For establishment size, relevant indicators include references to multiple branches, franchise structures, team size, or expressions such as “employees,” “locations,” “family-run craft business,” or “international corporation.” For training activity, terms such

as “apprenticeship,” “trainee,” “dual training,” “training occupations,” chambers such as “IHK,” (short for Industrie- und Handelskammer — Chambers of Industry and Commerce) “HWK,” (short for Handwerkskammer — Chambers of Crafts) or “we provide training” are directly informative.

Generative AI should not operate in isolation, but as part of a hybrid pipeline: Deterministic rules and classical Natural Language Processing (NLP) components should first handle the straightforward cases, such as legal-form recognition, domain extraction, pattern matching for training activity, counting location lists, and detecting explicit employee ranges or size information. Only more complex or ambiguous cases should be passed to the LLM, for example when industry must be inferred from several text fragments or when size class must be derived from weaker “signals” (correlations). The procedure may become especially robust when the model does not generate freely, but instead works with retrieval-augmented prompting based only on previously extracted text snippets and is required to return textual evidence for its decision rather than on its internal knowledge alone.

The results should then be filtered using confidence thresholds and validated against external reference sources, such as register-based industry information or known indicators of training activity.

### *C. Assessing Coverage Bias against a Third “Ground Truth” Source*

Once the data have been brought into a comparable form, the analysis for assessing coverage bias should begin with a descriptive coverage matrix. For each source, the first step is to report what share of establishments in the reference data from official Federal Employment Agency (BA) statistics appears overrepresented in KASI, in OSM, in both sources, or in neither. This should be followed by stratified coverage rates by industry, size class or federal state. Raw proportions and comparison of overcoverage and undercoverage makes it possible to show whether a source includes many establishments overall but systematically captures the wrong types of units.

The next step should be a multivariate model, such as a logistic (e.g. training/non training establishments) or multinomial regression (e.g. economic sector). For KASI, the model should explicitly refer to the subpopulation of potentially training-active establishments. For OSM, the general establishment population can also be analyzed.

Three expectations follow from the design. First, KASI is likely to cover the general establishment population much more selectively than OSM, because by design it is restricted to apprenticeship-providing establishments that recruit online and provide an email address. Second, OSM is expected to reach a broader range of establishments, but also to contain more incorrect inclusions, since visible POIs do not necessarily correspond to employee-bearing establishments. Third, the overlap between the two sources is likely to be greatest among larger, publicly visible, service-oriented, or customer-facing establishments, whereas small, industrial, business-to-

business, or informally recruiting establishments are likely to remain underrepresented in at least one source. For KASI, findings on regionally uneven online recruitment patterns point to systematic coverage differences. For OSM, earlier research suggests that visibility, urbanity, and local socioeconomic context may matter [2].

## V. CONCLUSION: LINKING KASI AND OSM WITH AI: METHODOLOGICAL INSIGHTS, DATA QUALITY, AND FUTURE DIRECTIONS

This article suggested an approach how in labor market research found establishment data sources can be linked and compared. It focused how generative AI methods could be used to construct variables from text. The proposed methodological advantage is that generative AI not only can accelerate classification, but also can bridge the gap between unstructured web and advertisement texts and an analyzable multivariate establishment dataset.

Linking found establishment data from online job advertisements (project KASI) and georeferenced data from open street map (OSM) at the individual establishment level promises a double analytical gain: Substantively, it makes it possible to identify which types of establishments are visible in both sources and which are systematically missing. Methodologically, it shows that coverage is not merely a question of quantity, but a question of the unit of observation and the process through which data are generated.

Especially when combined with additional data (e. g. an establishment survey) KASI is likely to be especially valuable for analyzing apprenticeship-providing establishments that recruit online, but not as a general establishment register. OSM, by contrast, is attractive as an open infrastructure for establishment-related found-data research, yet it requires much stricter cleaning and validation before it can serve as a reliable sampling frame.

In this way, the proposed approach would extend the empirical literature in two respects: first, by moving from aggregated coverage comparisons against official statistics to establishment-level error measurement; and second, by systematically combining job advertisement data with open accessible OSM based found data so that the contents of both data sources can be linked.

For the research question on how geodata on establishments can be linked, the linkage of KASI and OSM should be understood not only as a technical task, but also as a legally and ethically conditioned research practice. Establishment data may often be less sensitive than personal data, but they are not free of regulatory and ethical obligations some of which are particularly specific to establishment data. A responsible design therefore requires a careful distinction between organisational and personal information, clear limitation to what is necessary for the research objective, appropriate safeguards for processing and storage, and a cautious approach to any later use of scraped contact details for establishment surveys. For the proposed approach, this leads to a practical conclusion and data protection implications: linking

KASI and OSM is feasible in principle, provided that a clear distinction is maintained between non-personal establishment data and personal contact data or sole proprietorship data, that a sound legal basis is documented, and that research purpose, data minimization, transparency, and the right to object are properly implemented. For subsequent establishment surveys, contacting firms by post or via generic functional addresses is legally and ethically far more defensible than unsolicited phone calls or personalized cold emails. Where additional web scraping of company websites is envisaged, terms of service, technical protection measures, information requirements, a data protection impact assessment for larger-scale processing, and approval by data protection officers or ethics committees should be integrated into the study design.

The main limitation of the design is that two imperfect data sources cannot be fully validated against one another without an external reference. The establishment register of the Federal Employment Agency is the most convincing third source and reference. Another limitation is the focus on Germany-specific data from proprietary or project-specific datasets (OSM and KASI) which may limit the generalisability of the findings to other countries or contexts. Nevertheless, in many countries, OSM data on establishments and online job advertisement data are publicly available or available by web-scraping, making the approach potentially transferable beyond the German context.

A next step in the practical implementation of the proposed approach lies in semi-automated web enrichment, active-learning-based matching procedures, and validation against Federal Employment Agency (BA) reference data. It should be noted: Even the reference source of the BA official register, however, is not perfect. It does not cover solo self-employed persons without relevant employment notifications, and its definition of an establishment may differ from the meaning of an OSM point of interest (POI)/node or a KASI advertisement.

#### REFERENCES

- [1] Bundesinstitut für Berufsbildung (BIBB), “Kombinierte Ausbildungsstellenanzeigenanalyse und Inserierendenbefragung (KASI),” (engl. Federal Institute for Vocational Education and Training (BIBB), “Combined Analysis of Apprenticeship Vacancy Advertisements and Survey of Advertisers (KASI),”) BIBB project documentation, 2025. <https://www.bibb.de/de/210741.php>
- [2] C. Gerhards, “Coverage of OpenStreetMap data at firm level. Assessing the suitability for social science representative firm samples and found data firm analyses,” in *Proc. 20th Conf. Computer Science and Intelligence Systems*, vol. 43, 2025, pp. 653–658.
- [3] Christen, Peter. “The data matching process.” *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. 23-35.
- [4] M. Sostero, E. Fernández-Macias, F. D. Baciocco, and E. Cantalini, “The Professional Lens: What Online Job Advertisements Can Say about Occupational Task Profiles,” *JRC Working Papers Series on Labour, Education and Technology*, no. 2021/02, European Commission, Joint Research Centre, Seville, 2021.
- [5] J. Brückner, M. Schott, H. Fan, and A. Zipf, “Assessing shop completeness in OpenStreetMap for two federal states in Germany,” *AG-ILE: GIScience Series*, vol. 2, 2021, Art. no. 20.
- [6] H. Senaratne, A. Mobasheri, L. Ali, C. Capineri, and M. Haklay, “A review of volunteered geographic information quality assessment methods,” *International Journal of Geographical Information Science*, vol. 31, no. 1, pp. 139–167, 2017.
- [7] M. A. Brown, A. Gruen, G. Maldoff, S. Messing, Z. Sanderson, and M. Zimmer, “Web scraping for research: Legal, ethical, institutional, and scientific considerations,” *Big Data & Society*, vol. 12, no. 1, 2025, doi: 10.1177/20539517251381686
- [8] European Union, “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 (General Data Protection Regulation),” *Off. J. Eur. Union*, L 119, May 4, 2016.
- [9] Federal Republic of Germany, “Federal Data Protection Act (BDSG), Section 27: Data processing for purposes of scientific or historical research and for statistical purposes,” official English version, 2025.
- [10] European Data Protection Supervisor, “Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research,” *EDPS*, Brussels, Jan. 2022.
- [11] BVM Berufsverband Deutscher Markt- und Sozialforscher e.V., ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., and DGOF Deutsche Gesellschaft für Online-Forschung e.V. (engl. BVM Professional Association of German Market and Social Researchers, ADM Working Group of German Market and Social Research Institutes, and DGOF German Society for Online Research.), “Guideline on Online Surveys,” Berlin, 2021.
- [12] Rat der Deutschen Markt- und Sozialforschung (engl. Council of German Market and Social Research), “Guideline on Telephone Surveys,” Berlin, 2021.
- [13] J. Breuer, S. Stier, J. Lukito, F. Mangold, M. Wieland, D. Radovanović, K. Weller, and C. Wagner, *Overview of Ethical Considerations when Working with Digital Behavioral Data*. Cologne, Germany: GESIS – Leibniz Institute for the Social Sciences, 2025.
- [14] Papadakis, George, et al. “Blocking and filtering techniques for entity resolution: A survey.” *ACM Computing Surveys (CSUR)* 53.2 (2020): 1-42.
- [15] A. Primpeli and C. Bizer, “ALMSER: A Graph-Boosted Active Learning Method for Multi-Source Entity Resolution,” in *Proc. 20th Int. Semantic Web Conf. (ISWC 2021)*, 2021.
- [16] J. Dagdelen, A. Dunn, S. Lee, N. Walker, S. Rosen, G. Ceder, and A. Jain, “Structured information extraction from scientific text with large language models,” *Nature Communications*, vol. 15, 2024, Art. no. 1418, doi: 10.1038/s41467-024-45563-x



# A Holistic CPS for Smart Precision Livestock Farming: Multi-Actor Collaborative Decision-Making Across Ecosystem Domains

Sonia Hajri-Gabouj

Computer Science and Telecommunications lab.,

ECAM Louis de Broglie, Bruz, France

University of Carthage, INSAT, Tunisia,

Computer Science for Industrial Systems lab., LR11ES26

Email: sonia.hajri-gabouj@ecam-ldb.fr

Imen Harbaoui

0009-0007-4009-6448

University of Carthage, INSAT, Tunisia

Analysis, Design and Control of Systems lab., LR11ES20,

University of Tunis El Manar, ENIT

Email: imen.harbaoui@insat.ucar.tn

Rabaa Youssef-Douss

0000-0002-0254-9270

University of Carthage, INSAT, Tunisia

Communication, Signal and Image (COSIM) lab., LR11TIC01, Sup'COM

Email: rabaa.youssef@insat.ucar.tn

**Abstract**—Precision Livestock Farming (PLF) increasingly relies on integrated technologies to improve animal health, welfare, reproduction, and productivity within a complex ecosystem. However, current approaches often deal with specific use cases, single domain, and remain fragmented, relying on siloed analytics tools and unilateral decision-making. This paper proposes a holistic Cyber-Physical System (CPS) for smart PLF that enables seamless end-to-end processes, spanning several stakeholders' perspectives and farm domains: animals, environment, production, and logistics. The proposed layered and modular architecture leverages adaptive AI capabilities enabling intelligence-driven functionalities within an integrated farm workflow, and supports automated, human-in-the-loop and hybrid modes. Both domain-specific and shared views are provided to foster collaborative decision-making, and control across operations. The CPS capabilities are illustrated through various use cases, highlighting their relevance across different operational settings. By integrating sensing, data, analytics, human perspectives, and collaborative decision-making, the proposed CPS enhances visibility, transparency, and efficiency in livestock management and monitoring.

## I. INTRODUCTION

**P**RECISION Livestock Farming (PLF) increasingly relies on integrated technologies to support farm management and monitoring in order to enhance animal health, welfare, reproduction, and production [1], [2]. In fact, modern PLF systems collect and analyze data from sensors and environmental monitors to provide actionable insights and support real-time decision-making [3], [4]. Despite these developments, PLF systems still face significant challenges among them, the limited collaborative decision support among multiple stakeholders since data access and alert sharing are typically controlled by the farmer or herd manager [5]. Only a few initiatives, such as ClearFarm [6], integrate external stakeholders at the value-chain level, but not at the operational level where

the decision-making remains unilateral and primarily farmer-centered. Although differences in actor perspectives are recognized, existing approaches partially address role-specific views and lack structured processes enabling stakeholders the opportunity to interact, express their viewpoints, validate outcomes, and co-construct decisions [7], [8], [9], [6]. In addition, the PLF ecosystem is addressed partially, and most systems focus on isolated aspects of a specific domain rather than covering multiple domains including animals, environment, infrastructure, production, logistics and supply chain [11]. Furthermore, AI capabilities remain task-specific, addressing particular and fragmented prediction or optimization problems instead of providing scalable and adaptive support for intelligence-driven functionalities [12]. Finally, many initiatives focus narrowly on a single decision-making or control mode while neglecting multilevel autonomy embedding automated, human-in-the-loop and hybrid modes. The coordination between decision-making and command execution processes for the implementation of approved recommendations into concrete tasks also remains largely absent [3], [5]. The fragmented structures of existing PLF systems hinder end-to-end visibility and transparency, restrict the autonomy of specialized actors, and may delay timely responses to health, welfare, reproduction or production issues. To address these challenges, this paper presents a holistic Cyber-Physical System (CPS) for smart PLF ecosystems. Our main contributions include:

- The integration of multiple PLF domains including animals, environment, infrastructure, production, processes and logistics, ensuring comprehensive integration.
- The integration of several stakeholder with various perspectives through domain-specific views, enabling coordinated decision-making across roles.

- A modular and adaptive AI capabilities, tailored to processes' needs in perception, diagnosis, prediction, decision-making and adaptive control, to support system-wide observation, reasoning, action and collaboration functionalities within an integrated farm-scale workflow.
- A modular, multi-level architecture integrating flexible end-to-end workflows supporting automated, human-in-the-loop, and hybrid modes of control and monitoring, while fostering visibility, transparency, and collaboration among stakeholders.

This CPS provides an integrated solution addressing identified challenges in current PLF implementations.

The remainder of this paper is organized as follows. Section II reviews related work on PLF systems and highlights existing challenges. Section III presents the functional and non-functional requirements for smart PLF environments. Section IV describes the proposed holistic CPS architecture, including its layers, modules, and workflows. Section V details the proposed Intelligent and collaborative decision support modules. Section VI presents the compliance with the requirements and illustrative use-cases. Finally, Section VII concludes the paper, summarizes key findings, and outlines insights and directions for future research.

## II. RELATED WORKS

### A. Multi-Actor Collaboration Limitations

In PLF, multiple actors must collaborate to ensure effective farm management and monitoring efficiency [5]. Some existing approaches enable the sharing of sensor-derived data and welfare indicators, but the collaboration remains informal and lacks structured mechanisms and decision support architectures, limiting coordinated decision-making [9]. ClearFarm Project [6] emphasizes multi-actor co-design at the value-chain level, but operational roles such as veterinarians, nutritionists, and maintenance staff are not formally integrated, and some interactions still rely on the farmer as intermediary and unilateral decision-maker. Consequently, stakeholders' key expertise is not taken into account and their proposals are not coordinated, leading to delayed, inconsistent, or conflicting on-farm decisions [7].

### B. Multi-View and Multi-Perspective Integration

Current PLF systems provide dashboards for monitoring farm indicators, allowing actors to interpret situations according to their roles. However, these views are generally isolated, limiting sharing of variety of perspectives and experiences [9]. The ClearFarm Project [6] illustrates harmonized data sharing at the value-chain level and emphasizes combining perspectives for decision-making. Nevertheless, the operational multi-view integration on the farm has not been addressed yet. The involved actors have distinct perceptions and priorities regarding the same situation, depending on a variety of factors such as risk and trust in technology [7], [13]. Different users assume distinct roles in collaborative decision-support systems, which require both role specific and shared interfaces to ensure expertise, visibility, transparency and coordination [14]. Thus,

capturing and managing multi-views and multi-perspectives remains a challenge for informed decision-making.

### C. Partial Integration Across Farm Domains

Most existing decision-support systems in PLF focus on isolated domains, such as animal health, welfare, or production, without integrating data across animals, environment, infrastructure, production, logistics, and supply chain processes [11]. This siloed approach can lead to conflicting and/or non-actionable decisions, thus limiting holistic coordination, management and monitoring efficiently livestock operations. For example, a PLF alert detects a health problem in a cow. The veterinarian validates an immediate change in treatment. When logistics providers receive the order, they report that they cannot fulfill the request. Effective coordination between actors in the animal and the logistics domains could have reduced delays and enabled a single, consistent, and immediately applicable decision. Although IoT-based sensing systems enable domain-level data collection in livestock environments, these solutions remain fragmented and are deployed separately within each domain [26], [27], [28].

### D. AI Capabilities in Precision Livestock Farming

AI techniques have been widely applied for behavior recognition [15], [16], health monitoring [9], [10], lameness detection [23], estrus detection [24], disease and stress discrimination [25] and welfare assessment [17]. In [12], the authors present a comprehensive overview of AI approaches. In this context, AI applications are typically developed for each emerging need as isolated, task-specific solutions that are not designed as modular and adaptive components integrated into a farm-wide workflow. Furthermore, adaptive control and learning, whereby outcomes are fed back to refine models, remain largely unexplored, particularly in terms of mechanisms and continuous workflow.

### E. Multi-Level Autonomy and Command Execution

Some PLF systems enable automated actions, such as feeding or environmental control adjustments, but these are generally confined to isolated subsystems [3], [5]. Existing work rarely considers multi-level autonomy across automated, human-in-the-loop and hybrid decision-making and control modes, and often overlooks execution, leading to a gap between recommendations and their implementation.

### F. Cyber-Physical System Integration

Research on cyber-physical systems in PLF has mainly addressed the integration of sensing, computation, and control to support decision-making [18]. However, related CPS are often limited to isolated applications or subsystems related to a single domain, and comprehensive workflows spanning all farm operations are still largely absent [9], [19]. Prototype systems combining AI and modular sensors have been developed for applications such as rumen monitoring in dairy cows. Unfortunately, multi-actor collaboration and human-in-the-loop integration remain rare [20]. Overall, these studies

illustrate that current CPS are fragmented and that holistic, end-to-end workflows across PLF domains have not yet been implemented [9], [19].

### III. CPS REQUIREMENTS FOR SMART PLF ENVIRONMENTS

This section presents the functional (**RQ1** to **RQ14**) and non-functional (**RQ15** to **RQ17**) requirements of the CPS architecture for smart PLF environments. These requirements are derived from Section II, with traceability to the corresponding subsections (§) and references in the related work.

**RQ1** The system shall integrate several domains within a smart PLF ecosystem (§ II-C [11], § II-F [19]).

**RQ2** The system shall support cross-domain stakeholders with diverse roles and expertise through human-in-the-loop and bidirectional interactions for action execution and system feedback (§ II-A [5], [7], [8], § II-B [14]).

**RQ3** Users will be able to interact with a dedicated dashboard to support continuous performance monitoring through KPI-based visualizations and feedback loops (§ II-B [9], [14]).

**RQ4** Users will be able to interact with a dedicated dashboard to manage, monitor and track system alerts in real time (§ II-B [9], [10]).

**RQ5** The system shall provide a dedicated dashboard enabling stakeholders to access specific and shared interfaces to express viewpoints, share recommendations, and collaboratively validate decisions under operational constraints (§ II-A [7], § II-B [6], [14]).

**RQ6** The system shall provide a configuration dashboard enabling actors to manage system-wide parameters and interact bidirectionally with system feedback (§ II-A [5], § II-B [9], § II-D [12], § II-F [18]).

**RQ7** The system shall generate alerts related to performance, real-time system status, AI, execution, and human/decision-related dimensions (§ II-D [10], [23], [24], [25]).

**RQ8** The system shall provide a data acquisition capability across all ecosystem domains with real-time transmission of collected data and/or information (§ II-C [26], [27], [28]).

**RQ9** The system shall support transforming raw data, storing results for historical tracking and traceability, accessing AI models, and distributing processed data to relevant system modules (§ II-D [12], § II-F [19]).

**RQ10** The system shall provide AI-driven perception, diagnosis, prediction, decision optimization, and adaptive control and learning using data-driven methods, supporting performance monitoring, alert generation, decision-making, and AI agent functions as needed (§ II-D [15], [16], [17]).

**RQ11** The system shall provide decision support using contextual analysis and multi-source data fusion, generating consistent decision proposals and shared views, triggered by alerts or user requests (§ II-A [5], § II-B [13]).

**RQ12** The system shall support decision-making and control processes in automated, AI-assisted, human-operated, and hybrid modes, triggered by system alerts or user requests (§ II-E [3], § II-E [20]).

**RQ13** The system shall support multi-level autonomy in

command execution, enabling automated, manual, and hybrid execution of system commands depending on operational constraints (§ II-E [3], § II-F [19]).

**RQ14** The system shall incorporate execution feedback to ensure consistency, alignment, and continuous improvement across system operations, with feedback propagated to the originating components (§ II-C [2], [3], [9], § II-D [12]).

**RQ15** The system shall ensure end-to-end traceability and consistency across data, decision, control, execution, and human interaction flows (§ II-C [11], § II-F [19]).

**RQ16** The system shall adopt a modular architecture to support flexibility and scalability, with clearly defined interactions between components (§ II-C [4], [28], § II-F [19]).

**RQ17** The system architecture shall support applicability across different PLF systems without being constrained to a specific operational setting (§ II-C [11], § II-F [19]).

### IV. PROPOSED CPS ARCHITECTURE

The proposed CPS architecture consists of a human layer, a physical layer, a communication layer, and three cyber layers.

#### A. Human Layer

It represents stakeholders interacting with the CPS through dashboards and the PLF ecosystem, contributing to collaborative decision-making. Key actors include veterinarians (animal health and welfare), farm managers (farm operations and resource coordination), nutritionists (animal nutrition), reproduction specialists (animal reproduction), technicians (CPS infrastructure and maintenance), data analysts (AI and data analytics), agricultural consultants (technical expertise and innovation support), regulatory inspectors (compliance and certification), and supply chain representatives (quality and logistics). Each actor accesses role-specific information to contribute to, validate, and refine system decisions.

#### B. Physical Layer

It encompasses animals, sensors, actuators, physical environments, infrastructure, production, and logistics within the PLF ecosystem. It includes four main domains[9]:

- *Animals*: focuses on health, welfare, growth, and reproduction using wearable sensors (heart rate, accelerometers, rumination), biosensors (temperature, pH), computer vision [24], [25], and RFID for identification and tracking. Actuators support automated feeding and precision treatment delivery, enabling closed-loop decision-making.
- *Environment & Infrastructure*: covers housing and operational conditions monitored via temperature and humidity sensors, with actuators for ventilation, heating, and lighting control to ensure welfare and efficiency [26].
- *Production & Processes*: includes sensors for milking (yield, conductivity), feed intake, motion, and weighing [27], with actuators such as robotic milking, precision feeding, automated sorting, and herd routing for workflow automation.
- *Logistics & Supply Chain*: relies on GPS, RFID/barcodes, and cold-chain sensors, with actuators for automated

refrigeration and smart routing to ensure traceability and optimized logistics.

### C. Communication Layer

Positioned between the Physical and Data Treatment & Storage Layers, it connects sensing/actuation components to data processing modules via networks, IoT protocols, and transmission infrastructures [28]. It transmits: (i) Sensor data from the Physical Layer to the Data Treatment & Storage Layer. (ii) Control commands/feedback in the opposite direction to actuators. This bidirectional exchange enables continuous interaction between the physical and cyber systems. It ensures reliable, secure transmission using standardized protocols (MQTT, CoAP, HTTP), maintaining synchronization and data integrity for real-time decision-making and system coordination.

### D. Cyber Layer

It consists of three layers. The Data Treatment & Storage Layer handles data collection, preprocessing, and storage of structured and historical data, including AI models. The Intelligent Decision and Control Layer provides AI-driven analytics, generating alerts, decisions, and control commands for both automated and human-in-the-loop operation. The Application Layer offers dashboards for monitoring, visualization, KPI tracking, alerts, and collaborative decision-making. Table I summarizes the modules and their functions across these layers.

### E. CPS layers, modules and flows

Figure 1a illustrates the proposed layered CPS architecture, its modules, and the exchanged flows between components, while Figure 1b details these interactions. The architecture integrates perception, data processing, decision-making, and execution within a generic framework supporting human-in-the-loop collaboration and AI deployment across multiple application domains. Its generic, extensible design ensures coherent cross-layer information flows while remaining independent of any specific application context.

## V. INTELLIGENT AND COLLABORATIVE DECISION SUPPORT MODULES AND WORKFLOWS

### A. AI Core Module: Capabilities and Macro-Functions

We propose an AI Core Module that provides modular, adaptive, and context-aware AI capabilities for anomaly detection, alert generation, information aggregation and analysis, decision support, collaboration, operational control, decision-to-command translation, and system performance monitoring. These capabilities enable system-wide observation, reasoning, and action across multiple use cases, including key AI macro-functions such as:

- *Perception*: infers animal behavior, physiological indicators, and environmental conditions from multimodal sensor and system data.
- *Diagnosis*: identifies anomalies, deviations, and root causes.

- *Prediction*: anticipates risks, events, and performance degradation.
- *Decision & Optimization*: generates recommendations and constructs actor views, while fusing information, resolving conflicts, and prioritizing optimal actions.
- *Adaptive Control & Learning*: continuously updates models and control strategies to address data drift, model decay, and performance degradation.

### B. Alert generation

The CPS generates multi-source alerts indicating abnormal or anticipated conditions derived from raw and processed data, AI inferences, operational decisions, and manual observations. Alerts are defined across domains, including animal health and welfare (e.g., lameness, estrus, disease detection [23], [25]), environment and infrastructure (e.g., thermal stress, air quality, ventilation issues [26]), production and processes (e.g., milk quality changes, production drops, feeding or milking disruptions [27]), logistics and supply chain (e.g., delivery delays, stock shortages), and system-level issues (e.g., sensor failures, communication loss, AI degradation). AI can be used for alert generation, such as predictive risk detection, early warning signals, cross-domain correlation, operational optimization, and strategic or economic impact assessment [29]. The alert life cycle begins when an issue is detected through sensors, KPIs, system events, or AI outputs. It is then validated against historical data and domain rules to reduce false positives, before triggering AI-driven or rule-based actions. The cycle ends with resolution once normal conditions are restored, while persistent or worsening issues may be escalated for further analysis.

### C. Decision-Making Module: Intelligent Operational Support

We propose a Decision-Making Module (DMM) including the following key functionalities:

- *Contextual Analysis and Recommendation Generation*: produces contextual recommendations from predictive models and analytics, enabling actor review and validation.
- *System-wide and Actor-specific Fusion*: fuses CPS and actor inputs to generate coherent, actionable proposals. Fusion occurs at three levels: automatic fusion for rule-based routine events; AI-assisted fusion for complex cases using ML, ensembles, or consensus methods; and hybrid fusion, where actors refine or adjust outputs using contextual knowledge and constraints. Methods to support information fusion include rule-based aggregation, MCDA (AHP, TOPSIS, VIKOR), outranking (ELECTRE), and fuzzy/probabilistic approaches [22].
- *Multi-Actor, Multi-Perspective, and Shared Views Generation*: generates actor-specific and shared views by integrating CPS and actor inputs into a coherent system-wide representation, supporting coordinated decision-making through aggregated recommendations, fused inputs, and proposed actions. It interprets events, scores, and recom-

TABLE I: Functional overview of Cyber Layer modules in the proposed CPS architecture.

| Cyber Layer                          | Modules                          | Function(s)                                                                                                                                                                                                                                                      |
|--------------------------------------|----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data Treatment & Storage Layer       | Data Processing & Integration    | Collects, cleans, and integrates data from multiple sources (sensors, systems, actors).                                                                                                                                                                          |
|                                      | Data Storage (Databases)         | Stores structured, historical, and contextual data for analysis, traceability, and learning.                                                                                                                                                                     |
|                                      | AI Models Repository             | Stores trained AI models and makes them available for inference and updates.                                                                                                                                                                                     |
| Intelligent Decision & Control Layer | AI Core Module                   | Provides adaptive, data-driven intelligence that supports perception, diagnosis, prediction, decision-making, optimization, and learning processes. Delivers actionable insights to enable intelligence and advanced functionalities across intelligent modules. |
|                                      | Alert Generator                  | Detects anomalies or threshold violations and generates alerts. Classifies alerts into automatic or decision-required.                                                                                                                                           |
|                                      | Decision-Making Module           | Serves as the core component for making decisions by analyzing heterogeneous data from users and other modules, providing recommendations, and preparing validated decisions for execution to the Command Generator.                                             |
|                                      | Automated Systems/AI Agents      | Translates validated decisions or automatic alerts into executable commands for actuators or operational actions. Manages execution logic.                                                                                                                       |
|                                      | Command Generator                | Translates validated decisions or automatic alerts into executable commands for actuators or operational actions. Manages execution logic.                                                                                                                       |
| Application Layer                    | Performance & KPI Dashboard      | Calculates performance indicators for system monitoring and performance tracking, and supports two-way interaction with actors to adjust views, filter data, or provide feedback.                                                                                |
|                                      | Alerts Dashboard                 | Displays alerts, allows actors to create, acknowledge, deactivate, or prioritize alerts, and provides situational awareness. Supports two-way interaction between actors and the dashboard for alert management and response.                                    |
|                                      | Collaborative Decision Dashboard | Provides actors with personalized views of system recommendations and a shared space to view peers' proposals, submit, validate, or adjust decisions, enabling two-way interaction between actors and the system.                                                |
|                                      | Configuration & Roles Dashboard  | Allows actors to define system parameters (thresholds, AI models, rules, priorities) applied across all modules, and supports two-way interaction to view current settings, adjust parameters, and receive system feedback.                                      |

mendations from multiple actor perspectives to produce coherent, domain-informed proposals.

- *Decision Consolidation and Prioritization*: consolidates validated recommendations into final decisions and supports automatic, human, or hybrid decision-making, while prioritizing actions under operational constraints for command execution.
- *AI-Driven Intelligence*: the DMM interacts with the AI Core Module to enable adaptive, context-aware reasoning, improving information integration and decision support. It also enables predictive analytics and machine learning to enhance decision quality and responsiveness.

#### D. Alert-Driven Collaborative Decision Workflow

This section presents the Collaborative Decision Dashboard through an alert-driven workflow, illustrating how multiple actors interact and how shared decision support facilitates coordination and prioritization. The workflow represents a specific decision scenario, while additional use case-driven workflows are presented in the following section.

- 1) *Alert Generation*: The system generates and distributes alerts to the DMM and relevant actors.
- 2) *Collaborative interaction and validation*:
  - The Collaborative Dashboard displays actor-specific and shared views from the DMM, providing system recommendations and aggregated proposals;

- Actors provide recommendations, feedback, and adjustments based on their domain perspective;
- The dashboard enables requests for additional analyses, scenario exploration, and alternative evaluations;
- The dashboard supports coordination, integration of perspectives, and validation of recommendations to produce coherent decisions.

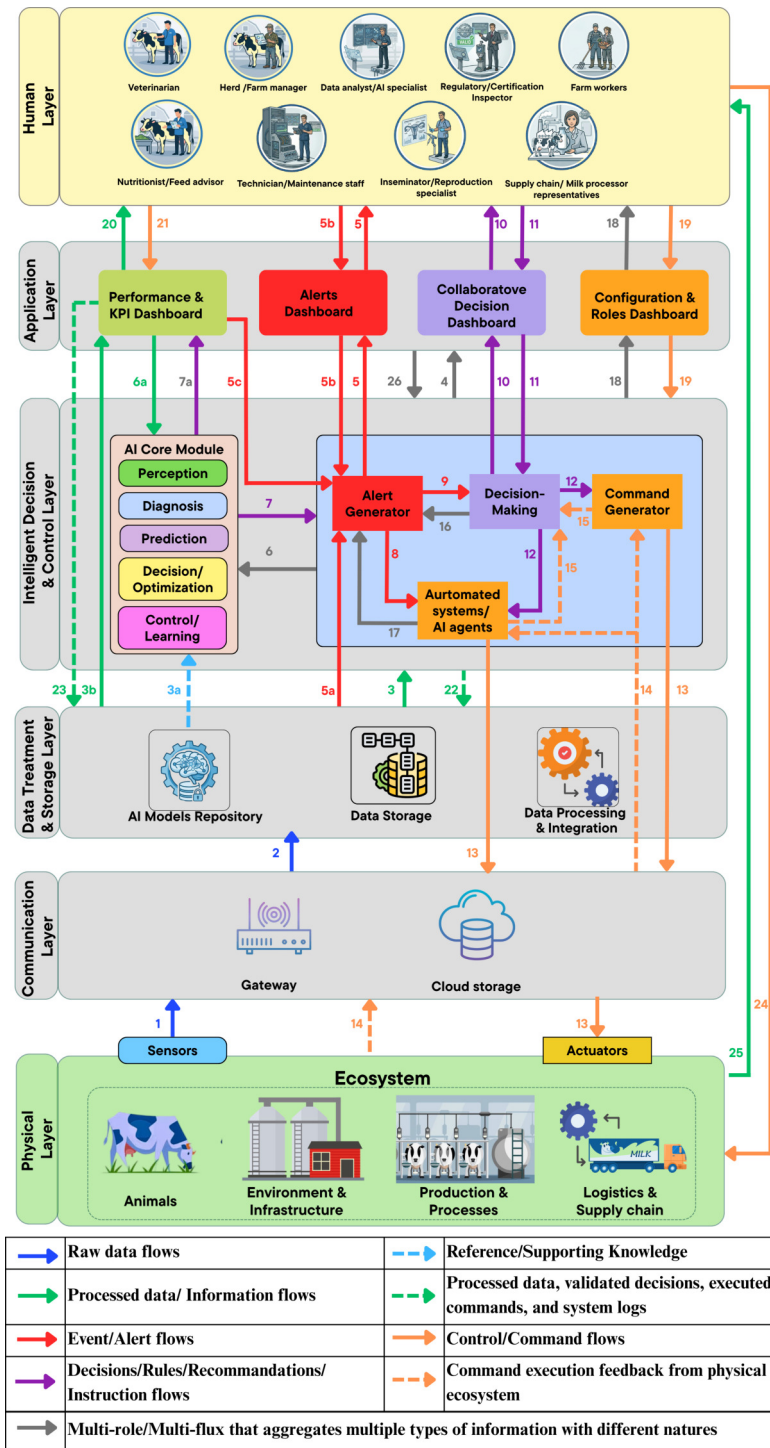
- 3) *Decision preparation & Command execution*: The DMM prepares validated decisions and transmits them to the Command Generator or automated CPS systems for execution.
- 4) *Feedback loop*: Feedback from execution is processed to update system modules (command generator, AI agents, decision-making module, alert generator, and dashboards) and stored as system state data.

## VI. REQUIREMENT COMPLIANCE & USE CASE-DRIVEN WORKFLOWS

This section presents the compliance of the proposed holistic CPS architecture with the requirements defined in Section III. Use case-driven workflows are introduced to illustrate its application in different operational scenarios, highlighting the end-to-end system behavior. The satisfaction of all requirements is demonstrated through the illustrations and details provided in Table II.

TABLE II: Requirement compliance &amp; use case-driven workflows.

| Requirement Compliance and use cases |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| RQ1                                  | <p>The Physical Layer corresponds to the PLF ecosystem, composed of four domains (see Section IV-B).</p> <ul style="list-style-type: none"> <li>· [Ecosystem - Sensors] → (1) → [Communication Layer]</li> <li>· [Ecosystem] → (25) → [Human Layer-Actors] → (24) → [Ecosystem]</li> <li>· [Command Generator] → (13) → [Communication Layer] → (13) → [Actuators-Ecosystem] → (14) → [Communication Layer] → (14) → [Command Generator]</li> </ul>                                                                                                                                                                                                                                                                                                                                                |
| RQ2                                  | <p>The Human Layer integrates multiple stakeholders. Their roles, domains, and actions are defined in Section IV-A</p> <p>Bidirectional interactions: (5; 5b), (10;11), (18;19) and (24;25).</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| RQ3                                  | [Human Layer-Actors] → (21) → [Performance & KPI Dashboard] → (22) → [Human Layer-Actors]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| RQ4                                  | [Human Layer-Actors] → (5b) → [Alerts Dashboard] → (5) → [Human Layer-Actors]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| RQ5                                  | [Human Layer-Actors] → (11) → [Collaborative Decision Dashboard] → (10) → [Human Layer-Actors]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| RQ6                                  | [Human Layer-Actors] → (19) → [Configuration & Roles Dashboard] → (18) → [Human Layer-Actors]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| RQ7                                  | <p>The Alert Generator module processes multi-source data originating from various system modules and actors (see Section V-B).</p> <ul style="list-style-type: none"> <li>· [Human Layer-Actors] → (5b) → [Alert Generator]</li> <li>· [Performance &amp; KPI Dashboard] → (5c) → [Alert Generator]</li> <li>· [Data Treatment &amp; Storage Layer (real-time event streams)] → (5a) → [Alert Generator]</li> <li>· [Data Treatment &amp; Storage Layer (complex events)] → (3) → [Alert Generator]</li> <li>· [AI Core Module] → (7) → [Alert Generator]</li> <li>· [DMM] → (16) → [Alert Generator]</li> </ul>                                                                                                                                                                                  |
| RQ8                                  | <ul style="list-style-type: none"> <li>· Sensor-based: [Ecosystem - Sensors] → (1) → [Communication Layer] → (2) → [Data Treatment &amp; Storage Layer]</li> <li>· Human-based: [Ecosystem] → (25) → [Human Layer-Actors] → (11) → [DMM]</li> <li>· Human-based: [Ecosystem] → (25) → [Human Layer-Actors] → (5b) → [Alerts Dashboards]</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| RQ9                                  | <ul style="list-style-type: none"> <li>· Raw data transformation: [Ecosystem - Sensors] → (1) → [Communication Layer] → (2) → [Data Processing &amp; Integration]</li> <li>· Results storage: [Intelligent Decision &amp; Control Layer] → (22) → [Data Storage (Databases)]</li> <li>· Results storage: [Performance &amp; KPI Dashboard] → (23) → [Data Storage (Databases)]</li> <li>· AI models access: [AI Models Repository] → (3a) → [AI Core Module]</li> <li>· Processed data distribution: 3b, 3, 5a</li> </ul>                                                                                                                                                                                                                                                                          |
| RQ10                                 | <p>The AI Core Module, supported by AI models stored in the Data Layer, consists of five components covering perception, diagnosis, prediction, decision-making/optimization, and adaptive control/learning (see Section V-A).</p> <ul style="list-style-type: none"> <li>· AI performance requests and responses: [Performance &amp; KPI Dashboard] → (6a) → [AI Core Module] → (7a) → [Performance &amp; KPI Dashboard]</li> <li>· Alert requests and responses: [Alert Generator] → (6) → [AI Core Module] → (7) → [Alert Generator]</li> <li>· AI Decision requests and responses: [DMM] → (6) → [AI Core Module] → (7) → [DMM]</li> <li>· AI Agents requests: [Automated Systems/AI Agents] → (6) → [AI Core Module] → (7) → [Automated Systems/AI Agents]</li> </ul>                         |
| RQ11                                 | <ul style="list-style-type: none"> <li>· [Alert Generator] → (9) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM]</li> <li>· [Actors] → [Collaborative Decision Dashboard] → (11) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM]</li> </ul> <p>The loop (10-11) is repeated iteratively until validation is achieved.</p>                                                                                                                                                                                                                                                                                                                                                                                                              |
| RQ12                                 | <ul style="list-style-type: none"> <li>· Automatic mode : [Alert Generator] → (9) → [DMM]</li> <li>· AI-assisted mode : [Alert Generator] → (9) → [DMM] → (6) → [AI Core Module] → (7) → [DMM]</li> <li>· Human mode : [Alert Generator] → (9) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM]</li> <li>· Hybrid mode : [Alert Generator] → (9) → [DMM] → (6) → [AI Core Module] → (7) → [DMM] → (10) → [Collaborative Decision Dashboard] → [Actors] → (11) → [DMM]</li> </ul>                                                                                                                                                                                                                                                                                      |
| RQ13                                 | <ul style="list-style-type: none"> <li>· Automated Command Execution: [Alert Generator] → (8) → [Automated systems] → (13) → [Actuators]</li> <li>· Human Command Execution: [Alert Generator] → (9) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (24) → [Ecosystem]</li> <li>· Code-generated Command Execution: [Alert Generator] → (9) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM] → (12) → [Command Generator] → (13) → [Actuators]</li> <li>· Hybrid Command Execution: [Alert Generator] → (9) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM] → (12) → ([Command Generator], [Automated systems]) → (13) → [Actuators]</li> </ul> |
| RQ14                                 | [Ecosystem] → (14) → [Command Generator] → (15) → [DMM] → (16) → [Alert Generator] → (5) → ([Alerts Dashboard],[Actors])                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| RQ15                                 | <p>Raw data transformation:</p> <p>[Ecosystem - Sensors] → (1) → [Communication Layer] → (2) → [Data Processing &amp; Integration] → (5a) → [Alert Generator] → (9) → [DMM] → (6) → [AI Core Module] → (7) → [DMM] → (10) → ([Collaborative Decision Dashboard],[Actors]) → (11) → [DMM] → (12) → [Command Generator] → (13) → [Ecosystem-Actuators] → (14) → [Command Generator] → ((15) → [DMM] → (16) → [Alert Generator] → (5) → ([Alerts Dashboard],[Actors])) and ((22) → [Data Storage (Databases)])</p>                                                                                                                                                                                                                                                                                    |
| RQ16                                 | This requirement has been addressed. The system adopts a modular architecture ensuring flexibility and scalability, with clear interactions between components and well-defined data flows (see Figure 1a and Figure 1b).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| RQ17                                 | This requirement has been addressed. The system adopts a generic layered CPS architecture not tied to any specific operational setting, ensuring adaptability across diverse PLF environments.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |



(a)

| Flow | Description                                                                                |
|------|--------------------------------------------------------------------------------------------|
| 1    | Raw sensor data acquisition                                                                |
| 2    | Raw data transmission                                                                      |
| 3    | Processed, structured, and contextualized data                                             |
| 3a   | Trained AI models, ML assets, reference data                                               |
| 3b   | Processed KPI data                                                                         |
| 4    | System outputs and insights                                                                |
| 5a   | Filtered data for alert generation                                                         |
| 5b   | Human-created or adjusted alerts                                                           |
| 5c   | KPI threshold / metric-based alerts                                                        |
| 5    | Active alerts and criticality                                                              |
| 6    | Information and requests for AI processing                                                 |
| 6a   | KPI requests and data for advanced analytics                                               |
| 7    | Intelligence and recommendations for modules                                               |
| 7a   | AI capabilities to support the KPI module                                                  |
| 8    | Alert-driven automated action execution                                                    |
| 9    | Alerts needing decision-making                                                             |
| 10   | System proposals, exploration results, AI-assisted analyses and views for actor validation |
| 11   | Actor decisions and requests                                                               |
| 12   | Validated decisions / instructions for execution                                           |
| 13   | Action command execution                                                                   |
| 14   | Execution feedback / Action confirmation                                                   |
| 15   | Feedback and operational metrics                                                           |
| 16   | Decision results and feedback                                                              |
| 17   | Execution feedback and insights                                                            |
| 18   | Confirmation, settings and guidance                                                        |
| 19   | Parameters, rules and thresholds                                                           |
| 20   | Indicators, status and metrics                                                             |
| 21   | Customization, filtering and queries                                                       |
| 22   | Processed data, logs, decisions, actions, alerts and predictions                           |
| 23   | Processed indicators                                                                       |
| 24   | Manual operations on the ecosystem                                                         |
| 25   | Operator observation and results                                                           |
| 26   | Actor inputs, proposals and validation                                                     |

(b)

Fig. 1: (a): Proposed holistic CPS architecture, (b): Module interactions and flow descriptions in the CPS architecture.

## VII. CONCLUSION

This paper presents a holistic Cyber-Physical System (CPS) architecture for Smart Precision Livestock Farming, bridging existing gaps in cross-domain integration, AI capabilities, and multi-stakeholder coordination. By adopting a multi-level, modular design, the framework structures perception, data transmission, decision-making, execution, and feedback into clearly defined components, enabling flexible end-to-end workflows and human-in-the-loop interactions across system operations. In contrast to traditional farmer-centric or centrally driven approaches, we introduced a shared collaborative dashboard and workflows that enable diverse stakeholders to interact, coordinate, and collectively validate through multi-view and multi-perspective views tailored to their roles. While this enhances visibility, transparency, and efficiency in livestock management, future work requires the further refinement of collaboration mechanisms, including role-based interactions, approval sequences, and coordination rules. Additionally, addressing interoperability challenges stemming from the heterogeneity of IoT devices, sensor networks, and software services remains a key priority, motivating the need for standardized mechanisms for data exchange, alignment, and system compatibility to ensure consistent implementation and reuse.

## REFERENCES

- [1] C. Aquilani, A. Confessore, R. Bozzi, F. Sirtori, & C. Pugliese, "Review: Precision Livestock Farming technologies in pasture-based livestock systems", *Animal*, vol. 16, 100429, 2022. <https://doi.org/10.1016/j.animal.2021.100429>
- [2] C.J. Rutten, A.G.J. Velthuis, W. Steeneveld, & H. Hogeveen, "Invited review: sensors to support decision-making in precision livestock farming", *J. Dairy Sci.*, vol. 96, no. 4, pp. 1928–1952, 2013. <https://doi.org/10.3168/jds.2012-6107>
- [3] P. Niloofar, I. Halachmi, & D. Berckmans, et al., "Data-driven decision support in livestock farming for improved animal health, welfare and greenhouse gas emissions: Overview and challenges", *Computers and Electronics in Agriculture*, vol. 190, no. 5, 106406, 2021. <https://doi.org/10.1016/j.compag.2021.106406>
- [4] S. Wolfert, L. Ge, C. Verdouw, & M. J. Bogaardt, "Big data in smart farming—A review", *Agricultural Systems*, vol. 153, pp. 69–80, 2017. <https://doi.org/10.1016/j.agsy.2017.01.023>
- [5] Krampe, C., Ingenbleek, P. T., Niemi, J. K., & Serratos, J., "Designing precision livestock farming system innovations: A farmer perspective", *Journal of Rural Studies*, vol. 111, 103397, 2024. <https://doi.org/10.1016/j.jrurstud.2024.103397>
- [6] ClearFarm Project, "Co-designed Welfare Monitoring Platform for Pig and Dairy Cattle", Horizon 2020 Research and Innovation Programme, Grant Agreement No. 862919, 2019-2024. <https://clearfarm.eu>
- [7] C. L. Sumner, M. A. G. von Keyserlingk, & D. M. Weary, "Perspectives of farmers and veterinarians concerning dairy cattle welfare", *Animal Frontiers*, vol. 8, no. 1, pp. 8–13, 2018. <https://doi.org/10.1093/af/vfx006>
- [8] B. E. Akinyemi, L. Jessiman, S. P. Turner, & J. M. Siegford, "Beyond the farm: stakeholder perspectives on precision livestock farming in the swine industry", *Frontiers in Animal Science*, vol. , 1710969, 2025. <https://doi.org/10.3389/fanim.2025.1710969>
- [9] B. Jiang, W. Tang, L. Cui, & X. Deng, "Precision livestock farming research: A global scientometric review", *Animals*, vol. 13, no. 13, 2096, 2023. <https://doi.org/10.3390/ani13132096>
- [10] R. Mkadmi, R.Y. Douss, & A. Benazza-Benyahia, "Cow monitoring system based on deep learning models for multiple objects detection and tracking". In *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, 1–8. <https://doi.org/10.1109/AICCSA63423.2024.10912629>
- [11] J. L. Kleen & R. Guatteo, "Precision Livestock Farming: What Does It Contain and What Are the Perspectives?", *Animals*, vol. 136, no. 5, 779, 2023. <https://doi.org/10.3390/ani13050779>
- [12] D. Distanto, C. Albanello, H. Zaffar, S. Faralli, and D. Amalfitano, "Artificial intelligence applied to precision livestock farming: a tertiary study", *Smart Agricultural Technology*, vol. 11, 100889, 2025. <https://doi.org/10.1016/j.atech.2025.100889>
- [13] Boyer, C. N., Cavasos, K. E., Greig, J. A., & Schexnayder, S. M., "Influence of risk and trust on beef producers' use of precision livestock farming". *Computers and Electronics in Agriculture*, vol. 218, pp. 108641, 2024. <https://doi.org/10.1016/j.compag.2024.108641>
- [14] X. Du, M. Yu, Z. Zhang, M. Tong, Y. Zhu & C. Xue, "A Task- and Role-Oriented Design Method for Multi-User Collaborative Interfaces". *Sensors*, vol. 25, no. 6, 1760, 2025. <https://doi.org/10.3390/s25061760>
- [15] A. Rohan, M. S. Rafiq, M. J. Hasan, F. Asghar, A. K. Bashir, & T. Dottorini, "Application of deep learning for livestock behaviour recognition: a systematic literature review", *Computers and Electronics in Agriculture*, vol. 224, 109115, 2024. <https://doi.org/10.1016/j.compag.2024.109115>
- [16] Y. Wang, S. Muecher, W. Wang, L. Guo, & L. Kooistra, "A review of three-dimensional computer vision used in precision livestock farming for cattle growth management", *Computers and Electronics in Agriculture*, vol. 206, 107687, 2023. <https://doi.org/10.1016/j.compag.2023.107687>
- [17] H. Zhao, D. Hong, J. Wang, & R. Ma, "Advancing livestock facial recognition with AI: from algorithm innovation to end-to-end precision farming application", *AgriEngineering*, vol. 8, no. 3, 77, 2026. <https://doi.org/10.1016/j.compag.2023.107687>
- [18] A. Montalvo, O. Camacho, & D. Chavez, "The impacts of precision livestock farming tools on the greenhouse gas emissions of an average Scottish dairy farm", *Sustainability*, vol. 17, no. 14, 6393, 2025. <https://doi.org/10.3389/su.2024.1385672>
- [19] A. Nsabiyeze, M. Zhang, J. Li, Q. Zhao, & X. Zhang, "Precision livestock farming for climate-resilient livestock management: a review of real-time monitoring and decision support systems", *Journal of Cleaner Production*, vol. 524, 146454, 2025. <https://doi.org/10.1016/j.jclepro.2025.146454>
- [20] U. Kaur, "Cyber-Physical Systems with robots and AI for precision dairy farming", *Journal of Animal Science*, vol. 102, pp. 297–298, 2024. <https://doi.org/10.1093/jas/skae234.340>
- [21] S. J. Russell & P. Norvig, *Artificial Intelligence: A Modern Approach*, (4th ed.), Pearson, 2020.
- [22] S. Greco, M. Ehrgott, & J. R. Figueira, Eds., *Multiple Criteria Decision Analysis: State of the Art Surveys*, New York, NY, USA: Springer, 2016.
- [23] S. S. Narli, H. Schmidt, A. Firouzabadi, L. Schönnagel, M.S. Reich, & S. Reitmaier, "Automated detection of lameness in dairy cattle through computer vision analysis of back shape characteristics", *Computers in Biology and Medicine*, vol. 197, 111038, 2025. <https://doi.org/10.1016/j.combiomed.2025.111038>
- [24] T. Lodkaew, K. Pasupa, & C.K. Loo, "CowXNet: An automated cow estrus detection system". *Expert Systems with Applications*, vol. 211, 118550, 2023. <https://doi.org/10.1016/j.eswa.2022.118550>
- [25] R. Lardy, Q. Ruin, Q. I. & Veissier, "Discriminating pathological, reproductive or stress conditions in cows using machine learning on sensor-based activity data". *Computers and Electronics in Agriculture*, vol. 204, 107556, 2023. <https://doi.org/10.1016/j.compag.2022.107556>
- [26] G. Provolto, C. Brandolese, M. Grotto, A. Marinucci, N. Fossati, O. Ferrari, E. Beretta, & E. Riva, "An Internet of Things Framework for Monitoring Environmental Conditions in Livestock Housing to Improve Animal Welfare and Assess Environmental Impact". *Animals*, vol. 15, no.5, 644, 2025. <https://doi.org/10.3390/ani15050644>
- [27] F. M. Tangorra, E. Buoio, A. Calcante, A. Bassi, & A. Costa, "Internet of Things (IoT): Sensors Application in Dairy Cattle Farming". *Animals*, vol. 14, no.21, 3071, 2024. <https://doi.org/10.3390/ani14213071>
- [28] M. Mansour, A. Gamal, A. I., Ahmed, L. A. Said, A., Elbaz, N., Herencsar, & A. Soltan, "Internet of Things: A Comprehensive Overview on Protocols, Architectures, Technologies, Simulation Tools, and Future Directions". *Animals*, vol. 16, no.8, 3465, 2023. <https://doi.org/10.3390/en16083465>
- [29] S. Karatsiolis, P. Panagi, V. Vassiliades, A. Kamilaris, N. Nicolaou & E. StavrakisR, "Towards understanding animal welfare by observing collective flock behaviors via AI-powered Analytics". In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 643–648. <http://dx.doi.org/10.15439/2024F2064>

# Semantic Matching of IT Vocational Training Offers with the ESCO Ontology

Katerina Kostadinovska\*<sup>†</sup> Kristine Hein\*

0009-0002-0258-0857 0009-0005-6414-4294

\*Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany

<sup>†</sup>TH Köln – University of Applied Sciences, Cologne, Germany

Email: katerina.kostadinovska@bibb.de, kristine.hein@bibb.de

**Abstract**—This paper presents a pipeline for semantically matching IT vocational training offers from the KURSNet database to skills defined in the ESCO ontology. The proposed approach combines semantic IT filtering, LLM-based keyword cleaning, and embedding-based matching using Sentence Transformer models. Applied to a dataset of 6,128 IT training courses with 16,372 unique keywords, the pipeline successfully maps 63.4% of keywords to ESCO skills at a cosine similarity threshold of 0.75. Evaluation against a manually annotated gold standard of 1,000 keyword-ESCO pairs yields a precision of 61.8%, with an inter-annotator agreement of  $\kappa = 0.91$ . The results demonstrate that embedding-based methods are suitable for bridging semi-structured national training data with standardized European competency frameworks.

## I. INTRODUCTION

THE alignment of national vocational training data with standardized European competency frameworks represents a key challenge for labor market transparency and qualification comparability [1]. In Germany, KURSNet – operated by the Federal Employment Agency – serves as the central database for continuing education offers, describing courses through heterogeneous, non-standardized keywords [2]. The European Skills, Competences, Qualifications and Occupations ontology (ESCO) provides a multilingual reference framework of over 13,000 skill concepts, enabling structured competency modeling across European labor markets [3].

Despite the complementary nature of these systems, no systematic mapping between KURSNet keywords and ESCO skills exists. Embedding-based approaches using Sentence Transformers have shown promising results for semantic similarity tasks across domains [4], [5], suggesting their potential for bridging this gap. This paper presents a multi-step pipeline that combines semantic filtering, LLM-based keyword cleaning, and bi-encoder matching to automatically align IT vocational training offers with ESCO skills.

## II. BACKGROUND AND RELATED WORK

### A. KURSNet and ESCO

KURSNet, operated by the German Federal Employment Agency (BA), serves as the central database for continuing education offers in Germany. Training courses are described through semi-structured fields including titles, content descriptions, and freely assigned keywords [2]. These keywords represent the primary semantic units for content description,

yet exhibit significant terminological heterogeneity – varying in granularity, language (German/English), and specificity. The European Skills, Competences, Qualifications and Occupations (ESCO) ontology provides a multilingual reference framework of over 13,000 skill concepts, organized in a three-pillar structure covering occupations, skills, and qualifications [3]. ESCO follows Linked Open Data principles, using RDF and SKOS to model hierarchical (*broader/narrower*) and associative (*related*) relations between concepts [6], [7]. Each concept is identified by a unique URI, enabling interoperability across systems. The structural gap between KURSNet’s free-text keywords and ESCO’s formal ontology motivates the need for semantic matching approaches.

### B. Structural Gap Between KURSNet and ESCO

Although both systems address vocational content, they differ fundamentally in their structural, terminological, and semantic design, as summarized in Table I. KURSNet stores information in field-structured format with hybrid keyword assignment – combining systematic classification, machine-generated terms, and manual additions by training providers [2] – resulting in terminological heterogeneity and varying levels of granularity. ESCO, by contrast, models occupations, skills, and qualifications as nodes in a formal semantic network, connected through hierarchical (*broader/narrower*) and associative relations [3].

TABLE I: Structural comparison of KURSNet and ESCO

| Dimension        | KURSNet                               | ESCO                                        |
|------------------|---------------------------------------|---------------------------------------------|
| System type      | Operational information system        | Ontological knowledge structure             |
| Data structure   | Field-based, semi-structured keywords | Graph-based with explicit relations         |
| Terminology      | Hybrid, partly free-form              | Controlled multilingual vocabulary          |
| Relations        | No explicit semantic relations        | broader/narrower, essential/optional skills |
| Interoperability | Limited                               | High (RDF, SKOS, URI)                       |

This structural gap motivates the need for semantic matching approaches that capture meaning similarity beyond exact lexical overlap [8], [9]. For example, the KURSNet keyword *Teamführung* (team leadership) would not match the ESCO

skill *Mitarbeiter führen* (manage employees) under a purely lexical approach, despite their semantic equivalence. Similarly, *Machine Learning* would fail to match *maschinelles Lernen einsetzen* (apply machine learning) due to cross-lingual variation. These examples illustrate that lexical methods are insufficient for bridging the terminological gap between KURSNet and ESCO keywords.

### C. Semantic Matching Approaches

Semantic matching of competency data has been approached through both ontology-based and embedding-based methods. Ontology-based approaches leverage explicit knowledge structures and hierarchical relations between skill concepts [10], [11], while Fernández-Sanz et al. demonstrate how different competency frameworks can be systematically mapped to the ESCO taxonomy [12]. Snijder et al. investigate ontology alignment methods between labor market ontologies such as ESCO and O\*NET [13].

Embedding-based approaches model semantic similarity through learned vector representations [8], [14]. SentenceBERT [4] enables efficient similarity computation between text segments, while domain-adapted models such as VacancyBERT [15] demonstrate improved matching quality for job market data. Bhola et al. show that embedding-based methods outperform lexical approaches for skill extraction from unstructured text [16]. However, the systematic mapping of national vocational training databases with semi-structured keyword data to ESCO remains underexplored, motivating the present work.

## III. METHODOLOGY

The proposed pipeline consists of four sequential steps for semantically aligning IT vocational training keywords from KURSNet with ESCO skills, as illustrated in Fig. 1.

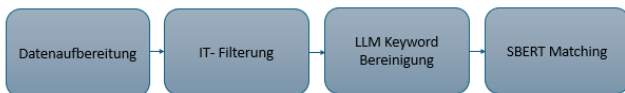


Fig. 1: Multi-step pipeline for semantic matching of IT vocational training offers to ESCO skills.

### A. Data Preprocessing and Deduplication

The raw KURSNet dataset contains 60,540 training offers described through freely assigned keywords. As multiple offers may share identical keyword sets, deduplication is performed based on keyword set identity – two offers are considered identical if they share exactly the same combination of keywords, regardless of order. This keyword-based deduplication is justified by the analytical focus of this work: since the subsequent matching pipeline operates exclusively on keywords, two offers with identical keyword sets are semantically indistinguishable for the purpose of this analysis. This reduces the dataset to 26,872 distinct offers. The deduplicated data is

subsequently transformed into long format, where each keyword constitutes an individual observation, yielding 851,626 rows representing 98,841 unique keywords. This format serves as the basis for all subsequent steps.

### B. Semantic IT Filtering

Since KURSNet covers all domains of continuing education, a semantic filtering step is applied to identify IT-relevant offers. The multilingual Sentence Transformer *paraphrase-multilingual-MiniLM-L12-v2* [4] is used to encode each course title against ten manually defined IT-domain anchor texts (e.g., *Software Development*, *Cybersecurity*, *Cloud Computing*). Each course is assigned the maximum cosine similarity score across all anchors. The score distribution shows a mean of  $\bar{x} = 0.373$  and a 75th percentile of  $P_{75} = 0.468$ . A threshold of  $\tau = 0.48$  is selected through iterative inspection of borderline cases – a qualitative review of ten boundary cases confirmed correct IT classification in eight instances – yielding 6,128 IT-relevant courses with 16,372 unique keywords.

### C. LLM-Based Keyword Cleaning

The raw keywords exhibit significant heterogeneity: generic meta-terms (e.g., *Weiterbildung*, *Lehrgang*, *Fachkraft*), redundant entries, synonyms, and varying levels of specificity coexist within the dataset. Direct use of these raw keywords for ESCO matching would be problematic, as irrelevant terms dilute the semantic signal of the text representations.

Rule-based approaches such as stopwords lists or heuristic filters were considered but deemed insufficient, as they cannot capture semantic relationships between terms and perform poorly under terminological variance or ambiguity. Statistical methods such as clustering or frequency analysis reduce data volume but fail to account for the specific meaning context of individual courses [17], [18].

GPT-4o-mini was selected for its balance between model performance and processing efficiency. For the cleaning task – which focuses on selection, normalization, and semantic condensation – high-complexity reasoning is not required, making smaller models sufficient. Compared to larger language models, GPT-4o-mini enables substantially more cost-efficient and faster processing, which is particularly relevant when scaling to thousands of training offers.

A structured prompt provides the model with the course title as semantic context alongside the full list of original keywords. The course title serves as an interpretive frame, preventing keywords from being evaluated in isolation without reference to the actual course topic. The model is instructed to remove generic or organizational terms such as *Kurs* (course) or *Lehrgang* (training program), merge terminological variants and synonyms to reduce redundancy, prefer precise technical terms or short technical phrases over generic designations, and extract semantically relevant core phrases from longer keyword strings – for instance, reducing *Einführung in die digitale Fertigung* (Introduction to digital manufacturing) to *digitale Fertigung* (digital manufacturing). Output is constrained ex-

clusively to JSON format, enabling automated downstream processing.

A negative list of generic terms provides additional post-processing filtering, retry logic handles API failures automatically, and checkpoint storage ensures fault-tolerant processing of large batches. This step intentionally introduces a methodological trade-off: while synonym merging reduces redundancy, it may also conflate terms that represent distinct ESCO skills – for example, merging *SPS Technik* and *Speicherprogrammierbare Steuerungen* into a single representation. The LLM-based cleaning therefore represents a deliberate compromise between redundancy reduction and potential information loss.

#### D. Bi-Encoder Semantic Matching

The cleaned keywords are matched against ESCO skills using the multilingual Sentence Transformer *paraphrase-multilingual-mpnet-base-v2* [5], which produces 768-dimensional embeddings optimized for semantic similarity tasks. This model was selected over the MiniLM variant used in the filtering step, as the higher dimensionality enables finer-grained semantic differentiation – particularly relevant for distinguishing closely related ESCO skill concepts such as *Softwareentwicklung* and *Anwendungsentwicklung*. Embeddings are computed separately for all unique keywords and for all ESCO preferred labels (*preferredLabel*). Cosine similarity is computed via dot product on L2-normalized vectors. Only the preferred label is used for ESCO representation, as KURSNet keywords are typically short phrases, making label-level comparison structurally appropriate. Matching is performed on unique keywords to avoid redundant computation; results are subsequently mapped back to all corresponding courses.

The threshold  $\tau = 0.75$  was selected based on a systematic evaluation across values from 0.65 to 0.85, as summarized in Table II. Lower thresholds yield higher coverage but lower average similarity, indicating a higher proportion of weaker matches. At  $\tau = 0.70$ , coverage increases to 12,687 keywords but mean similarity drops to 0.79. At  $\tau = 0.80$ , coverage drops substantially to 7,712 keywords, excluding a significant portion of relevant IT keywords. The value  $\tau = 0.75$  represents the optimal trade-off between coverage (63.4%) and semantic quality ( $\bar{x} = 0.82$ ). For each keyword, up to  $K = 10$  ESCO skills with the highest similarity scores are retained (Top-K).

TABLE II: Threshold evaluation results

| Threshold   | Keywords matched | $\bar{x}$ Similarity |
|-------------|------------------|----------------------|
| 0.65        | 14,284           | 0.76                 |
| 0.70        | 12,687           | 0.79                 |
| <b>0.75</b> | <b>10,388</b>    | <b>0.82</b>          |
| 0.80        | 7,712            | 0.85                 |
| 0.85        | 4,992            | 0.89                 |

## IV. RESULTS AND EVALUATION

### A. Matching Results

The pipeline was applied to the filtered dataset of 6,128 IT-relevant training courses. At threshold  $\tau = 0.75$ , 10,388 out of 16,372 unique keywords (63.4%) were successfully matched to at least one ESCO skill, generating 54,390 unique keyword-ESCO pairs across 5,109 distinct ESCO skills. The mean cosine similarity of matched pairs is  $\bar{x} = 0.819$  (median = 0.808), indicating that the majority of matches lie well above the threshold. A rank analysis shows a shallow decline from  $\varnothing 0.85$  (rank 1) to  $\varnothing 0.79$  (rank 10), suggesting that lower-ranked matches remain semantically plausible and justify the Top-K approach.

The most frequently matched ESCO skills reflect the thematic focus of the IT training landscape, with Cloud technologies, project management, and data analysis dominating the top assignments. A total of 213 keyword-ESCO pairs achieve a cosine similarity of  $\geq 0.99$ , typically arising from lexically identical or near-identical label pairs such as *DevOps*  $\rightarrow$  *DevOps* or *agiles Projektmanagement*  $\rightarrow$  *agiles Projektmanagement*.

Table III illustrates the Top-K matching results for the keyword *Data Science*, demonstrating how a single keyword can be associated with multiple semantically related ESCO skills.

TABLE III: Top-K ESCO matches for keyword *Data Science*

| ESCO Skill                | Cosine Similarity |
|---------------------------|-------------------|
| Datenanalyse              | 0.851             |
| Datenanalyse durchführen  | 0.808             |
| Daten sammeln             | 0.798             |
| Datenmodelle              | 0.784             |
| Datensätze erstellen      | 0.777             |
| Forschungsdaten verwalten | 0.760             |
| Daten untersuchen         | 0.759             |
| Daten verarbeiten         | 0.758             |

Qualitative inspection reveals systematic error patterns: morphologically similar but semantically distinct pairs such as *Hafenordnung* (harbor regulations)  $\rightarrow$  *Ports* (network ports) or *Auftragsbearbeitung* (order processing)  $\rightarrow$  *Brauverfahren* (brewing process) demonstrate the limitations of purely similarity-based approaches without additional conceptual modeling.

### B. Evaluation

To assess matching quality, a gold standard of 1,000 keyword-ESCO pairs was manually annotated by three independent annotators using a three-point scale: *yes* (relevant), *approximately* (partially relevant), and *no* (not relevant). Both *yes* and *approximately* were treated as positive matches. The dataset was stratified across five cosine similarity intervals (250/250/250/162/88 pairs), ensuring representative coverage of the full similarity range.

Inter-annotator agreement was measured using Cohen's Kappa [19], yielding  $\kappa = 0.91$ , indicating almost perfect agreement [20]. The overall observed agreement rate was

95.6%, with 44 disagreements resolved by a third annotator whose judgment served as the final gold standard label.

Overall precision of the matching procedure is **61.8%** (618 out of 1,000 pairs rated as relevant). As shown in Table IV, precision increases substantially with cosine similarity score, demonstrating a clear positive correlation between similarity score and match quality.

TABLE IV: Precision by cosine similarity interval

| Interval       | Precision    | n            |
|----------------|--------------|--------------|
| [0.75, 0.80)   | 39.6%        | 250          |
| [0.80, 0.85)   | 57.2%        | 250          |
| [0.85, 0.90)   | 63.2%        | 250          |
| [0.90, 0.95)   | 82.1%        | 162          |
| [0.95, 1.00]   | 96.6%        | 88           |
| <b>Overall</b> | <b>61.8%</b> | <b>1,000</b> |

Matches in the lowest interval [0.75, 0.80) are frequently thematically related but not semantically equivalent, while matches above 0.90 are predominantly correct. The highest interval approaches near-perfect precision, driven by lexically identical label pairs.

## V. DISCUSSION AND LIMITATIONS

### A. Discussion

The proposed pipeline achieves a precision of 61.8%, which represents a solid baseline given the challenging input data. Unlike existing work focusing on structured competency profiles or job postings [15], [16], [10], this work addresses semi-structured, terminologically heterogeneous keyword data – a significantly more demanding setting.

The score-dependent precision analysis confirms that cosine similarity is a reliable indicator of match quality. A clear trade-off is visible: higher thresholds yield substantially better precision but reduce coverage. For practical applications, threshold selection depends heavily on whether precision or recall is prioritized.

The thematic concentration of matches on Cloud computing, IT security, and project management reflects both actual trends in the IT labor market [21] and data-driven effects from keyword frequency distributions. Results should therefore be interpreted as reflecting both content trends and dataset-specific characteristics.

Error analysis reveals that the method’s limitations lie primarily at the boundary between conceptual proximity and semantic equivalence. Morphological similarity – such as *Hafenordnung* (harbor regulations) → *Ports* (network ports) or *Schraubenschlüssel verwenden* (use wrench) → *Primärschlüssel* (primary key) – represents a systematic error source that purely similarity-based approaches cannot reliably resolve, underscoring the need for complementary methods incorporating contextual or ontological information.

### B. Limitations

Several limitations should be acknowledged. First, the LLM-based keyword cleaning was not systematically evaluated; relevant content may be lost or distorted, and the

non-deterministic nature of LLMs limits reproducibility. The chosen prompt formulation further influences results, as synonym merging may conflate terms that represent distinct ESCO skills. Second, the high terminological heterogeneity and varying granularity levels of KURSNet keywords directly affect embedding quality and constrain achievable matching performance – similar concepts are frequently described using different formulations (e.g., *Webentwicklung* vs. *Frontend*), leading to inconsistent matching results. Third, the bi-encoder approach does not model fine-grained interactions between keywords and ESCO skills; cross-encoder models [22] could improve precision for borderline cases by jointly processing both texts, though at substantially higher computational cost for large-scale matching tasks. Fourth, matching relies exclusively on short label texts, disregarding contextual information from course descriptions or metadata; this may cause semantic nuances to be lost, particularly for polysemantic or highly generic terms. Fifth, the dataset is restricted to IT-relevant offers, limiting generalizability to other vocational domains such as healthcare or skilled trades. Finally, evaluation is based solely on precision; recall cannot be computed as the gold standard covers only model-proposed pairs, precluding assessment of false negatives and thus providing only a partial quality evaluation of the matching procedure.

## VI. CONCLUSION AND OUTLOOK

This paper presented a multi-step pipeline for semantically matching IT vocational training keywords from the KURSNet database to skills defined in the ESCO ontology. The approach combines semantic IT filtering, LLM-based keyword cleaning, and bi-encoder matching using multilingual Sentence Transformers.

The results confirm that embedding-based methods are suitable for bridging semi-structured national training data with standardized European competency frameworks, even when matching is performed exclusively on short label texts. From a practical perspective, the approach offers concrete benefits for multiple stakeholders: research institutions such as BIBB can leverage the mappings for data-driven analysis of competency trends within the German continuing education landscape – for instance, identifying which ESCO skills are underrepresented in current IT training offers, or tracking shifts in demand for Cloud computing and cybersecurity competencies over time; public employment agencies could integrate the method into digital training portals to enable competency-based search; and training providers may benefit from automated, standardized competency tagging of their offers. On a policy level, the approach supports data-driven assessment of the alignment between national training landscapes and European competency frameworks, contributing to labor market transparency and qualification comparability.

Several directions for future work emerge. Incorporating full ESCO skill descriptions and alternative labels (*non-preferred terms*) could increase semantic expressiveness, particularly for generic or ambiguous keywords. Extracting competencies

directly from course descriptions rather than assigned keywords may capture implicit skills not reflected in explicit keyword fields, with context-based representations potentially improving matching quality for polysemantic terms. Domain-specific fine-tuning of embedding models on labor market and education-specific texts [15] and cross-encoder re-ranking [22] offer promising avenues for improving precision for borderline cases. A more systematic, data-driven optimization of the similarity threshold  $\tau$  – explicitly balancing the trade-off between precision and coverage – represents another valuable direction. Extending the pipeline to non-IT domains would assess its generalizability and enable broader evaluation of ESCO’s semantic coverage. Finally, a cluster analysis of derived ESCO assignments could reveal latent competency structures, thematic groupings, and potential occupational transition pathways within the continuing education landscape, representing the next step toward a comprehensive data-driven competency analysis of national vocational training systems.

#### ACKNOWLEDGMENTS

During the preparation of this manuscript, the authors used DeepL to check the language and grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

#### REFERENCES

- [1] OECD, *Getting Skills Right: Future-Ready Adult Learning Systems*. OECD Publishing, 2019.
- [2] Bundesagentur für Arbeit, “Eingabehilfe für die erfassung von weiterbildungsangeboten in kursnet,” Bundesagentur für Arbeit, Nürnberg, Tech. Rep., 2025, stand: 31.03.2025.
- [3] European Commission, *ESCO Handbook: European Skills, Competences, Qualifications and Occupations*, 2nd ed. Luxembourg: Publications Office of the European Union, 2019.
- [4] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP*, 2019.
- [5] —, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. [Online]. Available: <https://arxiv.org/abs/2004.09813>
- [6] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data – the story so far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.
- [7] World Wide Web Consortium (W3C), “Skos simple knowledge organization system reference,” <https://www.w3.org/TR/skos-reference/>, 2009, w3C Recommendation.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *ICLR*, 2013.
- [9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.
- [10] T. Petrican and O. Stan, “Ontology-based skill matching algorithms,” in *Proceedings of CSCS*, 2015.
- [11] S. Paudel and S. Shakya, “Ontology based job-candidate matching using skill sets,” in *IOE Graduate Conference*, 2017.
- [12] L. Fernández-Sanz *et al.*, “e-skills match: A framework for mapping skills into esco,” *Computer Standards & Interfaces*, vol. 51, pp. 30–42, 2017.
- [13] L. Snijder *et al.*, “Advancing ontology alignment in the labor market,” in *AAAI-MAKE 2024*, 2024.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019.
- [15] N. Bocharova *et al.*, “Vacancybert: A domain-adapted language model for job matching,” *Applied Sciences*, 2021.
- [16] A. Bhola *et al.*, “Extracting skills from job descriptions using embeddings,” in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.
- [17] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>
- [18] 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 Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [19] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [20] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [21] OECD, *OECD Skills Outlook 2021: Learning for Life*. Paris: OECD Publishing, 2021.
- [22] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. [Online]. Available: <https://arxiv.org/abs/2104.08663>



# Incorporation of *Min/Max*, *With/Without* and *Pairing* Restrictions in Column Generation Solution Algorithms for Aircrew Rostering

George Kozanidis  
0000-0002-4325-2975  
Department of Mechanical  
Engineering, University of  
Thessaly, Volos 38334, Greece  
Email: gkoz@uth.gr

Nikos Polychronopoulos  
Department of Mechanical  
Engineering, University of  
Thessaly, Volos 38334, Greece  
Email:  
n.polychronopoulos@hotmail.com

Andreas Gavranis  
0000-0001-8102-516X  
Department of Mechanical  
Engineering, University of  
Thessaly, Volos 38334, Greece  
Email: agavra@uth.gr

**Abstract**—We consider the column generation solution framework handling the crew rostering problem in the context of commercial aviation. Motivated by practical requirements arising in realistic environments, we modify the algorithmic methodology to enable the incorporation of three special restriction types termed *min/max*, *with/without* and *pairing*, respectively. *Min/max* restrictions impose suitable lower/upper bounds on the number of a specific category crew members assigned to certain flight duties. *With* restrictions ensure that crew members of a specific category assigned to certain flight duties are coupled with crew members of another specific category. *Without* restrictions forbid the simultaneous assignment of crew members belonging to two distinct specific categories to certain flight duties. Finally, *pairing* restrictions dictate that the number of crew members of each of two mutually exclusive, but collectively exhaustive, categories assigned to certain flight duties with even total crew complement is also even. After properly adjusting the algorithmic methodology for accommodating these restrictions, we present computational results demonstrating how their incorporation affects algorithmic behavior and computational performance.

**Index Terms**—Column generation, branch & price, commercial aviation, airline management, crew rostering, *min/max* restrictions, *with/without* restrictions, *pairing* restrictions.

## I. INTRODUCTION

COLUMN generation comprises one of the most popular solution frameworks for the efficient treatment of complex decision-making problems arising in the context of airline management. Focusing on the algorithmic variation that handles the crew rostering problem, we incorporate into the column generation methodology three special types of restrictions for accommodating practical user requirements arising in realistic environments. These restrictions are applicable to flight duties with crew complement strictly greater than 1, i.e., flight duties that must be assigned to more than 1 crew members.

The first restriction type, termed *min/max*, imposes a lower/upper bound on the number of crew members of a specific category assigned to some flight duty. The second restriction type, termed *with/without*, is distinguished into

two sub-types. *With*-type restrictions ensure that the assignment of a specific category crew member to a certain flight duty is only allowed if a crew member of another specific category is also assigned to the same flight duty. *Without*-type restrictions, on the other hand, forbid the simultaneous assignment of crew members belonging to two distinct specific categories to the same flight duty.

The last restriction type, termed *pairing*, necessitates the existence of some attribute/characteristic with only two crew categories involved, such that any crew member always belongs to exactly one of the two by default. Such a restriction dictates that the number of crew members of each of these two categories assigned to some specific duty with even total crew complement must also be even, in order to minimize relevant crew handling costs which increase considerably otherwise.

We develop the methodology for incorporating the above restrictions into the column generation algorithmic variation handling the crew rostering problem. This includes the modification of the master problem formulation for ensuring their satisfaction, as well as the suitable adjustment of the column generation subproblem for ensuring they are properly taken into consideration within the pricing algorithmic procedure.

The main contribution of the present work is the extension of the fundamental column generation framework for enabling the accommodation of special restrictions arising in realistic environments. Besides illustrating how these restrictions can be properly handled, the current research provides valuable insight for the treatment of similar restriction types that may arise in different situations.

The remainder of this work is structured as follows. Section 2 presents a literature review on solution algorithms for aircrew rostering, while Section 3 presents the problem definition and develops the model formulation. Section 4 develops the associated algorithmic implementation, while Section 5 presents computational results drawn from realistic experiments. Finally, Section 6 summarizes this work and suggests fruitful directions for future research.

## II. LITERATURE REVIEW

A large number of decision-making problems arising in the regular operation of commercial aviation, such as tail assignment [1] and trainee assignment [2], necessitate the employment of sophisticated optimization models and solution algorithms. Aircrew rostering [3] comprises one of the most widely studied relevant problems, due to its high complexity, its tremendous significance, and its wide applicability.

Typically, the crew rostering problem is considered within the context of the crew scheduling framework [4], which, besides the crew rostering problem, also includes the problem of constructing pairings through the integration of flight legs, off days, rests, training duties, etc., termed crew pairing [5].

Three are the most common design schemes for the aircrew rostering problem. The first one, often referred to as *fair assignment* in the related literature, optimizes an objective encompassing the duty non-coverage cost, as well as the aggregate soft quality cost of the crew rosters [6]. The second one, typically termed *preferential bidding system* in the related literature, also minimizes the duty non-coverage cost as primal objective, but instead of secondarily optimizing soft quality, it optimizes the cost related to preference satisfaction with respect to specific crew requests that have been expressed in advance [7]. These two schemes comprise two different versions of the so-called *personalized crew rostering framework*. The third one, termed *bidline rostering*, builds a set of anonymous rosters in advance, with the final group schedule obtained after each crew member expresses specific preference for roster assignment [8]. The special restriction types introduced above are considered in the context of the personalized version of the crew rostering problem; therefore, we focus on this problem variation in what follows.

Column generation [9] comprises one of the major breakthroughs for the solution of the aircrew rostering problem. Ryan [10] formulates a set partition optimization model, which employs a specialized column generation approach to solve the LP relaxation of the master problem to near optimality, and then employs branch & bound to obtain an integer solution using the column variables that have been identified so far. Gamache et al. [11] formulate a set-partition optimization model for the aircrew rostering problem, utilizing a heuristic column generation-based methodology for identifying integer solutions.

Fahle et al. [12] develop a column generation heuristic solution algorithm for the problem, which exploits constraint programming in the shortest path pricing subproblem for handling the computational difficulties raised by the complex rules/regulations each crew roster needs to comply with.

The solutions returned by heuristic algorithms often exhibit moderate quality; this has motivated the development of exact methodologies such as branch & price [13] for the treatment of the problem. Branch & price comprises an innovative extension of branch & bound that ensures exact optimality by employing column generation in each node of the branching tree. It has been exploited for the solution of numerous deci-

sion-making problems arising in the regular operation of commercial aviation, such as logistics planning [14], [15], gate assignment [16] and vacation planning [17].

One of the earliest works on branch & price is that of Desrosiers et al. [18], although the specific term for this methodology was later introduced by Barnhart et al. [19] who illustrate its application on the generalized assignment and the crew scheduling problem. Gamache and Soumis [20] present an exact branch & price solution methodology for a rostering problem variation that includes crew requests, which, instead of assigning in advance a specific subset of these requests, incorporates them into the optimization procedure. More information on column generation methodologies for aircrew rostering can be found in the survey of Kasirzadeh et al. [21].

## III. PROBLEM DESCRIPTION & MATHEMATICAL FORMULATION

We consider the crew rostering problem in the context of commercial aviation, which aims to build rosters (schedules) for assignment to a crew member group, while optimizing an aggregate objective function. A crew roster is formed by selecting a proper subset of duties respecting the active legality restrictions imposed by air safety regulations and government/union agreements. The specific duty selection prescribes an associated cost, which reflects the extent to which the roster exhibits certain desirable quality characteristics aimed at increasing crew satisfaction and operational efficiency. The problem objective includes the duty non-coverage cost and the roster quality cost, with the former exhibiting higher priority, due to its indisputable superior significance.

The activities subject to coverage include flight duties called *pairings* and ground duties called *standbys*. Each pairing is formed by connecting a subset of pairwise temporally and locally compatible flight legs, coupled with rest periods. Temporal compatibility necessitates that the arrival time of each flight leg precedes the departure time of the immediately next one, while local compatibility necessitates that the arrival airport of each flight leg coincides with the departure airport of the immediately next one. Each pairing must begin and end at the corresponding crew group airport home base. Each standby, on the other hand, is a stand-alone independent duty, typically characterized by its service location (home, airport), as well as by its service time (morning, evening, night, etc.).

Column generation is a specialized algorithmic procedure for handling linear programming problems with an exponentially large number of decision variables. In the aviation context under consideration, each of these variables corresponds to a crew roster. Instead of explicitly enumerating all of them, a task that would necessitate an extraordinary amount of time, the column generation algorithmic approach formulates a so-called *restricted master problem* including a very small subset of them, which is moderately being enriched with additional beneficial ones until the optimal solution is obtained.

The criterion exploited for identifying these extra variables is their reduced cost with respect to the optimal solution of the master problem LP relaxation. The involved computations are

performed by a shortest path network optimization problem termed *pricing subproblem*. This problem is invoked numerous times along algorithmic execution, each time identifying new beneficial rosters for addition to the master problem.

The pricing subproblem ensures satisfaction of the complex hard rules that each roster needs to respect for the associated crew member in isolation. Typical examples of such restrictions are the imposition of an upper bound on the total flight time, the requirement that a rest period of specific minimum duration is ensured after certain long-lasting duties, and the imposition of a lower bound on the number of off-work days. In addition, each roster must respect the set of preassigned activities which may have been appended in advance to the monthly plan of the crew member under consideration.

On the other hand, rules applicable to the entire crew group are handled at the master problem level. An example of such a restriction is the satisfaction of a specific upper bound on the total number of days off from work assigned collectively to the crew group for a given day of the planning horizon. The special restrictions we consider in this research work must also be satisfied combinedly by all crew members.

For the formulation of the master problem, we introduce the following mathematical notation:

**Sets:**

$I$ : set of crew members, indexed by  $i$ .

$R_i$ : set of rosters of crew member  $i$ , indexed by  $j$ .

$D$ : set of standby/pairing duties, indexed by  $d$ .

**Parameters:**

$c_{ij}$ : quality cost of roster  $j$  of crew member  $i$ ,  $i \in I, j \in R_i$ ,

$h$ : cost for each uncovered unit of duty complement,

$b_d$ : crew complement of duty  $d$ ,  $d \in D$ ,

$a_{ijd}$ : binary parameter that takes the value 1 if roster  $j$  of crew member  $i$  includes (covers) duty  $d$ , and 0 otherwise,  $i \in I, j \in R_i, d \in D$ .

**Decision variables:**

$x_{ij}$ : binary decision variable that takes the value 1 if roster  $j$  is assigned to crew member  $i$ , and 0 otherwise,  $i \in I, j \in R_i$ ,

$y_d$ : integer slack decision variable denoting the number of complement units of duty  $d$  that remain uncovered,  $d \in D$ .

Utilizing the above notation, the master problem is formulated as follows:

$$\text{Min } \sum_{i \in I} \sum_{j \in R_i} c_{ij} \cdot x_{ij} + \sum_{d \in D} h \cdot y_d \quad (1)$$

$$\text{s.t. } \sum_{j \in R_i} x_{ij} = 1, i \in I \quad (2)$$

$$y_d + \sum_{i \in I} \sum_{j \in R_i} a_{ijd} \cdot x_{ij} = b_d, d \in D \quad (3)$$

$$x_{ij} \in B = \{0,1\}, y_d \in Z^+, i \in I, j \in R_i, d \in D \quad (4)$$

The objective function (1) minimizes the total system cost, which includes the cost of uncovered duties plus the roster quality cost. Cost coefficient  $h$  is always much larger than cost coefficients  $c_{ij}$ , imposing the lexicographic priority between the two objective terms. Constraint set (2) includes one multiple choice constraint for each crew member, ensuring the assignment of exactly one among all the alternative rosters. Constraint set (3) monitors duty coverage based on the as-

signed rosters; slack variable  $y_d$  is equal to the number of uncovered complement units, imposing the corresponding objective cost in case it becomes positive. Finally, constraint set (4) imposes integrality restrictions on the decision variables.

#### IV. SPECIAL RESTRICTIONS' INCORPORATION

The crew rostering problem is typically solved separately for each crew position. The complement of a duty is equal to 1 when only one crew member holding the position under consideration needs to be assigned to it. This is common for positions such as CP (captain) and FO (first officer). On the other hand, the number of FAs (flight attendants) that need to be assigned to a duty is naturally an integer greater than 1.

In realistic commercial airline environments, crew members typically belong to specific categories, based on their particular attributes/characteristics. Some examples include categories reflecting experience level (junior, senior, etc.), language knowledge (Chinese, French, German, etc.), and origin (Asian, European, etc.). The three restriction types under consideration pertain exclusively to duties with complement strictly greater than 1, i.e., duties  $d \in D$  for which  $b_d > 1$ . In the following subsections, we elucidate in detail the corresponding requirements and setting parameters for each of them. For this analysis, we extend the mathematical notation introduced in the previous section with two additional sets:

**Sets:**

$G$ : set of crew categories, indexed by  $g$ .

$G_i$ : set of crew categories of crew member  $i$ ,  $i \in I$ .

##### A. Min/max Restriction Type

A *min/max* restriction type imposes a lower/upper bound on the number of crew members of a specific category assigned to some particular duty. For example, a *min*-type restriction may impose the requirement that at least two Chinese speaking crew members are assigned to some pairing, while a *max*-type restriction may impose the requirement that at most one junior crew member is assigned to it.

To accommodate a *max*-type restriction, we add to the master problem formulation a less or equal ( $\leq$ ) constraint with right-hand side equal to the corresponding upper bound. In the left-hand side, this constraint is meant to include the rosters of crew members belonging to the specific category for which the restriction is imposed, provided that they include (cover) the duty to which the rule applies. Let  $M \in G$  and  $p \in D$  denote this crew category and duty, respectively, and  $u$  be the specified upper bound. Mathematically, the *max* restriction is expressed with the following constraint:

$$\sum_{i \in I: M \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \leq u \quad (5)$$

A generated roster is appended to the left-hand side of this constraint, if and only if it satisfies the category-duty combination specified by the rule. Clearly, the inclusion of this constraint and its proper population with rosters ensure satisfaction of the *max*-type restriction upon algorithmic termination.

Consider a *min*-type restriction next. Following a similar approach, one could add to the master problem a greater or

equal ( $\geq$ ) constraint with right-hand side equal to the corresponding lower bound. The decision of whether a generated roster should be appended to the left-hand side of this constraint would similarly necessitate the check of whether it satisfies the corresponding category-duty combination. A difficulty that arises in this case, however, is that this modeling choice leaves open the possibility of infeasibilities, since the presence of crew rosters which are necessary for constraint satisfaction in the left-hand side cannot be guaranteed. To give a particular example, the solution that consists of empty rosters (no duties covered), commonly used for initializing the master problem, cannot be utilized under this modeling approach, because it renders the master problem infeasible.

To circumvent this difficulty, we express a *min*-type restriction as a less or equal constraint with right-hand side equal to the difference between the total complement of the associated duty and the corresponding lower bound imposed on the number of crew members of the specific category. Using a similar but reverse logic as the one adopted in the case of a *max*-type restriction, the left-hand side of this constraint is meant to include rosters covering the duty in question which pertain to crew members that do not belong to the category specified by the rule. This adjustment is clearly equivalent to the imposition of the desirable lower bound on the number of crew members of the specific category prescribed by the rule. This modeling approach handles properly potential master problem infeasibilities, since the constraint is still satisfied even when its left-hand side is empty. Let  $M \in G$  and  $p \in D$  denote the crew category and duty, respectively, for which the restriction applies, and  $l$  be the specified lower bound. Mathematically, this restriction is expressed with the following constraint:

$$\sum_{i \in I: M \notin G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \leq b_p - l \quad (6)$$

### B. With/without Restriction Type

Moving on to the next case, a *with*-type restriction imposes the requirement that for each crew member of a specific category assigned to a particular duty, a crew member of another specific category must be assigned to the same duty, too. For example, a *with*-type restriction may dictate that each junior crew member assigned to some duty must be accompanied by a senior crew member. This one-to-one coupling restriction works only in one direction, i.e., the total number of junior crew members assigned to the duty cannot exceed the total number of senior crew members, whereas the opposite is clearly permissible. For simplicity, we call the former crew members *dependent* and the latter ones *independent*. As far as the involved crew categories are concerned, the convention is that any given crew member can be dependent, independent, or neither of the two, but a crew member cannot be both dependent and independent simultaneously.

To incorporate a *with* restriction, we add to the master problem formulation a less or equal ( $\leq$ ) constraint. This constraint is meant to include in its left-hand side the rosters of dependent crew members covering the duty in question, and in its right-hand side the rosters of independent crew members

covering this duty. This ensures that the number of dependent crew members to which the duty will be assigned will not exceed the number of independent ones, without imposing any restriction, whatsoever, on the latter ones. Let  $M_1 \in G$  denote the dependent crew category,  $M_2 \in G$  denote the independent one, and  $p \in D$  denote the duty for which the restriction applies. Mathematically, this restriction is expressed with the following constraint:

$$\sum_{i \in I: M_1 \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \leq \sum_{i \in I: M_2 \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \quad (7)$$

A *without*-type restriction, on the other hand, imposes the requirement that the number of crew members assigned to a given duty must be equal to 0 for at least one of two specified crew categories, the convention being again that any crew member can belong to neither or one of the two, but no crew member can belong to both of them at the same time.

To incorporate this restriction, we add to the master problem formulation two less or equal ( $\leq$ ) constraints, as well as a binary auxiliary decision variable. The right-hand side of the first constraint is set equal to the duty complement multiplied by this variable, while that of the second constraint is set equal to the duty complement multiplied by one minus this variable. Each of the two constraints is meant to include in its left-hand side rosters of crew members belonging to one of the two specified categories covering the duty in question. Let  $M_1 \in G$  and  $M_2 \in G$  denote the corresponding crew categories,  $p \in D$  denote the duty for which the restriction applies, and  $z$  be the auxiliary decision variable. Mathematically, this restriction is expressed with the following two constraints:

$$\sum_{i \in I: M_1 \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \leq b_p \cdot z \quad (8)$$

$$\sum_{i \in I: M_2 \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} \leq b_p \cdot (1 - z) \quad (9)$$

The binary nature of  $z$  ensures that the right-hand side of one of these two constraints is equal to 0, while that of the other one is equal to the corresponding duty complement. This way, the number of crew members assigned to this duty is always equal to 0 for one of the two specified categories, with no restriction, whatsoever, imposed on the other one.

### C. Pairing Restriction Type

The final restriction type we consider necessitates the existence of two mutually exclusive but collectively exhaustive crew categories, such that any crew member always belongs to exactly one of the two. An extra prerequisite for the activation of this restriction is that the crew complement of the duty in question is even; when this is the case, the *pairing*-type restriction requires that the number of crew members assigned to this duty for each of the two specified categories is even, too. The aim is to avoid extra costs which arise when the crew members of these two categories are not treated in pairs.

To express a *pairing*-type restriction, we add to the master problem formulation an equality ( $=$ ) constraint whose right-hand side is equal to an integer auxiliary variable multiplied by 2. The left-hand side of this constraint is meant to include the rosters of crew members belonging to one of the two specified categories which cover the duty in question. This way,

the sum of these crew members is ensured to be an even number (0, 2, 4, etc.). Let  $M \in G$  denote any of the two corresponding crew categories,  $p \in D$  denote the duty to which the restriction applies, and  $w$  be the auxiliary decision variable. Mathematically, the *pairing*-type restriction is expressed with the following constraint:

$$\sum_{i \in I: M \in G_i} \sum_{j \in R_i} a_{ijp} \cdot x_{ij} = 2 \cdot w \quad (10)$$

#### D. Algorithmic Modifications

After incorporating the special restrictions as illustrated above, some proper algorithmic modifications are deemed necessary for the column generation methodology, as explained next. The first one concerns the proper initialization of the master problem formulation. Although the master problem is typically initialized with an empty roster for each crew member, rosters containing a proper subset of duties may alternatively be utilized to this end in some cases ([22], [23]). When this is the case, the special constraints need to be populated with the initial roster variables, based on the associated crew categories and duties covered.

The next algorithmic modification regards the shortest path pricing subproblem. Each time a partial roster is extended through the inclusion of an additional duty, a check of whether the inclusion of this duty necessitates the population of some special constraints with the associated roster variable needs to be performed. In such a case, the dual variables associated with those constraints need to be taken into consideration for the computation of this roster variable's reduced cost. In addition, if this roster is deemed promising upon conclusion of the pricing subproblem and its addition to the master problem is decided, the associated roster variable should be appended to the proper special constraints.

#### V. COMPUTATIONAL RESULTS

In this section, we present computational results from realistic experiments for assessing how the inclusion of the special constraint types affects the computational performance of the column generation solution framework. For the execution of these experiments, we have exploited our long-term research collaboration with AIMS INTL DWC LLC [24], one of the largest worldwide providers of commercial airlines' operations management software.

Since 1983, AIMS has been engaged in the development of a specialized software package which facilitates the support and coordination of numerous airline operations, such as air-crew management, fleet planning, flight scheduling, etc. Among others, the AIMS software includes sophisticated solution algorithms for the treatment of complex decision-making problems encountered in commercial aviation, such as pairing generation, crew rostering, flight and maintenance scheduling, vacation planning, etc.

The proposed model extension incorporating the special restrictions under consideration has been fully implemented within AIMS software. The execution of the computational experiments for which results are presented next was performed on live databases of AIMS' airline clients to which

AIMS holds access for maintenance and support purposes. The actual code includes several specialized techniques for avoiding excessive computational times, such as limits on the size of the branch & price tree, integrality gap tolerances, execution time limits, etc. These techniques are not fixed but have been properly customized depending on the special characteristics of each particular airline client.

The source code interacts with commercial optimization software IBM ILOG CPLEX [25], which performs various specialized operations necessitated for algorithmic execution. All the computational experiments were executed on an Intel (R) Core (TM) i7-1165G7 @ 2.80GHz Processor with 16 GB system memory. Computer processors employed in realistic environments exhibit superior technical characteristics, leading to significantly higher computational performance.

We tested the proposed methodology on 12 realistic crew rostering problem instances, each time with Flight Attendants (FA) as the crew position subject to assignment. Table 1 presents the data parameters of these instances, i.e., the problem index ( $i$ ), the number of crew members ( $N$ ), the number of duties (standbys + pairings) subject to assignment ( $D$ ) and the total complement ( $TC$ ), computed as the weighted sum of the duties, with weights the corresponding crew complements.

TABLE I.  
DATA PARAMETERS OF THE PROBLEM INSTANCES

| $i$ | $N$  | $D$ | $TC$ |
|-----|------|-----|------|
| 1   | 1061 | 787 | 3675 |
| 2   | 1089 | 801 | 3918 |
| 3   | 1104 | 832 | 4194 |
| 4   | 1073 | 799 | 3876 |
| 5   | 1128 | 867 | 4233 |
| 6   | 1102 | 809 | 4051 |
| 7   | 1239 | 984 | 5332 |
| 8   | 1255 | 991 | 5498 |
| 9   | 1247 | 965 | 5154 |
| 10  | 1118 | 863 | 4349 |
| 11  | 1094 | 796 | 4042 |
| 12  | 1139 | 880 | 4426 |

The next 5 tables present the results of these experiments. Specifically, Table 2 presents the results when no special restrictions are present, Table 3 only with *min/max*-restrictions activated, Table 4 only with *with*-restrictions activated, Table 5 only with *without*-restrictions activated, and finally, Table 6 only with *pairing*-restrictions activated. Note that the same 12 problem instances were tested in all 5 model variations.

All these tables, besides the first one, include a column labeled "# SR", which presents the actual number of activated special restrictions. The next 3 columns present the obtained results, i.e., the time required for the solution of the master LP relaxation at the root node of the branch & price tree (*LP Time*), the time required for obtaining the final integer solution (*IP Time*) and the number of nodes of the branch & price tree (*Nodes*). The computational time results are presented in the form HH:MM:SS, where HH shows the number of hours, MM the number of minutes and SS the number of seconds. Note that the *IP Time* results are not reported in isolation, i.e., they include the associated *LP Time* counterparts.

The computational times reported in the following tables pertain exclusively to the column generation algorithmic methodology. The actual computational times are practically 10-15% higher, since they also include processing time for database access, parameter information loading, as well as several pre-processing procedures which need to be executed before the column generation solution framework is invoked.

TABLE II.  
COMPUTATIONAL RESULTS - NO SPECIAL RESTRICTIONS

| <i>i</i> | <i>LpTime</i> | <i>IpTime</i> | <i>Nodes</i> |
|----------|---------------|---------------|--------------|
| 1        | 10:25:48      | 20:36:12      | 1821         |
| 2        | 11:08:03      | 21:49:57      | 1833         |
| 3        | 11:27:13      | 22:47:36      | 1871         |
| 4        | 11:03:36      | 22:55:12      | 1859         |
| 5        | 12:50:48      | 25:43:48      | 1921         |
| 6        | 11:52:31      | 24:40:49      | 1863         |
| 7        | 15:26:12      | 31:00:48      | 1997         |
| 8        | 16:06:29      | 33:00:36      | 2015         |
| 9        | 15:54:19      | 32:52:31      | 2005         |
| 10       | 15:12:30      | 31:23:44      | 1919         |
| 11       | 12:45:09      | 25:50:24      | 1905         |
| 12       | 16:03:14      | 32:03:18      | 1957         |

TABLE III.  
COMPUTATIONAL RESULTS - ONLY *MIN/MAX* SPECIAL RESTRICTIONS

| <i>i</i> | # <i>SR</i> | <i>LpTime</i> | <i>IpTime</i> | <i>Nodes</i> |
|----------|-------------|---------------|---------------|--------------|
| 1        | 57          | 10:38:17      | 20:49:36      | 1835         |
| 2        | 65          | 11:21:42      | 22:45:25      | 1867         |
| 3        | 83          | 11:39:28      | 23:28:11      | 1893         |
| 4        | 78          | 11:21:03      | 23:38:26      | 1861         |
| 5        | 69          | 13:15:33      | 26:40:49      | 1939         |
| 6        | 81          | 12:00:15      | 25:29:18      | 1893         |
| 7        | 70          | 16:07:01      | 32:34:18      | 2009         |
| 8        | 97          | 16:23:08      | 33:43:23      | 2027         |
| 9        | 95          | 16:16:12      | 33:46:44      | 2045         |
| 10       | 80          | 15:44:48      | 32:11:27      | 1945         |
| 11       | 76          | 13:18:12      | 26:51:07      | 1925         |
| 12       | 82          | 16:38:17      | 33:32:04      | 1985         |

TABLE IV.  
COMPUTATIONAL RESULTS - ONLY *WITH* SPECIAL RESTRICTIONS

| <i>i</i> | # <i>SR</i> | <i>LpTime</i> | <i>IpTime</i> | <i>Nodes</i> |
|----------|-------------|---------------|---------------|--------------|
| 1        | 73          | 10:53:09      | 21:07:16      | 1849         |
| 2        | 88          | 11:39:33      | 23:12:05      | 1859         |
| 3        | 64          | 11:54:51      | 23:18:34      | 1881         |
| 4        | 71          | 11:31:10      | 24:22:56      | 1877         |
| 5        | 69          | 13:04:22      | 26:31:33      | 1919         |
| 6        | 70          | 12:14:18      | 25:38:04      | 1897         |
| 7        | 63          | 15:46:29      | 32:53:10      | 2019         |
| 8        | 82          | 16:49:13      | 33:04:51      | 2015         |
| 9        | 57          | 16:31:47      | 34:37:23      | 2079         |
| 10       | 62          | 15:41:56      | 32:18:19      | 1927         |
| 11       | 70          | 13:23:04      | 27:40:53      | 1937         |
| 12       | 75          | 16:31:39      | 34:14:41      | 1997         |

TABLE V.  
COMPUTATIONAL RESULTS - ONLY *WITHOUT* SPECIAL RESTRICTIONS

| <i>i</i> | # <i>SR</i> | <i>LpTime</i>   | <i>IpTime</i>   | <i>Nodes</i> |
|----------|-------------|-----------------|-----------------|--------------|
| 1        | 26          | 10:41:11        | 21:28:32        | 1859         |
| 2        | <b>35</b>   | <b>11:49:08</b> | <b>25:31:27</b> | <b>2011</b>  |
| 3        | 19          | 11:35:16        | 23:50:44        | 1897         |
| 4        | <b>38</b>   | <b>11:59:23</b> | <b>26:07:16</b> | <b>1977</b>  |
| 5        | 21          | 13:11:28        | 26:33:55        | 1943         |
| 6        | 12          | 12:01:30        | 25:36:47        | 1881         |
| 7        | 15          | 15:51:39        | 32:09:36        | 2011         |

|           |           |                 |                 |             |
|-----------|-----------|-----------------|-----------------|-------------|
| 8         | 26        | 16:25:27        | 35:11:58        | 2053        |
| 9         | 18        | 16:18:40        | 33:52:09        | 2053        |
| <b>10</b> | <b>33</b> | <b>15:27:12</b> | <b>34:56:19</b> | <b>2033</b> |
| <b>11</b> | <b>24</b> | <b>13:14:22</b> | <b>28:01:33</b> | <b>1989</b> |
| 12        | 17        | 16:24:04        | 33:17:29        | 1993        |

TABLE VI.  
COMPUTATIONAL RESULTS - ONLY *PAIRING* SPECIAL RESTRICTIONS

| <i>i</i>  | # <i>SR</i> | <i>LpTime</i>   | <i>IpTime</i>   | <i>Nodes</i> |
|-----------|-------------|-----------------|-----------------|--------------|
| 1         | 12          | 10:49:41        | 21:32:37        | 1853         |
| 2         | 13          | 11:37:52        | 22:59:20        | 1871         |
| 3         | 10          | 11:47:02        | 23:37:52        | 1881         |
| 4         | 13          | 11:39:14        | 23:11:33        | 1871         |
| 5         | 12          | 13:01:29        | 26:30:53        | 1963         |
| 6         | 13          | 12:08:29        | 25:51:37        | 1891         |
| 7         | <b>12</b>   | <b>15:45:18</b> | <b>46:39:33</b> | <b>5305</b>  |
| 8         | 13          | 16:34:11        | 33:51:21        | 2045         |
| <b>9</b>  | <b>10</b>   | <b>16:24:03</b> | <b>37:17:44</b> | <b>2193</b>  |
| <b>10</b> | <b>10</b>   | <b>16:09:57</b> | <b>41:45:56</b> | <b>3627</b>  |
| 11        | 12          | 13:06:09        | 27:57:29        | 1957         |
| 12        | 13          | 16:23:14        | 34:43:05        | 2009         |

As the above results demonstrate, the *min/max*, and the *with* restriction types do not appear to increase the computational times considerably; the *without* and the *pairing* restriction types, on the other hand, result in substantially higher computational times for some problem instances (highlighted in bold). An interesting relevant observation regards the fact that the size of the branch & price tree also appears to be significantly larger in those problem instances, as depicted by the corresponding number of tree nodes.

An important observation that provides valuable insight into this behavior is the fact that the imposition of the *without* and the *pairing* restriction types necessitates auxiliary decision variables, whereas that of the *min/max* and the *with* restriction types does not. Combined with the fact that a solution in which all roster variables satisfy the integrality constraints may still fail to be feasible simply because the associated auxiliary decision variables do not, this observation helps justify why some computational times are significantly higher.

To provide a specific example for the *without* restriction type, consider the case of a duty with total complement equal to 4. A solution in which all roster variables are integer may assign to this duty two crew members of the first of the mutually exclusive crew categories and two of the second one, with the corresponding auxiliary variable being equal to 0.5, and thus failing to satisfy integrality.

Similarly, a solution in which all roster variables are integer may assign to a duty with an active pairing restriction three crew members of the specified crew category, with the corresponding auxiliary variable being equal to 1.5, and thus failing to satisfy integrality.

Of course, the algorithmic methodology can always backtrack and reverse previously made poor branching decisions that lead to such pitfalls, but this is both time consuming and tedious. The branching strategy (choice of the next decision variable on which branching is performed) is a vital element of the algorithmic methodology that can crucially affect computational performance; consequently, several alternative branching schemes have been proposed in the literature [26].

Based on proper experimentation, in our algorithmic implementation we adopt the strategy of branching on the largest fractional variable. For roster variables, the branch on 1 decision assigns a specific roster to the corresponding crew member, while the branch to 0 decision prohibits this assignment. For a *without* auxiliary decision variable, each of the two branching values determines which of the two mutually exclusive crew categories can be assigned to the duty in question. Finally, for a pairing auxiliary decision variable, each of the two branches imposes an even number as a lower/upper bound on the number of crew members of the associated crew category assigned to this duty.

To circumvent the computational effort challenges that can potentially arise in the presence of *without* and/or *pairing* restrictions, we adopt a slightly different branching strategy. Specifically, we give higher branching priority to the auxiliary decision variables associated with the special restrictions, and we do not choose to branch on a roster variable unless all of them are integer valued.

This is a very crucial algorithmic amendment that improves computational performance significantly, as justified by additional computational results presented next. The rationale motivating its adoption is that it avoids feeding the algorithmic implementation with misleading information that would significantly deteriorate its performance.

More specifically, in the case of a *without* restriction type, it clarifies in advance which of the two mutually exclusive crew categories will be freely assigned to the duty in question and which will be prohibited from assignment, thus assisting the algorithm in deciding earlier which particular roster variables will be used to this end. Similarly, in the case of a *pairing* restriction type, it excludes in advance the assignment of the intermediate odd number of crew members to the duty in question, which would otherwise appear as feasible if the auxiliary variable assumed a fractional value.

Failing to adopt this strategy in case of a *without* restriction type often leads the algorithm to assume that both mutually exclusive categories can be assigned to the duty in question simultaneously, which results in a large number of roster assignment decisions that need to be reversed in order to reach a feasible solution. Similarly, failing to adopt it in case of a *pairing* restriction often leads the algorithm to assume that an odd number of the specified category crew members can be assigned to the duty, which, in a similar way, necessitates the reversion of a large number of roster assignment decisions in order to reach a feasible solution.

The next two tables present computational results for the above problem instances with the same *without* and *pairing* restriction types activated, the only difference being that the strategy of not branching on a roster variable unless all *without/pairing* auxiliary decision variables are integer-valued is adopted (we use the term *without/pairing priority branching* for this strategy). In the presence of many fractionally valued auxiliary decision variables, the algorithm chooses to branch on the one with the largest value.

TABLE VII.

 COMPUTATIONAL RESULTS - ONLY *WITHOUT* SPECIAL RESTRICTIONS - *WITHOUT* PRIORITY BRANCHING

| <i>i</i> | <i>LpTime</i> | <i>IpTime</i> | <i>Nodes</i> |
|----------|---------------|---------------|--------------|
| 1        | 10:37:11      | 21:21:17      | 1841         |
| 2        | 11:44:08      | 23:53:52      | 1893         |
| 3        | 11:38:27      | 24:01:16      | 1885         |
| 4        | 12:01:37      | 24:41:28      | 1879         |
| 5        | 13:14:52      | 26:24:08      | 1951         |
| 6        | 12:00:19      | 25:21:12      | 1869         |
| 7        | 15:54:42      | 32:13:29      | 2017         |
| 8        | 16:24:13      | 33:51:05      | 2029         |
| 9        | 16:16:04      | 33:08:59      | 2031         |
| 10       | 15:24:43      | 33:17:33      | 1951         |
| 11       | 13:07:14      | 27:08:27      | 1947         |
| 12       | 16:18:40      | 32:37:11      | 1977         |

TABLE VIII.

 COMPUTATIONAL RESULTS - ONLY *PAIRING* SPECIAL RESTRICTIONS - *PAIRING* PRIORITY BRANCHING

| <i>i</i> | <i>LpTime</i> | <i>IpTime</i> | <i>Nodes</i> |
|----------|---------------|---------------|--------------|
| 1        | 10:45:04      | 21:17:02      | 1837         |
| 2        | 11:35:52      | 22:26:11      | 1841         |
| 3        | 11:43:03      | 23:39:28      | 1875         |
| 4        | 11:41:35      | 23:23:57      | 1885         |
| 5        | 13:00:37      | 26:16:27      | 1961         |
| 6        | 12:05:43      | 24:58:50      | 1869         |
| 7        | 15:48:33      | 32:17:29      | 2005         |
| 8        | 16:37:13      | 33:58:40      | 2037         |
| 9        | 16:27:10      | 34:03:15      | 2041         |
| 10       | 16:17:26      | 35:06:22      | 1989         |
| 11       | 12:59:26      | 26:44:18      | 1937         |
| 12       | 16:25:57      | 34:31:36      | 2003         |

The above results demonstrate that the computational challenges posed by the *without/pairing* restriction types are handled quite satisfactorily by the *without/pairing priority branching* strategy. Its adoption avoids the excessively large branch & price trees, and in turn the dramatically increased computational times encountered sometimes otherwise. For this reason, it has been adopted as the standard branching strategy in the actual algorithmic implementation.

Note that the computational time required for the solution of the master LP relaxation at the root of the branch & price tree is independent of the adopted branching strategy. Consequently, the corresponding *Lp Time* values should coincide in Tables 5 and 7, as well as in Tables 6 and 8. The small differences observed are negligible and can be explained by the inevitable variance in the computer processor's task workload.

## VI. SUMMARY AND FUTURE RESEARCH

In this research work, we have addressed the problem of incorporating special restrictions into column generation algorithms for aircrew rostering. We considered three types of restrictions typically encountered in commercial aviation, and we illustrated how the algorithmic methodology needs to be modified for accommodating them.

Computational results from realistic problem instances indicate that, in some special cases, the inclusion of some specific restriction types may lead to substantially increased branch & price trees and, consequently, computational times.

Mainly, this behavior appears to be present because these restriction types exploit auxiliary decision variables which may provide misleading information to the optimizer if their exact integer values have not been determined promptly.

To properly handle these difficulties, we devised a specialized branching strategy which gives higher branching priority to the auxiliary decision variables. Additional computational experiments confirm that its adoption addresses successfully the difficulties introduced by the increased computational times; consequently, it has been adopted as standard in the actual algorithmic implementation.

The present research can be extended in many fruitful ways. An interesting direction for future research involves the development of additional special techniques for handling the computational challenges potentially arising by the incorporation of special restrictions that involve auxiliary decision variables. An innovative idea in that direction is to determine the optimal integer values of these decision variables and temporarily enforce them to the master LP relaxation for identifying its optimal dual solution utilized for the solution of the shortest path pricing subproblem.

Another interesting direction for future research is the investigation of how the incorporation of the special restrictions affects the quality of the final solutions. As the present research is currently in progress, we solely investigate the corresponding effect on computational performance; such a future research extension will clearly render it more complete.

#### ACKNOWLEDGMENT

The research is conducted in the operating framework of the University of Thessaly Innovation, Technology Transfer Unit and Entrepreneurship Center “One Planet Thessaly”, under the “University of Thessaly Grants for Scientific Publication Support” action and is funded by the Special Account of Research Grants of the University of Thessaly. The authors have not used any AI tool, whatsoever, during the preparation of this work and take full responsibility for its content.

#### REFERENCES

- [1] M. Fuentes, L. Cadarso, V. Vaze, and C. Barnhart, “The tail assignment problem: A case study at Vueling Airlines,” *Transp. Res. Proc.*, 52, pp. 445-452, 2021. <https://doi.org/10.1016/j.trpro.2021.01.052>
- [2] G. Kozanidis, “Optimal assignment of aircrew trainees to simulator and classroom training sessions subject to seniority and preference restrictions,” *J. Air Transp. Manag.*, 59, pp. 143-154, 2017. <https://doi.org/10.1016/j.jairtraman.2016.11.012>
- [3] H. Dawid, J. König, and C. Strauss, “An enhanced rostering model for airline crews,” *Comput. Oper. Res.*, 28(7), pp. 671-688, 2001. [https://doi.org/10.1016/S0305-0548\(00\)00002-2](https://doi.org/10.1016/S0305-0548(00)00002-2)
- [4] B. Gopalakrishnan, and E. L. Johnson, “Airline crew scheduling: State-of-the-art,” *Ann. Oper. Res.*, 140, pp. 305-337, 2005. <https://doi.org/10.1007/s10479-005-3975-3>
- [5] A. Aydemir-Karadag, B. Dengiz, and A. Bolat, “Crew pairing optimization based on hybrid approaches,” *Comput. Ind. Eng.*, 65(1), pp. 87-96, 2013. <https://doi.org/10.1016/j.cie.2011.12.005>
- [6] P. R. Day, and D. M. Ryan, “Flight attendant rostering for short-haul airline operations,” *Oper. Res.*, 45(5), pp. 649-661, 1997. <https://doi.org/10.1287/opre.45.5.649>
- [7] M. Gamache, F. Soumis, D. Villeneuve, J. Desrosiers, and E. Gélinas, “The preferential bidding system at Air Canada,” *Transp. Sci.*, 32(3), pp. 246-255, 1998. <https://doi.org/10.1287/trsc.32.3.246>

- [8] A. I. Z. Jarrah, and J. T. Diamond, “The problem of generating crew bidlines,” *Interfac.*, 27(4), pp. 49-64, 1997. <https://doi.org/10.1287/inte.27.4.49>
- [9] J. Desrosiers, and M. E. Lübbecke, “A primer in column generation,” in *Column Generation*, G. Desaulniers, J. Desrosiers, and M.M. Solomon, Eds. Boston, MA, US: Springer, pp. 1-32, 2005. [https://doi.org/10.1007/0-387-25486-2\\_1](https://doi.org/10.1007/0-387-25486-2_1)
- [10] D. Ryan, “The solution of massive generalized set partitioning problems in air crew rostering,” *J. Oper. Res. Soc.*, 43, pp. 459-467, 1992. <https://doi.org/10.1057/jors.1992.72>
- [11] M. Gamache, F. Soumis, G. Marquis, and J. Desrosiers, “A column generation approach for large-scale aircrew rostering problems,” *Oper. Res.*, 47(2), pp. 247-263, 1999. <https://doi.org/10.1287/opre.47.2.247>
- [12] T. Fahle, U. Junker, S. E. Karisch, N. Kohl, M. Sellmann, and B. Vaaben, “Constraint programming based column generation for crew assignment,” *J. Heuristics*, 8(1), pp. 59-81, 2002. <https://doi.org/10.1023/A:1013613701606>
- [13] J. Desrosiers, and M. E. Lübbecke, “Branch-price-and-cut algorithms,” in *Encyclopedia of Operations Research and Management Science*. Chichester, UK: John Wiley & Sons, pp. 109-131, 2011. <https://doi.org/10.1002/9780470400531.eorms0118>
- [14] G. Kozanidis, “Branch and price for covering shipments in a logistic distribution network with a fleet of aircraft,” *Optim. Methods Softw.*, 33(2), pp. 221-248, 2018. <https://doi.org/10.1080/10556788.2017.1281923>
- [15] G. Kozanidis, “Column generation for optimal shipment delivery in a logistic distribution network,” in *Sustainable Logistics and Transportation*, D. Cinar, K. Gakis, and P.M. Pardalos, Eds. Springer Optimization and Its Applications, vol. 129, pp. 87-112. Berlin: Springer-Verlag, 2018. [https://doi.org/10.1007/978-3-319-69215-9\\_5](https://doi.org/10.1007/978-3-319-69215-9_5)
- [16] Z. Liu, and Q. Xiang, “A branch-and-price algorithm for the airport gate assignment problem considering the trade-off between robustness and efficiency,” *Transp. Res. C Emerg. Technol.*, 154, 104232, 2023. <https://doi.org/10.1016/j.trc.2023.104232>
- [17] G. Kozanidis, “An integrated column generation solution framework for optimal aircrew vacation planning subject to seniority ranking and priority preference satisfaction,” *Omega*, 135, 103224, 2025. <https://doi.org/10.1016/j.omega.2025.103324>
- [18] J. Desrosiers, F. Soumis, and M. Desrochers, “Routing with time windows by column generation,” *Networks*, 14(4), pp. 545-565, 1984. <https://doi.org/10.1002/net.3230140406>
- [19] C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. Savelsbergh, and P. H. Vance, “Branch-and-price: Column generation for solving huge integer programs,” *Oper. Res.*, 46(3), pp. 316-329, 1998. <https://doi.org/10.1287/opre.46.3.316>
- [20] M. Gamache, and F. Soumis, “A method for optimally solving the rostering problem” in *Operations Research in the Airline Industry*, G. Yu, Ed. Boston, MA, US: Springer, pp. 124-157, 1998. [https://doi.org/10.1007/978-1-4615-5501-8\\_5](https://doi.org/10.1007/978-1-4615-5501-8_5)
- [21] A. Kasirzadeh, M. Saddoune, and F. Soumis, “Airline crew scheduling: Models, algorithms, and data sets,” *EURO J. Transp. Logist.*, 6(2), pp. 111-137, 2017. <https://doi.org/10.1007/s13676-015-0080-x>
- [22] G. Kozanidis, and O. Moschopoulos, “Set-cover master problem formulations for maximum flight coverage in branch & price solution algorithms for optimal aircrew rostering,” *Oper. Res. Int. J.*, 25(3), 69, 2025. <https://doi.org/10.1007/s12351-025-00950-0>
- [23] G. Kozanidis, and O. Moschopoulos, “Set-cover master problem formulations in branch & price solution methodologies for optimal aircrew scheduling,” in *Lecture Notes in Operations Research, Annual International Conference of the German Operations Research Society*, Hamburg, Germany, August 29 - September 1, pp. 249-256, 2023. [https://doi.org/10.1007/978-3-031-58405-3\\_32](https://doi.org/10.1007/978-3-031-58405-3_32)
- [24] AIMS INTL DWC-LLC (2025). Airline solutions package and aviation software. <https://www.aimsairlinesoftware.com/> Last accessed on April 13, 2026.
- [25] IBM ILOG CPLEX Optimization Studio v. 22.1.1 (2022). <https://www.ibm.com/docs/en/icos/22.1.1> . Last accessed on April 13, 2026.
- [26] F. Vanderbeck, “On Dantzig-Wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm,” *Oper. Res.*, 48(1), pp. 111-128, 2000. <https://doi.org/10.1287/opre.48.1.111.12453>

## Software System for Optimization of Metallic Nanostructures

Rossen Mikhov  
0000-0001-7186-0110  
Institute of Information and  
Communication Technologies,  
Bulgarian Academy of Sciences  
Acad. G. Bonchev Str., bl. 2,  
1113 Sofia, Bulgaria

Vladimir Myasnichenko  
0000-0003-4355-7055  
Tver State University  
33, Zhelyabova Str.,  
170100 Tver, Russia  
Email: viplabs@ya.ru

Leoneed Kirilov<sup>1,2</sup>  
0000-0001-8424-9070  
<sup>1</sup> Institute of Information and  
Communication Technologies,  
Bulgarian Academy of Sciences  
Acad. G. Bonchev Str., bl. 2,  
1113 Sofia, Bulgaria  
Email: l\_kirilov\_8@abv.bg  
<sup>2</sup> Centre of Excellence in  
Informatics and Information and  
Communication Technologies,  
Sofia, Bulgaria

Nickolay Sdobnyakov  
0000-0003-1749-0336  
Tver State University  
33, Zhelyabova Str.,  
170100 Tver, Russia

Stefka Fidanova<sup>1,2</sup>  
0000-0002-8484-5849  
<sup>1</sup> Institute of Information and Communication  
Technologies,  
Bulgarian Academy of Sciences  
Acad. G. Bonchev Str., bl. 2,  
1113 Sofia, Bulgaria  
<sup>2</sup> Centre of Excellence in Informatics and Information  
and Communication Technologies, Sofia, Bulgaria

**Abstract**—A software system for numerical optimization of metallic nanostructures using Monte Carlo methods is presented in this paper. It is designed for high performance and flexibility, allowing the construction of complex experiment plans combining multiple Monte Carlo algorithms and for systematic variation of their parameters. The system runs on Linux and Windows operating systems, and is compatible with a wide range of existing software packages using the standard XYZ file format for atomic configurations. It has been used for modeling and optimization of various metallic and bimetallic nanostructures from 150 to 16,000 atoms, including nanoparticles, nanowires, nanofilms and nanocages.

**Index Terms**—nanostructures, simulation, software systems, Monte Carlo methods, optimization.

### I. INTRODUCTION

THE problem of optimizing the structure of nanoparticles and nanostructured materials is closely related to their stability, especially when it comes to nanoclusters containing only a few hundred atoms [1]. At the same time, optimization algorithms for various types of nanoobjects have recently been actively tested using the Monte Carlo method

[2]–[5]. This approach is interesting, including from a prognostic point of view for describing the manifestations of instability of nanoscale objects [6]. At the same time, the problem of structural optimization can be complicated by changes in external factors, primarily temperature [7]. Thus, despite some progress in the development of approaches to optimizing the structure of nanoparticles and nanostructured materials, a number of open questions remain, including those related to the optimization algorithm itself.

This paper presents a software system for Monte Carlo optimization of metallic nanostructures. We formulate the primary requirements for such a system, propose a software architecture suitable for meeting those requirements, and describe the components of the system in the way they have been implemented. A description is provided of the computational model on which the implemented Monte Carlo algorithms work, and an illustrative example is given using the software.

### II. REQUIREMENTS FOR THE SOFTWARE SYSTEM

For experimenting with Monte Carlo optimization of various types of metallic and bimetallic nanostructures, the software system should implement a number of Monte Carlo algorithms, providing flexibility in the manner they can be combined, and allowing systematic variation of their parameters. It should be easy to use not only for investigating the

This work was partially supported by the Centre of Excellence in Informatics and ICT under the Grant No BG16RFPR002-1.014-0018-C01, financed by the Research, Innovation and Digitalization for Smart Transformation Programme 2021-2027 and cofinanced by the European Union and by the IICT-BAS project No210386 “Models and methods for decision support and applications” with Virtech LTD, Sofia.

properties of the studied nanoparticles, but also for investigating the behavior of the algorithms themselves.

We formulated the primary requirements for the system as follows:

- It must be optimized for x86-64 processors, achieving performance sufficient to execute billions of Monte Carlo iterations within minutes, thereby enabling the algorithms to reach good solutions for nanoparticles with several hundred to several thousand atoms on standard and widely available hardware.
- It must use the standard XYZ data format for input and output of atomic configurations, ensuring compatibility with a wide range of existing software packages, such as those for 3D visualization, analysis of the results, etc.
- It must be compatible with Linux and Windows operating systems.
- It must support the construction of complex plans for numerical experiments that combine multiple processing steps, various algorithmic variants, parameter variations, output of intermediate data for animated 3D visualization, performance statistics, etc.
- It must be able to synchronize the execution of the experiment plan across multiple processor cores on a single computer or across multiple computers with different operating systems, while remaining robust and capable of continuing normal operation after unexpected interruptions (e.g., power failures or system restarts).
- It must provide a high level of operational flexibility, including the ability to modify the experiment plan after the experiment has already been started, in a way that results already obtained remain usable for the updated plan.

### III. SOFTWARE ARCHITECTURE

The software architecture consists of three main components: Computational core, a template for the Experiment Master Plan, and auxiliary functionalities for analysis of the results, testing, and visualization (Fig. 1).

The Computational core is concentrated in a single executable command written in the C programming language [8]. It incorporates all Monte Carlo algorithms and their variations. Each execution of this command corresponds to a single run of a single algorithm.

The Experiment Master Plan defines what nanoparticles are to be studied, what combinations of algorithms are to be used, whether parameters will be varied, what results are to be collected, where they will be stored, and how they will be aggregated for visualization. Essentially, the Experiment Master Plan consists of one Makefile with a specific structure. However, here we do not just use a standard Make command, but instead we combine some of the lesser-known advanced features of GNU Make [9] with a number of cus-

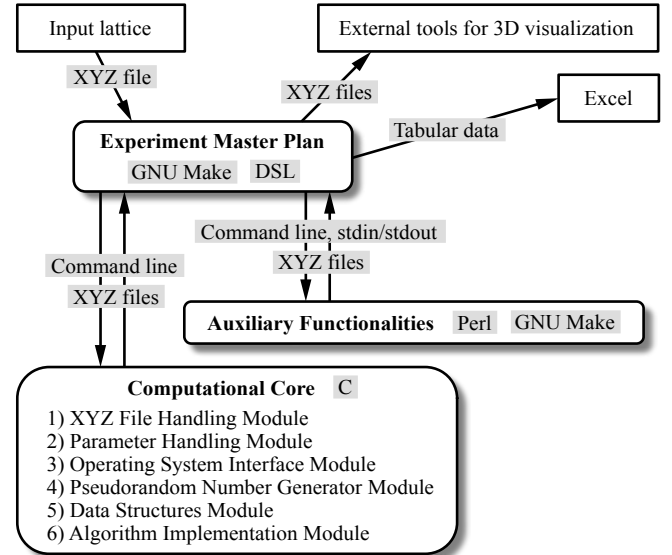


Fig 1. Main components of the software architecture

tom extensions developed specifically for this system. In this way Make is transformed from a cumbersome tool into a flexible and expressive domain-specific language (DSL), suitable for achieving the objectives outlined in Section II.

The auxiliary functionalities consist of a library of GNU Make functions implemented using Perl [10] scripts, which provide the tools necessary for expressing the Experiment Master Plan. Additionally, an automated regression testing system for the implementation of the algorithms of the Computational core is also included.

Graph visualization is done in Excel, using data automatically extracted in real time and formatted according to a table schema defined in the Experiment Master Plan. 3D visualization of atomic configurations is performed separately using standard widely available software packages capable of reading XYZ files.

This architecture isolates the core computational workload into a single command with a clearly defined interface, enabling careful code optimization for performance. At the same time, the experimental workflow of a whole experiment is centralized in a single Makefile, supported by an extensive library of specialized functions, ensuring both flexibility and transparency in experiment design and documentation. This architecture is also easily extensible, allowing for future enhancements as the need arises, such as graphical or web-based user interfaces built upon the existing components.

#### A. Computational Core

The Computational core is implemented as a single command, each execution of which corresponds to a single run of a single algorithm. The core consists of six modules:

##### 1) XYZ File Handling Module

This module provides all necessary functions for reading and writing data in the *de facto* standard XYZ file format for

atomic configurations. The format is structured as follows (Fig. 2):

- The first line contains a single integer  $n$ , which represents the number of nodes in the lattice.
- The second line has no standardized meaning, and is typically ignored when read by other software. For the present system, it contains a long list of variables and their values, used for input parameters, output results (e.g., energy, temperature; everything other than the  $(x, y, z)$  coordinates of the atoms), and miscellaneous internal state.
- Each of the next  $n$  lines represents a node in the lattice and contains the following columns, separated by spaces or tab characters: the chemical element (or the letter  $n$ , if the node is empty),  $x$ ,  $y$ ,  $z$  coordinates. Depending on the experiment, additional columns may be included, which represent the energy of the given atom, classification of the atom by layer group in the nanoparticle, etc.
- The file may contain multiple configurations (e.g., for animation), in which case the above  $n+2$  lines repeat the required number of times.

This module takes special care to store all parameters required to fully reconstruct the system's internal state with sufficient numerical precision, ensuring reproducibility and safe continuation after interruptions.

The module also reads parameters for the potential energy function from a separate file format.

### 2) Parameter Handling Module

This module defines all parameters used by the system, together with the rules for their manipulation. Parameters can be read both from the command line and from the input XYZ file, and are written to output XYZ files. They may be input or output parameters. Each parameter has a designated type (e.g., integer, floating point, string), default value, optionally minimum and maximum value, and help text. Example input parameters include the composition (type and proportion of chemical elements in the particle), maximum number of iterations, neighborhood radius, etc. Output parameters include algorithm stop reason, achieved energy, etc. The starting and current state of the random number generator are also stored as parameters, for reproducibility of the results. The system uses approximately 50 parameters in total.

The parameter system is designed in such a way that any generated XYZ file can serve as input for a subsequent run of the same or a different algorithm, with a simple modification of the command line. This, besides being useful when defining an Experiment Master Plan, also helps for the seamless continuation after intentional or unintentional interruptions.

### 3) Operating System Interface Module

This module contains low-level functions with platform-specific implementations for Linux and Windows.

```

3000
#Algo=Diffusion Task=AuAg-3000_4500~fcc_080M_Tprev1500_T1500_seed25
#Energy=-9583.0102125318499 MinEnergy=-9583.0102125318499
Temper=551.57657207140539 IterCnt=50618260 IterMax=80000000
#StopReason=IterMax JumpMax=80000000 JumpCnt=50618260
PbbyJumpCnt=25311619 EqlbrIterMax=80000000 EqlbrIterCnt=0
TemperChangelterMax=10000 TemperChangelterCnt=8260
SnapshotIterMax=80000000 SnapshotIterCnt=50618260 #Nodes=3004
Num=3000 Composition=AuAg Rneigh=5.26 #Rcut=6.7823411887046792
TemperDelta=-1.8739842480310037e-05 #ZoomX=1.002 #ZoomY=1
#ZoomZ=1 ZoomMin=0.8 ZoomMax=1.2 SaveNodes=0
RandState=2D43C0D10795E7100E4D586B2CFAB6D7
#InitialRandState=5250B2EA33F5F116E134B987F06DA224
Au -5.8165665933600001 -1.9214831400000001 -19.320616619999999
Au -9.8402655085199981 -17.709002340000001 11.800856399999999
Au 1.8900739026000002 -0.0004763400000000001 -17.068808520000001
Au -1.9485816044400002 21.585799979999997 7.8422465399999997
Au 7.9086129746400005 19.785529560000001 -11.86733892
[...]
Ag 3.7796173664399997 -11.328081239999999 -11.32314648
Ag 1.9835312842800001 13.8258042 19.75329348
Ag 3.99379606884 -23.70382386 -3.9047772599999999
Ag -7.6161807576000005 -11.41884288 11.415396300000001
Ag -9.7092440686800003 -15.49355826 -9.6807108599999996
[...]
    
```

Fig 2. Example XYZ file. “⋮” means continuation of the previous line

### 4) Pseudorandom Number Generator (PRNG) Module

As Monte Carlo simulations use an enormous number of pseudorandom numbers, a suitable choice of PRNG is essential. Default PRNGs typically have a number of disadvantages, so we specifically chose the xoroshiro128+ algorithm [11], which ensures high statistical quality for floating-point random number generation. The algorithm has a long period  $2^{128}-1$ , compact internal state (only 32 hex digits, which is convenient for use in XYZ files and on the command line), and has code that is especially optimized for x86/x86-64 processors.

### 5) Data Structures Module.

Defines the data structures and associated operations required for the implementation of the algorithms.

### 6) Algorithm Implementation Module

Implements the computational model and algorithms described in Section IV and their variants, with a strong emphasis on performance for x86-64 processors. To this end, the code is organized in a way that enables the compiler to analyze it as much as possible in order to apply built-in optimizations. The most critical code sections are implemented in multiple ways, then profiled on representative examples in real conditions, to select the most efficient coding approach.

## B. Experiment Master Plan

Each experimental study typically requires hundreds of algorithm runs under varying conditions. In the proposed architecture, the entire experiment is defined in a single Makefile.

The XYZ files are stored in a small number of directories (often one) on a shared filesystem. All the information about what the XYZ represents in the context of the Experiment Master Plan is encoded in the file name. The file name is parsed by the system into a list of variables that can be used to construct command-line parameters for the Computational core, in case the file will subsequently be used as input to some algorithm.

To facilitate expressing complex experiment plans in a concise and declarative manner, the variables and parameters can be combined using a custom domain-specific language (DSL), which supports arithmetic operators and other functions, type checking, regular expressions, and other features. This approach enables a high degree of flexibility to plan studies of different kinds of nanoparticles, or studies of the behavior of the algorithms themselves.

The developed DSL also functions as a query language over the results, enabling the extraction of specific information from the potentially thousands of generated XYZ files, and aggregation of data in tabular form for real-time visualization, according to a user-defined table schema.

The use of static descriptive file names in one shared directory, and GNU Make for generating those files, brings all classic advantages of GNU Make, such as the possibility of synchronized execution using several processor cores on one or several computers, possibly with different operating systems. The system has features for robustness that allow it to continue normal operation after unexpected interruptions, and features for operational flexibility that allow modifying the Master Plan after the experiment has already been started without loss of work. The developed DSL and other extensions compensate for many of the disadvantages of the Makefile language, specifically for use in experiments with nanoparticles.

#### C. Auxiliary Functionalities for Analysis and Testing

The auxiliary functionalities consist of a library of GNU Make functions, implemented using Perl scripts. They support the DSL described above, and provide additional tools for experiment management, including:

- various commands for XYZ file processing;
- progress visualization with color;
- logging;
- performance statistics.

An automated regression testing system for the Computational core is also included.

A graphical user interface is not provided, because our past usage of the system has always been very dynamic, with each experiment having a unique element in its design that evolves even during the process of its execution. It would be difficult to warp such usage in some standardized collection of buttons and menus. Instead, our system supports convenient batch execution of trials. Should the system ever become popular among non-technical users, the architecture allows for the addition of graphical or web-based interfaces on top of the existing components.

#### IV. COMPUTATIONAL MODEL AND ALGORITHMS

Two basic Monte Carlo algorithms, developed by the authors, are currently provided:

- a wide-lattice simulated annealing algorithm [12];
- an algorithm for the simulation of diffusion [13].

The system works with a pre-prepared model of the nanostructure. For example, in the case of a bimetallic

nanoalloy, the total number of atoms is known, as well as the number of atoms of each metal separately (proportion). The atoms are arranged on a predefined lattice consisting of empty nodes. The first goal is to find a configuration of atoms on this lattice so that the resulting nanoparticle is stable—i.e., to minimize the potential energy of the system.

The Cleri–Rosato many-body tight-binding (TB) potential [14] is used to calculate the interaction between the atoms. The total potential energy for the entire system has the following form:

$$E = \sum_i \sum_{j \neq i} A_{ab} \exp(-p_{ab}(\frac{r_{ij}}{r_{0,ab}} - 1)) - \sum_i \sqrt{\sum_{j \neq i} \xi_{ab}^2 \exp(-2q_{ab}(\frac{r_{ij}}{r_{0,ab}} - 1))}, \quad (1)$$

where  $i$  enumerates all atoms in the system;  $j$  enumerates all atoms other than  $i$ , but at a distance from atom  $i$  less than the predefined global parameter  $R_{cut}$ ;  $a$  and  $b$  denote the chemical element of atoms  $i$  and  $j$ ;  $r_{ij}$  is the distance between the atoms;  $r_{0,ab}$ ,  $A_{ab}$ ,  $p_{ab}$ ,  $\xi_{ab}$ , and  $q_{ab}$  are constants specific to the chemical elements under consideration.

The modeled nanoalloy can have nanoscale dimensions along some of the coordinate axes, and macroscale dimensions along the remaining axes. A distinction is made between 3 cases:

- nanoparticle, where all three axes are nanoscale;
- nanowire, where one of the axes is macroscale;
- nanofilm, where two of the axes are macroscale.

In case of a nanowire or a nanofilm, the model is analogous to that for a nanoparticle, but using periodic boundary conditions. This means that the used lattice is nanoscale along all three axes, but a periodic window is introduced along the axes where macroscale dimensions are modeled. The distance between two atoms is thus defined as follows:

$$r_{ij} = \sqrt{|\Delta x_{ij}|^2 + |\Delta y_{ij}|^2 + |\Delta z_{ij}|^2} \quad (2)$$

$$|\Delta x_{ij}| = \min(|x_i - x_j|, L_x - |x_i - x_j|), \quad (3)$$

where  $x_i$  and  $x_j$  are the  $x$ -coordinates of atoms  $i$  and  $j$  inside the periodic cell, and  $L_x$  is the  $x$ -size of the periodic cell. If macroscale dimensions are modeled along the  $y$  and  $z$  axes, then  $|\Delta y_{ij}|$  and  $|\Delta z_{ij}|$  are also calculated in a similar way. The non-periodic case corresponds to  $L_x = \infty$ .

The present model, as well as the algorithms for it, are suitable for the optimization of bimetallic nanoparticles with sizes between 100 and several tens of thousands of atoms.

#### V. ILLUSTRATIVE EXAMPLE

The system has been used for modeling various metallic and bimetallic nanostructures from 150 to 16,000 atoms. In this section, we give an example for using it to optimize six gold–silver nanocages of 3000 atoms each [15].

The Experiment Master Plan for this example is partially illustrated in Fig. 3. It combines both the wide-lattice simu-

lated annealing algorithm and the diffusion algorithm in a hybrid optimization approach consisting of four steps:

*Step 1. Parameter setting for the wide-lattice Monte Carlo algorithm.* The algorithm is run 30 times from random initial configurations for  $40 \times 10^6$  iterations in each run, to determine the optimal initial temperature.

*Step 2. Fixing of the geometric shape.* The wide-lattice algorithm is run again, 30 times from random initial configurations, but this time for  $1000 \times 10^6$  iterations in each run, using the best values from *Step 1* for the initial temperature.

*Step 3. Parameter setting for the diffusion Monte Carlo algorithm.* The algorithm is run 30 times on the results of *Step 2*, for  $8 \times 10^6$  iterations in each run, to determine the optimal initial temperature for diffusion.

*Step 4. Diffusion.* The diffusion algorithm is run again, 30 times on the results of *Step 2*, for  $80 \times 10^6$  iterations in each trial, using the best values from *Step 3* for the initial temperature.

The parameters of the Computational core are set as appropriate for each invocation. The results are queried and aggregated in tabular form for analysis and visualization (the table schema is not shown in the figure). A sample achieved configuration of one of the investigated nanocages is displayed in Fig. 4.

## VI. CONCLUSION

A software system for numerical optimization of metallic nanostructures using Monte Carlo methods has been presented. The architecture consists of a Computational core, designed for high performance, a template for the Experiment Master Plan, designed for flexibility, and auxiliary functionalities for analysis of the results, testing, and visualization. The system is suitable both for investigating the properties of the studied nanostructures, and for investigating the behavior of the algorithms. The system runs on Linux and Windows operating systems, and is compatible with external tools using the standard XYZ file format for atomic configurations. It has been used for modeling and optimization of various metallic and bimetallic nanostructures from 150 to 16,000 atoms, including nanoparticles, nanowires, nanofilms and nanocages.

## REFERENCES

- [1] V. S. Myasnichenko, P. M. Ershov, N. Yu. Sdobnyakov, R. Mikhov, and L. Kirilov, "Thermal stability of copper clusters with the size of 100 – 150 atoms," *Physical and chemical aspects of the study of clusters, nanostructures and nanomaterials*, vol. 9, pp. 330–336, 2017, <https://doi.org/10.26456/pcascnn/2017.9.330>
- [2] U. Gahn, "Atom distributions in face-centered cubic solid solutions from Monte Carlo computer simulations," *Journal of Physics and Chemistry of Solids*, vol. 43, no. 10, pp. 977–989, 1982, [https://doi.org/10.1016/0022-3697\(82\)90114-7](https://doi.org/10.1016/0022-3697(82)90114-7)
- [3] M. C. Giménez and W. Schmicker, "Monte Carlo simulation of nanowires of different metals and two-metal alloys," *The Journal of Chemical Physics*, vol. 134, no. 6, pp. 064707-1–064707-6, 2011, <https://doi.org/10.1063/1.3549900>

```
# All lists undergo Bash-like brace expansion.
S ?= 00..29
STEP1 ?= \
    AuAg--3000-4500_fcc_0040M_T1000,1500,2000,2500}_seed{$S} \
    AuAg--3000-4500_ico_0040M_T2000,2500,3000}_seed{$S} \
    Au3Ag--3000-4500_fcc_0040M_T1000,1500,2000}_seed{$S} \
    Au3Ag--3000-4500_ico_0040M_T2500,3000,4000}_seed{$S} \
    AuAg3--3000-4500_fcc_0040M_T1500,2000,2500}_seed{$S} \
    AuAg3--3000-4500_ico_0040M_T2000,2500,3000}_seed{$S}
STEP2 ?= \
    AuAg--3004-4500_fcc_1000M_T1500_seed{$S} \
    AuAg--3004-4500_ico_1000M_T2500_seed{$S} \
    Au3Ag--3004-4500_fcc_1000M_T1500_seed{$S} \
    Au3Ag--3004-4500_ico_1000M_T3000_seed{$S} \
    AuAg3--3004-4500_fcc_1000M_T2000_seed{$S} \
    AuAg3--3004-4500_ico_1000M_T2500_seed{$S}
STEP3 ?= \
    AuAg--3000-3004_4500-fcc_008M_Tprev1500_T1000,1500,2000}_seed{$S} \
    AuAg--3000-3004_4500-ico_008M_Tprev2500_T1500,2000,2500,3000}
    _seed{$S} \
    Au3Ag--3000-3004_4500-fcc_008M_Tprev1500_T1000,1500,2000}_seed{$S} \
    Au3Ag--3000-3004_4500-ico_008M_Tprev3000_T1000,1500,2000,2500,3000}
    _seed{$S} \
    AuAg3--3000-3004_4500-fcc_008M_Tprev2000_T1000,1500,2000}_seed{$S} \
    AuAg3--3000-3004_4500-ico_008M_Tprev2500_T1500,2000,2500}_seed{$S}
STEP4 ?= \
    AuAg--3000-3004_4500-fcc_080M_Tprev1500_T1500_seed{$S} \
    AuAg--3000-3004_4500-ico_080M_Tprev2500_T2500_seed{$S} \
    Au3Ag--3000-3004_4500-fcc_080M_Tprev1500_T1500_seed{$S} \
    Au3Ag--3000-3004_4500-ico_080M_Tprev3000_T2000_seed{$S} \
    AuAg3--3000-3004_4500-fcc_080M_Tprev2000_T1500_seed{$S} \
    AuAg3--3000-3004_4500-ico_080M_Tprev2500_T2000_seed{$S}
SRO ?= \
    AuAg--3000-3004_4500-fcc_080M_Tprev1500_T1500_seed25.ssh \
    AuAg--3000-3004_4500-ico_080M_Tprev2500_T2500_seed11.ssh \
    Au3Ag--3000-3004_4500-fcc_080M_Tprev1500_T1500_seed04.ssh \
    Au3Ag--3000-3004_4500-ico_080M_Tprev3000_T2000_seed21.ssh \
    AuAg3--3000-3004_4500-fcc_080M_Tprev2000_T1500_seed28.ssh \
    AuAg3--3000-3004_4500-ico_080M_Tprev2500_T2000_seed01.ssh

TASKS ?= $(STEP1) $(STEP2) $(STEP3) $(STEP4) $(SRO)

taskPtn = »Composition:SrtNat?Srt?Nat?«--»Num:Nat«--»nodes:Nat«
    _»ltnodes:Nat?«»ltnSep:Ch??«»ltnShp:Nm«_»it:Dbk«M»tprevSep:Ch*?«
    _»tprev:Nat?«_T»Temper:Dbk«_seed»seed:Nat«ssht:Ch*«
parmsCode = inputClsc=input_--»nodes«_»ltnShp«xyz \
    inputDfsn=trialsAll»Composition«--»nodes«_»ltnodes«_»ltnShp«_1000M_
    _»T»tprev«_seed»seed«xyz \
    InputFile=ltnSep?inputDfsn?inputClsc«_Rneigh»ltnSep?5.26|3.3« \
    IterMax»rnd(it*1e6)«:(it<=4000)«_JumpMax»IterMax« \
    EqIbIterMax»IterMax«_TemperChangeIterMax=10000 \
    TemperDelta»dbl(0-(Temper-1)/(IterMax-1e3))« \
    SnapshotIterMax»ssht?rnd(rnd(0-100)/(TemperDelta*TemperChangeIterMax)) *
    _»TemperChangeIterMax)IterMax« \
    SnapshotIterCnt»ssht?rnd(SnapshotIterMax-TemperChangeIterMax)|0« \
    MinEnergy=INF \
    IterCnt=0 JumpCnt=0 PbtyJumpCnt=0 EqIbIterCnt=0 TemperChangeIterCnt=0 \
    zoom»bool(it>=80)« \
    ZoomMin»zoom?it=320?0.99|0.8|1«_ZoomMax»zoom?it=320?1.01|1.2|1« \
    tgt=$$$$@ nul=$(NulNativ) SnapshotFile»ssht?tgt|nul« \
    MinEnergyFile»ssht?nul|zoom?nul|tgt«_ZoomFile»ssht?nul|tgt«_SaveNodes=0
```

Fig 3. Fragment of the Experiment Master Plan for the given example. The “\” symbol denotes continuation of the previous line.

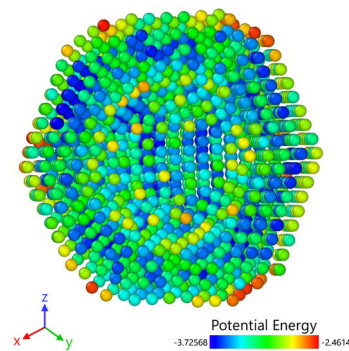


Fig 4. An achieved atomic configuration of the AuAg ico nanocage

- [4] M. Müller and K. Albe, “Kinetic lattice Monte-Carlo simulations on the ordering kinetics of free and supported FePt L10-nanoparticles,” *Beilstein Journal of Nanotechnology*, vol. 2, pp. 40–46, 2011, <https://doi.org/10.3762/bjnano.2.5>
- [5] V. Myasnichenko, N. Sdobnyakov, L. Kirilov, R. Mikhov, and S. Fidanova, “Monte Carlo approach for modeling and optimization of one-dimensional bimetallic nanostructures,” in *Numerical Methods and Applications: NMA 2018*, G. Nikolov, N. Kolkovska, and K. Georgiev, Eds. Springer, 2019, pp. 133–141, [https://doi.org/10.1007/978-3-030-10692-8\\_15](https://doi.org/10.1007/978-3-030-10692-8_15)
- [6] V. Myasnichenko, N. Sdobnyakov, L. Kirilov, R. Mikhov, and S. Fidanova, “Structural instability of gold and bimetallic nanowires using Monte Carlo simulation,” in *Recent Advances in Computational Optimization*, S. Fidanova, Ed. Springer, 2020, pp. 133–145, [https://doi.org/10.1007/978-3-030-22723-4\\_9](https://doi.org/10.1007/978-3-030-22723-4_9)
- [7] R. Mikhov, V. Myasnichenko, S. Fidanova, L. Kirilov, and N. Sdobnyakov, “Influence of the temperature on simulated annealing method for metal nanoparticle structures optimization,” in *Advanced Computing in Industrial Mathematics: BGSIAM 2018*, I. Georgiev, H. Kostadinov, and E. Lilkova, Eds. Springer, 2021, pp. 278–290, [https://doi.org/10.1007/978-3-030-71616-5\\_25](https://doi.org/10.1007/978-3-030-71616-5_25)
- [8] *Information technology — Programming languages — C*, ISO/IEC Standard 9899, ed. 3, 2011.
- [9] *GNU Make Manual*, Free Software Foundation, Feb. 2023, <https://www.gnu.org/software/make/manual/>
- [10] *Perl Documentation*, Perl 5 Porters, Apr. 2026, <https://perldoc.perl.org>
- [11] D. Blackman and S. Vigna, “Scrambled linear pseudorandom number generators,” *ACM Transactions on Mathematical Software*, vol. 47, no. 4, pp. 36:1–36:32, 2021, <https://doi.org/10.1145/3460772>
- [12] V. Myasnichenko, L. Kirilov, R. Mikhov, S. Fidanova, and N. Sdobnyakov, “Simulated annealing method for metal nanoparticle structures optimization,” in *Advanced Computing in Industrial Mathematics: BGSIAM 2017*, K. Georgiev, M. Todorov, and I. Georgiev, Eds. Springer, 2019, pp. 277–289, [https://doi.org/10.1007/978-3-319-97277-0\\_23](https://doi.org/10.1007/978-3-319-97277-0_23)
- [13] R. Mikhov, V. Myasnichenko, L. Kirilov, N. Sdobnyakov, P. Matrenin, et al., “On the problem of bimetallic nanostructures optimization: an extended two-stage Monte Carlo approach,” in *Recent Advances in Computational Optimization: WCO 2020*, S. Fidanova, Ed. Springer, 2022, pp. 235–250, [https://doi.org/10.1007/978-3-030-82397-9\\_12](https://doi.org/10.1007/978-3-030-82397-9_12)
- [14] F. Cleri and V. Rosato, “Tight-binding potentials for transition metals and alloys,” *Physical Review B*, vol. 48, no. 1, pp. 22–33, 1993, <https://doi.org/10.1103/PhysRevB.48.22>
- [15] V. Myasnichenko, R. Mikhov, N. Sdobnyakov, L. Kirilov, and A. Bazulev, “Lattice Monte Carlo simulation of atomic ordering in gold-silver nanocages,” *WSEAS Transactions on Electronics*, vol. 16, pp. 157–167, 2025, <https://doi.org/10.37394/232017.2025.16.16>

# A Reproducible Zero Trust Testbed for Dynamic Trust Evaluation Using Emulation and Virtualization Tools

Tiberius Nkinyili,  
0000-0002-1798-3176,  
Strathmore University,  
Nairobi, Kenya  
Email: ttabulu@strathmore.edu

Vincent Omwenga,  
0000-0002-9443-500X,  
Strathmore University,  
Nairobi, Kenya.  
vomwenga@strathmore.edu,

Solomon Ogara,  
0000-0002-3321-0448,  
Jaramogi Oginga Odinga  
University of Science and  
Technology, Kisumu Kenya.  
sogara@jooust.ac.ke.

**Abstract**—The practical evaluation of Zero Trust Architecture (ZTA) remains constrained by the lack of accessible, reproducible testbeds supporting dynamic trust experimentation. This paper presents a lightweight, containerized Zero Trust testbed integrating a Software Defined Perimeter (SDP), Software Defined Networking (SDN), network emulation, and policy-as-code enforcement. A pluggable Python trust engine implements progressively sophisticated trust models, culminating in Dempster–Shafer evidential fusion with temporal decay and residual inertia. Evaluation across six representative enterprise scenarios demonstrates low policy evaluation latency, modest session setup overhead, linear scalability, and rapid breach containment, while remaining deployable on commodity hardware. By combining reproducibility, open-source components, and advanced trust evaluation within a single platform, the proposed testbed provides an accessible environment for research, education, and the experimental validation of Zero Trust mechanisms.

**Index Terms**—Zero Trust testbed, Software-Defined Perimeter (SDP), Software-Defined Networking (SDN), Mininet, GNS3, Docker, dynamic trust evaluation, virtualization, emulation, containerization, reproducibility.

## I. INTRODUCTION

ZERO Trust Architecture (ZTA) has become the dominant security paradigm for modern enterprise networks due to the erosion of traditional network perimeters, widespread adoption of cloud services, and the proliferation of mobile and IoT devices [1], [2]. Guided by the principle of “never trust, always verify,” ZTA enforces continuous authentication, authorization, and monitoring of every access request [3]. Frameworks such as [4], [5] and [6], and recent cybersecurity initiatives have accelerated its adoption. However, despite the maturation of ZTA frameworks [6], a significant gap remains between theoretical trust models and their practical validation. This gap spans three dimensions: algorithmic opacity, where trust models remain mathematically specified but not reproducible [7]; infrastructure inaccessibility, due to the complexity of deploying realistic ZTA environments [8]; and scenario non-standardization, where poorly specified evaluation conditions hinder reproducibility and cross-study comparison [7], [9].

This paper addresses these challenges through a reproducible, lightweight testbed that maps the NIST architecture onto open-source, containerized components deployable on a commodity workstation. The testbed provides executable

trust engines, Infrastructure-as-Code deployment, and standardized scenarios with published telemetry and empirical datasets. Consistent with the resource-constrained philosophy of prior work [1], it enables reproducible experimentation with dynamic trust models using only commodity hardware.

### A. Contributions

The main contributions of this paper are: (1) a modular, lightweight Zero Trust testbed integrating Software-Defined Perimeter, Software-Defined Networking, and containerized services using open-source components; (2) a pluggable Python-based trust engine supporting progressive policy models, including static weighting and Dempster–Shafer evidence fusion with temporal decay and residual inertia; (3) experimental validation across six enterprise scenarios evaluating latency, breach containment, resource utilization, false-positive rates, and scalability; and (4) a reproducible infrastructure-as-code implementation to support replication and future research.

The remainder of the paper reviews related work (II), presents the system architecture (III) and implementation (IV), describes the experimental methodology (V), results (VI), and discussions (VII), and concludes with limitations (VIII), reproducibility (IX) and future research directions (X).

## II. RELATED WORK

### A. Zero Trust Testbeds and Dynamic Trust Evaluation

Literature contains several efforts to implement Zero Trust principles in testbed environments. Reference [1] demonstrated combining Software Defined Perimeter (SDP) with Software Defined Networking (SDN) and lightweight virtualization for zero trust experimentation with Mininet as a key enabler for resource-constrained prototyping. Reference [10] extended this foundation by building an SDP for network introspection, demonstrating how SDP components map to ZTA concepts in [3]. Their work validated that SDP 2.0 specifications directly support ZTA’s Single Packet Authorization (SPA). Reference [11] investigated identity-aware SDN controllers, while various commercial implementations have demonstrated Zero Trust Network

Access (ZTNA) as alternatives to traditional VPNs [12]. However, these efforts share common limitations: a lack of dynamic trust evaluation, dependence on proprietary components, or failure to provide reproducible artifacts.

Within the CARTA paradigm [13], AI and machine learning support continuous trust assessment through behavioral modelling, adaptive trust scoring, and predictive threat detection [14]. However, these approaches typically depend on labelled training data and often lack interpretability. The proposed engine uses deterministic variance-based weighting for unstable evidence, providing transparent and verifiable trust decisions suitable for adversarial environments. While explainability is prioritized, optimization of the variance parameter through reinforcement learning is identified as future work [15], [16].

### B. Simulation, Emulation, and Virtualization Approaches

The choice between simulation, emulation, and full virtualization has implications on testbed fidelity and resource requirements. Simulators model network behavior mathematically, offering high scalability but limited realism for protocol-specific and timing-sensitive behaviors. Emulators execute real network stacks in virtualized environments, providing a balance between fidelity and resource efficiency. Full virtualization runs unmodified operating systems and applications but incurs higher overhead. Table 1 compares common approaches with their respective strengths and limitations.

**Table 1.** Comparison of testbed approaches for zero trust experimentation

| Approach              | Examples              | Fidelity             | Resource Cost | Isolation | Reproducibility |
|-----------------------|-----------------------|----------------------|---------------|-----------|-----------------|
| Simulation            | ns-3, OMNeT++         | Low (models)         | Very low      | Complete  | High            |
| Lightweight emulation | Mininet               | Medium (Linux netns) | Low           | Partial   | High            |
| Full emulation        | GNS3, Eve-NG          | High (real images)   | Medium-High   | Good      | Medium          |
| Type 2 virtualization | Virtual Box, QEMU/KVM | Very high            | High          | Strong    | Medium          |
| Type 1 virtualization | VMware ESXi, Xen      | Native               | Very high     | Strong    | Low             |

Mininet remains the de facto standard for SDN prototyping and emulation of large-scale networks through lightweight Linux network namespaces. It supports efficient experimentation in resource-constrained environments while achieving aggregated bandwidths exceeding 10 Gb/s with over 1,000 hosts on commodity hardware [17]. Although its lightweight virtualization offers weaker isolation and introduces security challenges compared to Type 1 hypervisors [18], these limitations are generally acceptable for zero trust experimental

tion. Applications requiring hardware-level isolation or proprietary device emulation may instead benefit from platforms such as GNS3 or full virtualization solutions.

### C. Software Stack for Zero Trust Experimentation

Numerous closed and open-source projects have been used to implement parts of ZTA. Table 2 catalogs tools relevant to zero trust testbed construction, extending prior surveys with additional components for dynamic trust evaluation.

**Table 2.** Open-source software components for zero trust testbeds.

| Component               | Purpose                                                               |
|-------------------------|-----------------------------------------------------------------------|
| Keycloak                | Identity Provider (OAuth2, OIDC, SAML)                                |
| Open Policy Agent (OPA) | Policy Decision Point (PDP)                                           |
| Envoy Proxy             | L7 Policy Enforcement Point (PEP)                                     |
| OpenDaylight            | SDN Controller (OpenFlow)                                             |
| Open vSwitch (OVS)      | OpenFlow-enabled virtual switch                                       |
| Mininet                 | Network emulation                                                     |
| GNS3                    | Full network emulation                                                |
| Docker                  | Container platform for applications                                   |
| Pritunl Zero            | Overlay zero trust server                                             |
| OpenZiti                | Zero trust networking overlay                                         |
| SPIFFE/SPIRE            | Identity providers for non-user components such as services and nodes |
| eBPF Cilium/Tetragon    | Kernel level filters for processes and system calls                   |
| Kubernetes              | Network orchestration                                                 |

### D. Trust Evaluation Models

Trust evaluation has evolved from static, to reputation-based, and dynamic models. Static approaches assign fixed weights to trust attributes regardless of signal quality, making them insensitive to changing operating conditions [9]. Reputation-based models leverage historical interactions to estimate trust [19], but do not effectively distinguish transient compromise from persistent environmental variability. Dynamic approaches employ adaptive weighting, including entropy-based methods and machine-learning [15], [20].

### E. Dempster-Shafer Theory in Network Security

Dempster-Shafer theory [21] has been applied to intrusion detection [22], anomaly detection [23], and trust management [24]. Unlike Bayesian methods, it explicitly models epistemic uncertainty through belief mass assigned to a frame of discernment, avoiding forced probability assignments with incomplete evidence. However, existing approaches assume uniform source reliability, neglect the relationship between signal variance and reliability [25], and fuse evidence within a single domain rather than across heterogeneous domains.

### F. Temporal Trust Dynamics

Temporal trust employs linear or exponential decay to reduce the influence of stale evidence [26], [27]. While effective in capturing evidence freshness, they do not account for consistency of observations. Continuous Adaptive Risk

and Trust Assessment (CARTA) frameworks combine temporal and contextual information to balance trust revocation against prolonged trust retention [28]. The proposed ensemble model addresses this through a dual-horizon architecture that considers historical stability and recent observations.

### G. Gap Analysis

No prior testbed integrates a trust engine supporting Dempster-Shafer fusion, variance-based weighting and temporal decay, with SDP/SDN enforcement in a reproducible, low-resource manner. Existing implementations either treat trust as static binary rules or fixed weights, depend on proprietary policy engines, or lack end-to-end integration across control and data planes. This paper directly addresses this gap.

## III. TESTBED ARCHITECTURE AND DESIGN PRINCIPLES

The proposed testbed integrates Mininet for SDN-based network emulation, GNS3 for high-fidelity topology emulation and telemetry generation, and Docker for lightweight service virtualization. This hybrid architecture balances functional realism with the resource constraints of commodity hardware.

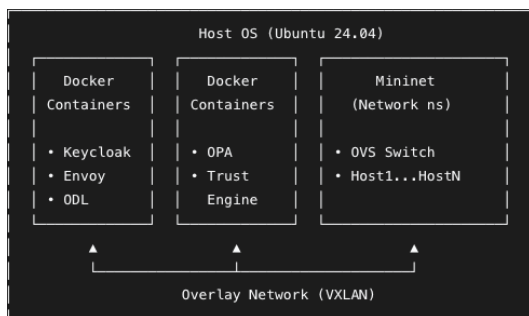
### A. Design Principles

The architecture is guided by five principles: (1) operation on commodity hardware without proprietary software for resource-constrained philosophy; (2) decoupled Policy Decision Point (PDP) and Policy Enforcement Point (PEP) for modular trust-policy evaluation; (3) automated deployment through Docker Compose and infrastructure-as-code for reproducibility; (4) progressively increasing sophistication of trust models, from static policies to Dempster-Shafer evidence fusion with temporal decay; and (5) complete version-controlled reproducibility of the testbed.

### B. Logical ZTA Component Mapping

Figure 1 presents the conceptual architecture, while Table 3 maps the implemented components to the logical entities defined in [3], demonstrating alignment with the Zero Trust

**Figure 1.** High-level architecture diagram (conceptual)



### C. Physical and Virtualization Layers

The testbed is hosted on Ubuntu 24.04 LTS and is designed to operate on commodity hardware (minimum 4 CPU cores,

16 GB RAM, and 256 GB SSD). Mininet with Open vSwitch provides SDN-based network emulation, with optional GNS3 integration for scenarios requiring greater network fidelity. Docker and Docker Compose orchestrate the control-plane services including: Keycloak, Open Policy Agent (OPA), Envoy, and OpenDaylight thus providing lightweight isolation, portability, and reproducible deployment.

**Table 3.** Mapping of NIST ZTA components to testbed implementations

| ZTA Component                         | Testbed Implementation                                         | Role                                                                                                     |
|---------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <b>Policy Decision Point (PDP)</b>    | Open Policy Agent (OPA) + Python trust engine                  | Computes trust scores and makes access decisions based on telemetry and policies.                        |
| <b>Policy Administrator (PA)</b>      | OpenDaylight SDN controller + custom orchestration             | Translates PDP decisions into network flow rules (OpenFlow) and service-mesh configurations.             |
| <b>Policy Enforcement Point (PEP)</b> | Open vSwitch (L3), Envoy Proxy (L7), eBPF(cilium and tetragon) | Enforces allow/deny decisions at network and application layers.                                         |
| <b>Identity Provider (IdP)</b>        | Keycloak, SPIFFE/SPIRE                                         | Provides authentication and device posture attributes (OAuth2/OIDC).                                     |
| <b>Telemetry Sources</b>              | Mininet hosts, OVS, Envoy, GNS3 nodes, Suricata, Zeek          | Stream real-time metadata (identity, device health, network variance, application behaviour, NIDS data). |
| <b>Subject (User/Device)</b>          | Mininet hosts or GNS3 virtual machines                         | Source of telemetry data                                                                                 |
| <b>Resource</b>                       | Containerised services (e.g., web servers, databases)          | Source of telemetry data                                                                                 |

## IV. IMPLEMENTATION DETAILS

### A. Testbed Preparation

The steps in Table 4 prepare a minimal installation for testbed deployment. These are later automated based on a yaml script for easier reproducibility.

**Table 4.** Testbed Deployment Workflow

#### Host Preparation

```
sudo apt update && sudo apt install -y git curl bridge-utils net-tools
```

**Resource utilization**

**vmstat-a**

**Docker**

```
sudo apt install -y docker.io docker-compose
sudo systemctl enable docker
sudo usermod -aG docker $USER
newgrp docker
```

**OpenVswitch-PEP-Network Layer**

```
sudo apt install -y openvswitch-switch
sudo systemctl start openvswitch-switch
ovs-vsctl show
```

**Mininet- Emulator**

```
sudo apt install -y mininet
sudo mn --test pingall
```

**OpenDaylight**

```
docker run -d \
    --name opendaylight \
    -p 8181:8181 -p 6653:6653 \
    opendaylight/odl:latest
```

**Keycloak-Identity and Device Trust Anchor**

```
docker run -d \
    --name keycloak \
    -p 8080:8080 \
    -e KEYCLOAK_ADMIN=admin \
    -e KEYCLOAK_ADMIN_PASSWORD=admin \
    quay.io/keycloak/keycloak:latest start-dev
```

**B. Architectural Workflow**

The proposed testbed follows a five-stage workflow. Firstly, Keycloak authenticates users, issuing JWTs containing identity and session claims. Secondly, the trust engine collects multi-domain telemetry (network traffic, device posture, application behaviour, and resource metadata). Thirdly, the Engine computes a trust score using the ensemble model. Fourthly, OPA evaluates the trust score against rego policy rego rules as in Table 5, to determine the access tier. Finally, the Policy Administrator enforces decisions through OpenDaylight and Open vSwitch's micro-segmentation and application-level access control.

The network is emulated in Mininet using a three-tier topology comprising corporate, DMZ, and remote segments interconnected through an OVS core switch. Fine-grained micro-segmentation is achieved through dynamically installed OpenFlow rules updated in response to trust evaluations. For reproducibility, Table 6 shows repository replication process.

**Table 5.** Access Tiers Rego policy

```
package trust.policy

default allow = false
default access_level = "no_access"
access_level = "full_access" {
    input.trust_score > 0.75
}
access_level = "limited_access" {
    input.trust_score >= 0.45
    input.trust_score <= 0.75
}
```

```
}
access_level = "no_access" {
    input.trust_score < 0.45
}
```

**Table 6.** Automated Deployment

```
# Clone repository and deploy
git clone https://github.com/[redacted]/dcta.git
cd dcta
docker compose up -d # IdP, OPA, ODL containers
sudo python3 deploy_mininet.py # Network topology
python3 run_ensemble_scenarios.py # Execute evaluation
```

**C. Design Considerations and Future Extensions**

The testbed employs Docker Compose to maximize reproducibility and minimize resource overhead, providing deterministic deployment and seamless integration with Mininet and OpenvSwitch. Although the architecture is readily portable to Kubernetes for production-scale deployments, Docker offers greater experimental repeatability.

**V. VALIDATION SCENARIOS AND EXPERIMENTAL METHODOLOGY****A. Ensemble Trust Engine**

The proposed Trust Engine integrates variance-based dynamic weighting, Dempster-Shafer (DS) evidential fusion, and dual-horizon temporal trust dynamics to support adaptive trust evaluation in a Zero Trust environment. This paper focuses on their integration within a reproducible testbed. The ensemble trust score is computed as:

$$T_{\text{ensemble}}(t) = \underbrace{W_{\text{short}}(t) \cdot T_{\text{instant}}(t)}_{\text{Fresh Signal (Verification)}} + \underbrace{(1 - W_{\text{short}}(t)) \cdot T_{\text{prev}} \cdot D_{\text{long}}(\Delta t)}_{\text{Historical Inertia (Continuity)}}$$

Where  $T_{\text{instant}}$  represents the current trust obtained through DS fusion,  $W_{\text{short}}$  controls the influence of recent evidence,  $T_{\text{prev}}$  is the previous trust state, and  $D_{\text{long}}$  models long-term trust decay.

Trust is evaluated across four domains. For each domain, a sliding-window variance estimates signal reliability, with higher variance reducing the corresponding evidence weight before fusion. Domain trust scores are converted into DS Basic Probability Assignments (BPAs), enabling explicit representation of uncertainty and robust multi-domain evidence fusion. The resulting belief is transformed into an instantaneous trust score using the Pignistic probability. To support continuous authentication, the engine combines this instantaneous assessment with historical trust through dual-horizon temporal weighting, balancing responsiveness to recent events against resilience to transient fluctuations. The final ensemble trust score is mapped to configurable policy thresholds that grant Full, Limited, or No Access, enabling risk-adaptive access control under uncertainty.

**B. Experimental Design**

Five trust evaluation models of increasing sophistication were compared: Baseline (no policy enforcement), Single Do-

main (binary trust based on one attribute), Multi-Domain Hierarchical (static rule-based trust), Static Weighted (fixed trust weights), and the proposed Ensemble, which integrates variance-based adaptive weighting, Dempster-Shafer evidence fusion, temporal decay, and residual trust. They were evaluated using six representative enterprise network scenarios in Table 7: Corporate Office, Remote VPN, Public Wi-Fi, BYOD, Compromised Host, and Untrusted Device/Geofence. These scenarios span the full trust spectrum, varying domain trust levels and variance characteristics to represent stable, noisy, and adversarial environments. Each scenario was executed over 30 evaluation epochs (1-minute intervals), with domain scores generated through controlled randomization around predefined means to ensure reproducible variance profiles.

**Table 7.** Scenario Configuration Matrix

| Scenario                  | Network Profile              | Device Profile               | Data/Identity                | Application                  | Variance Characteristics             | Ground Truth   |
|---------------------------|------------------------------|------------------------------|------------------------------|------------------------------|--------------------------------------|----------------|
| Corporate Office          | High ( $\approx 0.94$ )      | High ( $\approx 0.95$ )      | High ( $\approx 0.88$ )      | High ( $\approx 0.90$ )      | Stable ( $\sigma^2 < 0.02$ )         | Full Access    |
| Remote VPN                | Mode rate ( $\approx 0.84$ ) | High ( $\approx 0.98$ )      | High ( $\approx 0.92$ )      | High ( $\approx 0.91$ )      | Variable ( $\sigma^2 \approx 0.04$ ) | Full Access    |
| Public Wi-Fi              | Low ( $\approx 0.30$ )       | Mode rate ( $\approx 0.75$ ) | Mode rate ( $\approx 0.60$ ) | Mode rate ( $\approx 0.72$ ) | Chaotic ( $\sigma^2 > 0.20$ )        | Limited Access |
| BYOD                      | High ( $\approx 0.89$ )      | Low ( $\approx 0.40$ )       | Mode rate ( $\approx 0.49$ ) | Mode rate ( $\approx 0.62$ ) | $\sigma^2 \approx 0.15$              | Limited Access |
| Compromised Host          | Low ( $\approx 0.25$ )       | Low ( $\approx 0.21$ )       | Low ( $\approx 0.21$ )       | Low ( $\approx 0.27$ )       | Chaotic across all domains           | No Access      |
| Untrusted Device/Geofence | Mode rate ( $\approx 0.31$ ) | Low ( $\approx 0.34$ )       | Mode rate ( $\approx 0.30$ ) | Low ( $\approx 0.30$ )       | Unstable ( $\sigma^2 \approx 0.10$ ) | No Access      |

### C. Evaluation Methodology

Performance was assessed using classification accuracy, convergence time, ensemble stability, per-evaluation latency, and breach containment time. Each scenario was repeated across 50 independent simulation runs with different random

seeds, and results are reported as mean  $\pm$  standard deviation. Statistical significance was evaluated using the Wilcoxon signed-rank test ( $p < 0.01$ ), with effect sizes measured using Cliff's delta. The ensemble model employed a variance penalty factor of  $\alpha = 10.0$  and short- and long-term temporal decay rates of  $\lambda = 3.0$  and  $\lambda = 0.5$ , respectively.

### D. Implementation

The prototype was implemented in Python 3.12 using modular components:

- `dynamic_trust_weighting.py`: Variance computation, dynamic weight calculation, intra-domain score aggregation.
- `weighted_belief_fusion.py`: DS mass function construction, Dempster combination rule, Pignistic transformation.
- `ensemble_trust_simulator.py`: Dual-horizon temporal integration, ensemble computation.
- `ds_utils.py`: Generic Dempster-Shafer library (mass function manipulation, belief calculation).
- `run_ensemble_scenarios.py`: Scenario orchestration, data logging, and visualisation.

## VI. EXPERIMENTAL RESULTS

### A. Trust Classification Performance

The proposed Ensemble Trust Engine correctly classified all six evaluation scenarios at steady state ( $t \geq 10$ ) over 50 independent runs as in Table 8. High-trust environments (Corporate Office and Remote VPN) consistently received Full Access, uncertain environments (Public Wi-Fi and BYOD) were assigned Limited Access, while compromised and untrusted scenarios were correctly denied access. Trust trajectories remained stable throughout evaluation without crossing decision thresholds, demonstrating robust and consistent policy enforcement.

The results further show that variance-based adaptive weighting effectively suppresses unreliable domains while allowing stable evidence to dominate trust estimation. Consequently, transient network instability (Remote VPN and Public Wi-Fi) did not result in unnecessary access revocation, whereas persistent multi-domain compromise consistently produced sustained denial. The ensemble temporal model progressively reinforced trust under stable conditions while preventing rapid oscillations caused by short-term fluctuations.

**Table 8.** Ensemble Trust Scores and Access Decisions

| Scenario         | Tensem ble ( $t=0$ ) | Tensem ble ( $t=15$ ) | Tensem ble ( $t=29$ ) | Decision                 |
|------------------|----------------------|-----------------------|-----------------------|--------------------------|
| Corporate Office | 0.795 $\pm$ 0.003    | 0.796 $\pm$ 0.002     | 0.792 $\pm$ 0.002     | Full Access $\checkmark$ |
| Remote VPN       | 0.787 $\pm$ 0.005    | 0.787 $\pm$ 0.004     | 0.783 $\pm$ 0.004     | Full Access $\checkmark$ |

|                                  |             |             |             |                         |
|----------------------------------|-------------|-------------|-------------|-------------------------|
| <b>Public Wi-Fi</b>              | 0.570±0.012 | 0.606±0.011 | 0.604±0.010 | <b>Limited Access</b> ✓ |
| <b>BYOD</b>                      | 0.558±0.015 | 0.607±0.013 | 0.599±0.012 | <b>Limited Access</b> ✓ |
| <b>Compromised Host</b>          | 0.227±0.009 | 0.302±0.008 | 0.299±0.007 | <b>No Access</b> ✓      |
| <b>Untrusted Device/Geofence</b> | 0.360±0.011 | 0.343±0.010 | 0.346±0.009 | <b>No Access</b> ✓      |

### B. Performance Evaluation

The complete trust evaluation pipeline required an average of 18.5 ms per evaluation cycle, satisfying the design target of <20 ms as in Table 9. Mathematical trust computation accounted for 6.9 ms, with memory state management representing the primary overhead. Total session establishment latency increased from 8–12 ms (baseline) to 23–32 ms with the Ensemble model, remaining well within acceptable enterprise response times.

During simulated compromise scenarios, trust values fell below the revocation threshold within one evaluation interval, enabling access revocation in less than 15 seconds, compared with several minutes for baseline and static trust approaches. This substantially reduces the attack window for lateral movement while maintaining responsiveness.

### C. Stability and Scalability

Compared with a spatial-only Dempster-Shafer model, the Ensemble approach reduced trust score volatility by 33–56% across all scenarios ( $p < 0.001$ , Cliff's  $\delta > 0.80$ ), demonstrating statistically significant improvements in access stability and reducing false-positive revocations caused by transient fluctuations.

The prototype remained lightweight, consuming approximately 3.2 GB RAM and 10–15% CPU at idle, increasing to 4.1 GB RAM and 35–40% CPU under 100 concurrent sessions. Trust reevaluation scaled linearly ( $\approx 0.2$  ms per additional session), while SDN flow installation remained below 50 ms, confirming suitability for small- to medium-scale enterprise Zero Trust deployments.

**Table 9.** Latency Breakdown per Evaluation Epoch

| Component                         | Complexity                   | Latency (ms) |
|-----------------------------------|------------------------------|--------------|
| Variance computation              | $O(N)$ per domain, $N=10$    | 2.1          |
| Weight normalization              | $O(D)$ , $D=4$               | 0.3          |
| Mass function construction        | $O(D)$ , $D=4$               | 0.4          |
| DS pairwise fusion (3 iterations) | Closed form for binary frame | 3.8          |

|                                         |                 |      |
|-----------------------------------------|-----------------|------|
| Pignistic transformation                | $O(1)$          | 0.1  |
| Temporal integration (ensemble mixture) | $O(1)$          | 0.2  |
| memory state read/write                 | Network I/O     | 8.4  |
| OPA policy evaluation                   | Rego evaluation | 3.2  |
| Total                                   |                 | 18.5 |

## VII. DISCUSSION

The proposed testbed demonstrates that integrating identity-aware SDP enforcement with SDN enables continuous, low-latency micro-segmentation driven by dynamic trust evaluation. The results further demonstrate that trust should be modelled as a dynamic continuous property reflecting both historical stability and current behaviour.

Compared with existing Zero Trust testbeds, the proposed platform uniquely provides an integrated dynamic trust engine, multiple policy models, reproducible Infrastructure-as-Code deployment, and comprehensive empirical evaluation, including latency, false-positive, and breach-containment analyses as in Table 10. The substantially reduced false positives illustrate an effective security–usability trade-off. These capabilities establish a practical and extensible research platform for evaluating adaptive trust mechanisms under standardized conditions.

**Table 10.** Comparison of Results with Other works

| Feature                        | Horne [1] | Lefebvre [10] | This work                 |
|--------------------------------|-----------|---------------|---------------------------|
| SDP/SDN integration            | ✓         | ✓             | ✓                         |
| Dynamic trust engine           | ✗         | ✗             | ✓ (DS + variance + decay) |
| Multiple policy models         | ✗         | Partial       | ✓ (6 models)              |
| Reproducible artifacts         | Partial   | ✗             | ✓ (Full IaC)              |
| Empirical latency measurements | ✗         | ✓             | ✓                         |
| False positive analysis        | ✗         | ✗             | ✓                         |
| Breach containment measurement | ✗         | ✗             | ✓                         |

## VIII. LIMITATIONS AND DESIGN IMPROVEMENTS

The proposed framework is modular and vendor-agnostic, enabling interoperability with different implementations of identity providers, SDN controllers, and policy enforcement mechanisms without modifying the trust engine. Evaluation was conducted in a containerized Mininet environment using simulated telemetry and binary trust states, limiting direct generalization to production-scale deployments. Although the

framework demonstrates scalability for small and medium-sized enterprise environments, further validation using real security telemetry, hardware trust anchors (e.g., TPM 2.0), multi-state trust models, and distributed deployments is required. Design improvements to address these include integrating SPIFFE/SPIRE for workload identity, eBPF-based enforcement (e.g., Cilium and Tetragon), and real network telemetry from Zeek and Suricata to evaluate trust using operational protocol metadata, behavioral anomalies, and intrusion alerts. These are summarized in *Table 11*.

Each technology addresses a specific limitation identified in the paper while preserving the testbed's design principles: open-source, reproducible, and lightweight.

**Table 11.** Future Improvements to Challenges and Limitations

| Technology                            | Testbed Layer | Current State                    | Extension                           |
|---------------------------------------|---------------|----------------------------------|-------------------------------------|
| <b>NIST SP 1800-35</b> [6]            | Standards     | SP 800-207 mapping               | Cross-reference with 1800-35 builds |
| <b>AI/ML scoring</b> [20], [29], [30] | Trust Engine  | Deterministic variance weighting | Adaptive $\alpha$ via RL            |
| <b>Docker</b> [31], [32]              | Orchestration | Docker Compose                   | K8s for production scale            |
| <b>SPIFFE/SPIRE</b> [33], [34]        | Identity      | Keycloak (human IdP)             | Machine identity + mTLS             |
| <b>eBPF/Cilium</b> [35], [36]         | Enforcement   | OVS + Envoy                      | Kernel-level, sub-ms                |
| <b>Zeek + Suricata</b> [37], [38]     | Telemetry     | Simulated Gaussian               | Real network metadata + IDS         |

## IX. REPRODUCIBILITY AND REUSABILITY

### A. Code Repository Structure

To facilitate reproducibility, all source code, configuration files, topology definitions, policy scripts, and simulation outputs are provided in a public repository [39].

The repository includes implementations of the baseline, temporal decay, and ensemble trust models, together with automation scripts for provisioning the Docker–Mininet testbed and reproducing the experimental scenarios.

The testbed is designed to be easily extended: researchers can modify Rego policies, change trust engine parameters (decay rates, weights), add new telemetry sources, or integrate additional enforcement points (e.g., a service mesh).

## X. CONCLUSION AND FUTURE WORK

This paper presented a lightweight, reproducible Zero Trust testbed that bridges theoretical trust evaluation with

operational enforcement through an open-source, containerized implementation aligned with NIST SP 800-207. The architecture integrates SDP-based identity verification, SDN-enabled microsegmentation, and an Ensemble Trust Engine combining variance-based adaptive weighting, Dempster-Shafer evidential fusion, and dual-horizon temporal dynamics to support continuous, context-aware access control.

Experimental evaluation demonstrated that the proposed framework correctly enforced dynamic access policies while maintaining low computational overhead (approximately 18.5 ms per evaluation cycle) and significantly reducing trust-score volatility and breach containment time compared with static approaches. The modular separation of the trust engine from policy enforcement proved particularly effective, enabling independent evolution of trust algorithms without reconfiguring the control plane, while the containerized design ensures reproducibility on commodity hardware.

Future work will extend the framework with hardware-backed device attestation [40], reinforcement learning for adaptive trust parameter tuning, multi-state trust models, cloud-scale deployments [30], [41], and formal policy verification to further improve scalability, robustness, and real-world applicability.

## XI. ACKNOWLEDGEMENT

During the preparation of this work, the authors used Deepseek to summarize and aggregate individual docker container configurations to one yaml script for all containers and Claude (opus 4.6) to check inter-chapter cohesion and narrative flow. The authors take full responsibility for the content of this publication.

## REFERENCES

- [1] D. Horne, “Leveraging Software Defined Perimeter (SDP), Software Defined Networking (SDN), and Virtualization to Build a Zero Trust Testbed with Limited Resources,” 2022.
- [2] D. T. Van, T. B. Hung, T. D. Thang, and N. K. Giao, “Application of Machine Learning in Malicious IoT Classification and Detection on Fog-IoT Architecture,” presented at the The Seventh International Conference on Research in Intelligent and Computing in Engineering, Feb. 2022, pp. 299–303. doi: 10.15439/2022R39.
- [3] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “NIST Special Publication 800-207 -Zero Trust Architecture.” Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, Aug. 10, 2020. doi: <https://doi.org/10.6028/NIST.SP.800-207>.
- [4] R. Ward and B. Beyer, “BeyondCorp: A New Approach to Enterprise Security,” *login.*, vol. 39, No. 6, pp. 6–11, 2014.
- [5] “Executive Order 14028,” Improving the Nation’s Cybersecurity. Accessed: Jun. 29, 2026. [Online]. Available: <https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity>
- [6] National Institute of Standards and Technology, “Implementing a Zero Trust Architecture,” U.S. Department of Commerce, NIST Special Publication 1800-35, 2025.
- [7] N. Ngema, B. Nleya, and R. C. Maswanganyi, “A Review of Zero Trust Architecture: Principles, Applications, and Implementation Challenges in Communication, Navigation, and Surveillance (CNS) Systems,” *Sensors*, vol. 26, no. 12, 2026, doi: 10.3390/s26123813.
- [8] R. Vaughn, T. Adewale, and J. Ajayi, “Zero-Trust Architectures for Securing Cloud-Native Infrastructure,” Nov. 2024.

- [9] N. F. Syed, S. W. Shah, A. Shaghghi, A. Anwar, Z. Baig, and R. Doss, "Zero Trust Architecture (ZTA): A Comprehensive Survey," *IEEE Access*, vol. 10, pp. 57143–57179, 2022, doi: 10.1109/ACCESS.2022.3174679.
- [10] M. Lefebvre, S. Nair, D. W. Engels, and D. Horne, "Building a Software Defined Perimeter (SDP) for Network Introspection," in *2021 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2021, pp. 91–95. doi: 10.1109/NFV-SDN53031.2021.9665152.
- [11] Y. Palmo, S. Tanimoto, H. Sato, and A. Kanai, "A Consideration of Scalability for Software Defined Perimeter Based on the Zero-trust Model," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2021, pp. 717–724. doi: 10.1109/IIAI-AAI53430.2021.00127.
- [12] J. Adorno, "Evolution of Secure Access: From VPN to SDP-Enabled Zero Trust Network Access (ZTNA) | Zscaler." Accessed: May 19, 2026. [Online]. Available: <https://www.zscaler.com/blogs/product-insights/evolution-secure-access-vpn-sdp-enabled-zero-trust-network-access-ztna>
- [13] Trio.so, "CARTA: An Overview of Gartner's Continuous Adaptive Risk and Trust Assessment," Trio Security, 2025.
- [14] A. Tripathi *et al.*, "Real Time Adaptive Access Control with Behavioral Analytics for Enhanced Cybersecurity in IoT and Cloud Systems," presented at the The Ninth International Conference on Research in Intelligent Computing in Engineering, May 2025, pp. 151–155. doi: 10.15439/2024R75.
- [15] S. Munasinghe, N. Piyaarathna, E. Wijerathne, U. Jayasinghe, and S. Namal, "Machine Learning Based Zero Trust Architecture for Secure Networking," in *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, 2023, pp. 1–6. doi: 10.1109/ICIIS58898.2023.10253610.
- [16] M. Aaqib, A. Ali, L. Chen, and O. Nibouche, "Behaviour-based trust assessment for the Internet of Things systems using multi-classifier ensemble learning and Dempster-Shafer fusion," *Neural Computing and Applications*, vol. 37, no. 26, pp. 22191–22214, Sep. 2025, doi: 10.1007/s00521-025-11273-8.
- [17] B. Lantz and B. O'Connor, "A Mininet-based Virtual Testbed for Distributed SDN Development," *ACM SIGCOMM Computer Communication Review*, vol. 45, pp. 365–366, Aug. 2015, doi: 10.1145/2829988.2790030.
- [18] O. Flauzac, E. Robledo, and F. Nolot, "Is Mininet the Right Solution for an SDN Testbed?," Dec. 2019, pp. 1–6. doi: 10.1109/GLOBE-COM38437.2019.9013145.
- [19] F. Ullah *et al.*, "Deep Trust: A Novel Framework for Dynamic Trust and Reputation Management in the Internet of Things (IoT) Based Networks," *IEEE Access*, vol. PP, pp. 1–1, Jan. 2024, doi: 10.1109/ACCESS.2024.3409273.
- [20] D. Haider, S. Mushtaq, H. Ali, and M. Su'ud, "Enhancing Zero Trust Cybersecurity using Machine Learning and Deep Learning Approaches," *Journal of Informatics and Web Engineering*, vol. 4, pp. 24–34, Oct. 2025, doi: 10.33093/jiwe.2025.4.3.2.
- [21] G. Shafer, "A Mathematical theory of Evidence," 1976.
- [22] M. Maliha and M. Atiquzzaman, "Q-ID: A Reinforcement Learning Framework for Adaptive Intrusion Detection," presented at the 20th Conference on Computer Science and Intelligence Systems (FedC-SIS), Kraków, Poland, Oct. 2025, pp. 35–42. doi: 10.15439/2025F1820.
- [23] P. D. J. Fidalgo, A. Pasic, and S. G. Zarzosa, "Dynamic Threat Intelligence for Improvement of Resilience of Critical Infrastructure During Pandemics," presented at the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, Oct. 2024, pp. 591–596. doi: 10.15439/2024F8106.
- [24] N. Wang and D. Wei, "An Adaptive Dempster-Shafer Theory of Evidence Based Trust Model in Multiagent Systems," *Applied Sciences*, vol. 12, no. 15, p. 7633, Jul. 2022, doi: 10.3390/app12157633.
- [25] L. Xiong, X. Su, and H. Qian, "Conflicting evidence combination from the perspective of networks," *Inf. Sci.*, vol. 580, no. C, pp. 408–418, Nov. 2021, doi: 10.1016/j.ins.2021.08.088.
- [26] H. Wang, J. Jiang, and W. Li, "A Dynamic Trust Model Based on Time Decay Factor," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCCom/IOP/SCI)*, 2018, pp. 2048–2051. doi: 10.1109/SmartWorld.2018.00343.
- [27] J. D. L. C. Kiguigouléli Ki and B. Zerbo, "Modeling Dynamic Trust for Cooperative Operations in Edge Computing Environments," in *2025 IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD)*, 2025, pp. 1–8. doi: 10.1109/MNE3SD67637.2025.11323241.
- [28] IBM Security, "Cost of a Data Breach Report 2024," p. 46, 2024.
- [29] M. Ahmed, M. Mark, and M. Joy, "Intelligent Trust: Leveraging AI for Dynamic Policy Orchestration in Zero Trust Security Architectures," *International Journal of Artificial Intelligence*, Aug. 2021.
- [30] V. Kalekar, "Uncertainty-Aware Reinforcement Learning for Zero Trust: An Empirical Evaluation of PPO, MC Dropout, and Bayesian Neural Networks," in *Proceedings of the 12th International Conference on Information Systems Security and Privacy*, Marbella, Spain: SCITEPRESS - Science and Technology Publications, 2026, pp. 505–514. doi: 10.5220/0014245800004061.
- [31] Dockerdocs, "DockerDocs Manuals," Docker Documentation. Accessed: Apr. 14, 2026. [Online]. Available: <https://docs.docker.com/desktop/>
- [32] D. Silva, J. Rafael, and A. Fonte, "Toward Optimal Virtualization: An Updated Comparative Analysis of Docker and LXD Container Technologies," *Computers*, vol. 13, no. 4, p. 94, Apr. 2024, doi: 10.3390/computers13040094.
- [33] "The SPIFFE Standard | SPIFFE." Accessed: Jun. 29, 2026. [Online]. Available: <https://spiffe.io/docs/latest/spiffe-specs/>
- [34] L. C. Cardoso, M. A. Marques, P. H. B. Correia, H. Z. Cochak, C. C. Miers, and M. A. Simplicio, "Next-Generation SPIFFE/SPIRE Identity Management Systems with Post-Quantum Cryptography Algorithms," in *2025 IEEE 25th International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 2025, pp. 154–163. doi: 10.1109/CCGRID64434.2025.00033.
- [35] "eBPF Security: Top 5 Use Cases, Challenges & Best Practices." Accessed: Jun. 29, 2026. [Online]. Available: <https://www.oligo.security/academy/ebpf-security-top-5-use-cases-challenges-and-best-practices>
- [36] "Installation and Configuration," Tetragon - eBPF-based Security Observability and Runtime Enforcement. Accessed: Jun. 29, 2026. [Online]. Available: <https://tetragon.io/docs/installation/>
- [37] "Zeek: Get Zeek," Zeek. Accessed: Jun. 29, 2026. [Online]. Available: <https://zeek.org/get-zeek/>
- [38] "2. Quickstart guide — Suricata 9.0.0-dev documentation." Accessed: Jun. 29, 2026. [Online]. Available: <https://docs.suricata.io/en/latest/quickstart.html>
- [39] "tnkinyili/DCTA: Dynamic and Context Aware Trust Algorithm for Zero Trust Enforcement in Heterogeneous Enterprise Environments," GitHub. Accessed: Apr. 16, 2026. [Online]. Available: <https://github.com/tnkinyili/DCTA>
- [40] M. Alawneh and I. M. Abbadi, "Integrating Trusted Computing Mechanisms with Trust Models to Achieve Zero Trust Principles," in *2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, Milan, Italy: IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/IOTSMS58070.2022.10062269.
- [41] S. Thapaliya and S. Jha, "Harnessing AI for Enhanced Identity Management: Addressing Cybersecurity Challenges in the Digital Age," presented at the The Ninth International Conference on Research in Intelligent Computing in Engineering, May 2025, pp. 71–77. doi: 10.15439/2024R34.

# Cosine Colorization for Enhancement of Magnetic Resonance Angiography (MRA)

Modupe Odusami  
0000-0003-1082-185X  
Department of Applied Informatics  
Vytautas Magnus University, Akademija  
Lithuania  
modupe.olufunke.odusami@vdu.lt

Olusola Abayomi-Alli  
0000-0003-2513-5318  
Faculty of Fundamental Sciences  
Vilnius Gediminas Technical University  
Vilnius, Lithuania  
olusola-oluwakemi.abayomi-alli@vilniustech.lt

Robertas Damaševičius  
0000-0001-9990-1084  
Department of Applied Informatics  
Vytautas Magnus University, Akademija  
Lithuania  
robertas.damasevicius@vdu.lt

Rytis Maskeliūnas  
0000-0002-2809-2213  
Centre of Real Time Computer Systems  
Kaunas University of Technology  
Kaunas, Lithuania  
rytis.maskeliunas@ktu.lt

**Abstract**—Non-invasive imaging modalities play a crucial role in healthcare. Magnetic resonance angiography (MRA) has enhanced access to blood vessel information. MRA grayscale preserves anatomical fidelity but offers limited perceptual separation. This study investigates the use of cosine colorization to enhance feature discrimination in MRA. Cosine colorization utilizes the Frangi vesselness filter combined with statistical thresholding to isolate vascular regions. Quantitative and statistical evaluations were conducted on 20 MRA volumes, comparing grayscale baselines with full-volume and vessel-restricted cosine mappings. While full-volume colorization introduces structural distortion ( $SSIM \approx 0.85$  and  $NMSE \approx 0.018$ ), the vessel-restricted approach maintains near-perfect anatomical fidelity ( $SSIM \approx 1.0$  and  $NMSE \approx 0$ ). Statistical analysis confirms significant improvements in perceptual separability ( $p < 10^{-5}$ ) without degradation of structural metrics for the vessel-restricted cosine approach. The results show that this approach provides an interpretable enhancement of vascular structures.

**Index Terms**—Cosine Colorization, Enhancement, Magnetic Resonance Angiography, Vessel-Restricted Mapping.

## I. INTRODUCTION

THE Magnetic Resonance Angiography (MRA) imaging modality provides information on blood vessels without exposing patients to ionizing radiation [1]. MRA volumes are typically visualized using grayscale colormaps in clinical practice [2]. The application of windowing techniques in which image values are connected to corresponding shades of gray levels has been used to improve the intensity range of images [2]. The complexity of distinguishing MRA vessels from adjacent structures often poses high contrast resolution and limited perceptual clarity of MRA representations [3]. Colorization has long been understood as an effective way to enhance visual perception, as it enriches image representation by introducing chromatic information that adds a perceptual dimension [3]. The ability to colorize grayscale images can

greatly improve their visual quality and make them easier to interpret [4]. In medical imaging, however, colorization must be carefully designed to preserve anatomical structure and avoid introducing misleading artifacts [5], [6]. Nonlinear color-mapping approaches have been shown to improve visual distinction. By enhancing structural features, they often make anatomical information more noticeable than in standard grayscale images [7], [8]. However, their effectiveness in preserving anatomical integrity—particularly in volumetric angiographic data—has not yet been thoroughly investigated.

The use of colorization techniques in medical imaging shows that volume visualization capabilities could be improved, and a well-interpretable rendering of complex data could be achieved [9]. In multi-modal PET and CT imaging, dual-threshold colorization allows for better observation and a clear distinction between varying levels of metabolic activity. Similarly, the application of colorization to 3D MRI volumes facilitates the generation of photorealistic 3D renderings, bypassing the limitations of manual transfer functions while maintaining structural integrity [10]. Furthermore, in 3D spectral OCT of the retina, adaptive color mapping enhances instrument visibility and contrast, while visually setting apart semantic regions to inform clinicians of data reliability [10].

We focus specifically on cosine-based colorization as a minimal and interpretable transformation for MRA visualization. Given a normalized grayscale intensity field, we consider a cosine mapping [11] and a mixed cosine–grayscale formulation [12] that preserves luminance structure while introducing controlled chromatic variation. This study aims to characterize trade-offs inherent in periodic color mappings, thereby providing a foundation for subsequent work on constrained optimization and more flexible colorization models.

The main contributions of this work are as follows:

- Systematic evaluation of cosine-based colorization for MRA visualization.
- Analysis of the impact of full-volume versus vessel-restricted cosine mappings on anatomical fidelity and perceptual discriminability.
- Introduction of a quantitative evaluation framework combining structural and perceptual metrics for assessing medical image colorization.

## II. LITERATURE REVIEW

MRA volumes are conventionally visualized using grayscale colormaps, where voxel intensities are mapped directly to gray levels for display. In clinical practice, grayscale visualization remains the default mode for examining vascular morphology and structural continuity [13]. To improve visibility within grayscale displays, a windowing technique was utilized for maximum intensity projections to adjust the visual representation of the arterial segments for the diagnosis of acute ischemic stroke [14]. Furthermore, windowing was used to ensure that the representation of intra-aneurysmal flow was clear enough to be separated from surrounding structures for quantitative measurement [15]. However, windowing operates strictly through linear or piecewise-linear remapping of intensity values and does not introduce additional perceptual dimensions beyond luminance contrast.

Colorization techniques in medical imaging enhanced visual interpretation. Colorization with semantic segmentation, where region identity constrains color assignment and helps preserve structural boundaries. However, these methods rely on accurate segmentation and typically require annotated data. Another approach leverages domain-specific priors, such as region-of-interest (ROI) masks or anatomical constraints, to localize colorization. This can reduce global distortion and preserve contextual information, although it does not inherently guarantee consistency in luminance or intensity ordering within the colored regions.

Nonlinear extensions, such as gamma or logarithmic mappings, have been used to highlight particular intensity in selected regions [16], [17]. However, these approaches operate through monotonic remapping and only provide intensity scaling [18]. Hue-based ramps introduce chromatic variation and can improve perceptual separation. Nevertheless, they are usually designed for general-purpose visualization [19]. Periodic mappings, on the other hand, provide controlled and reproducible chromatic variation throughout the intensity domain by defining color as an oscillatory function of intensity. By parameterizing color channels using periodic functions like sine or cosine, these mappings allow for exact control over frequency and phase, providing a simplified and comprehensible way to modify color patterns [20]. However, the effect of cosine-based color mappings in medical imaging, particularly with respect to anatomical fidelity, luminance consistency, and intensity ordering, has not been comprehensively evaluated.

Since there is no ground-truth color reference, the assessment of colorization in medical imaging is based on how well the converted image preserves structural and functional

information from its original grayscale data. Among the most important factors to take into account are maintaining brightness consistency, intensity ordering, and visual aberrations, including distortions.

## III. METHODS

The pipeline used in this study is illustrated in Fig. 1. The pipeline includes dataset preparation and normalization of the input 3D MRA volume, production of representation variations, vessel-restricted masking for localized colorization, and quantitative evaluation. First, the input MRA volume is normalized and represented as a grayscale baseline. Then, cosine colorization is applied in two ways: as a full-volume mapping and as a vessel-restricted mapping. For the vessel-restricted representation, a vessel mask is obtained using multiscale Frangi vesselness filtering, anatomical foreground and central masking, and statistical thresholding. Finally, the resulting representations are compared using structural, ordering, and perceptual evaluation metrics.

### A. Dataset

The dataset used in this study is the Magnetic Resonance Angiography Atlas Dataset from <https://www.nitrc.org/projects/icbmmra/>, consisting of sixty-one images of healthy volunteers. This study used 20 three-dimensional MRA healthy volumes for the experiment. The study is concerned with the evaluation of cosine-based colorization; the subset was selected for a controlled paired comparison of the grayscale, full-volumetric, and vessel-restricted representations.

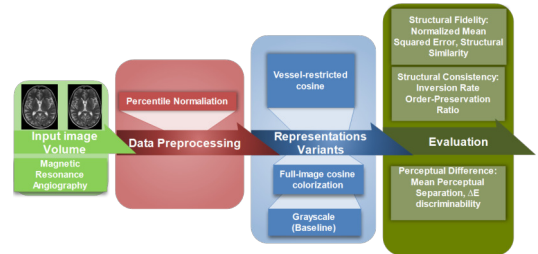


Fig. 1. The evaluation pipeline

### B. Data Representation

Let  $V(x, y, z) \in [0, 1]$  denote a normalized 3D MRA volume.

A percentile-based min-max scaling normalization technique was applied on the input image volume [21].

To enable direct comparison with colorized outputs, the grayscale volume is replicated across three channels as shown in Eq. (1)

$$\text{GrayB}(x) = [V(x), V(x), V(x)] \quad (1)$$

### C. Vessel-Restricted Application

To avoid unwanted color patterns in non-vascular tissue [22], cosine colorization is applied only within vessel regions using a vesselness-based region-of-interest (ROI) mask derived from multiscale filtering and statistical thresholding.

1) *Multiscale Frangi Vesselness*: Following the Hessian-based formulation described in [23]. The Gaussian filter  $G(x,s)$  is applied to the normalized image through convolution to suppress high-frequency noise. For each scale  $s \in \{1.0, 2.0, 3.0, \dots\}$ , the scale-space representation is:

$$V_s(x) = V_{01}(x) * G(x, s), \quad (2)$$

where  $G(x, s)$  is a Gaussian kernel with standard deviation  $s$ .

The Frangi parameters are fixed to:  $\beta = 0.5$ ,  $\gamma = 12$ .

Multiscale responses are averaged (mean) to improve robustness [24], as computed in Eq. (3).

$$S(x) = \frac{1}{N_s} \sum_s V_f(x, s) \quad (3)$$

where  $V_f(x, s)$  is the Frangi vesselness response at scale  $s$ , and  $N$  is the number of scales.

2) *Anatomical Foreground and Central Masking*: To reduce artifacts near the cortical rim and avoid false positives caused by image boundaries, vessel extraction is limited to the inner anatomical region [25]. A foreground mask  $M_{fg}$  is computed from  $V(x)$  using Otsu thresholding [25]. A 3D Euclidean distance transform is then computed inside  $M_{fg}$ , and a central mask is defined as

$$M_{\text{central}} = \{x \mid \text{dist}(x, \partial M_{fg}) > d\}, \quad (4)$$

where  $d = 20$  voxels.

3) *Statistical Vessel Extraction*: Vessel regions are defined using a statistical criterion computed within the central mask [26]. The vessel threshold is defined as

$$\tau = \mu_c + k\sigma_c, \quad (5)$$

where  $\mu_c$  and  $\sigma_c$  are the mean and standard deviation of  $S_c$ , and  $k = 0.4$ . The vessel mask is then defined as

$$M_{\text{vessel}} = \{x \mid S(x) > \tau\}. \quad (6)$$

#### D. Cosine Color Mapping

The cosine function is used to form smooth and continuous variations in color based on the input intensity values [11]. Cosine colorization embeds grayscale intensities into RGB channels using periodic cosine functions in Eq. (7).

$$\begin{aligned} R &= \frac{1 + \cos(2\pi f I)}{2} \\ G &= \frac{1 + \cos(2\pi f I + 0.5\phi)}{2} \\ B &= \frac{1 + \cos(2\pi f I + \phi)}{2} \end{aligned} \quad (7)$$

where  $f$  controls frequency and  $\phi$  controls phase offset. The parameters were fixed to  $f = \pi/3$  and  $\phi = 2\pi/3$  to provide a slow smooth color modulation with separate RGB phase responses. These values were used as a controlled base line, not an optimized setting.

#### E. Vessel-Only Cosine Application

Cosine mapping is applied only where  $x \in M_{\text{vessel}}$  shown.

$$V_{\text{RGB}}(x) = \begin{cases} (R(x), G(x), B(x)), & x \in M_{\text{vessel}} \\ (0, 0, 0), & \text{otherwise} \end{cases} \quad (8)$$

This isolates colorization to vascular structures and prevents visual distortion in surrounding tissue.

#### F. Mixed Cosine-Grayscale Representation

To preserve luminance structure while introducing chromatic variation, a mixed representation is defined:

$$V_{\text{mix}}(x) = (1 - \alpha) \text{GrayB}(x) + \alpha V_{\text{RGB}}(x) \quad (9)$$

where  $\alpha \in [0, 1]$  controls the contribution of colorization.

Each experiment's blending coefficient was set to  $\alpha = 0.10$  to enable a subtle chromatic enhancement.

#### G. Evaluation Metrics

Metrics were chosen to evaluate anatomical fidelity, intensity-order preservation, and vessel-background separability. Normalized Mean Squared Error (NMSE) and Structural Similarity (SSIM) measure luminance consistency with the original grayscale MRA; order-preservation ratio (OPR) and inversion rate assess preservation of vessel intensity ordering; and CIELAB-based  $\Delta E$ ,  $\Delta E_d'$ , and local contrast measure perceptual vessel distinction.

1) *Structural fidelity*: Structural fidelity was evaluated to determine how well the colorization process preserves the original grayscale appearance of the MRA volume. Since colorization modifies the RGB channels while the anatomical information is encoded primarily in image luminance, the colorized representation was first converted to a luminance image and then compared with the original normalized grayscale volume. Let  $V(x)$  denote the normalized grayscale volume and let  $T(x) = [R, G, B]$  denote the corresponding colorized output. The luminance of the colorized image was computed using the standard sRGB/Rec. 709 weighting coefficients [27]:

$$Y_T(x) = 0.2126 R(x) + 0.7152 G(x) + 0.0722 B(x). \quad (10)$$

Similarity between luminance image and original grayscale volume was first quantified using the Normalized Mean Squared Error (NMSE), which measures the average reconstruction error relative to the energy of the reference image:

$$\text{NMSE} = \frac{\text{mean}((Y_T - V)^2)}{\text{mean}(V^2) + 10^{-8}}. \quad (11)$$

Lower NMSE values indicate better preservation of the original luminance structure, with a value of zero corresponding to perfect agreement.

To complement this pixel-wise error measure, the Structural Similarity Index (SSIM) was computed to evaluate the preservation of local structural information, including luminance, contrast, and spatial organization:

$$\text{SSIM} = \text{SSIM}(Y_T, V). \quad (12)$$

Unlike NMSE, which measures absolute reconstruction error, SSIM emphasizes perceptually meaningful structural consistency between the original grayscale image and the luminance of the colorized representation. An SSIM value approaching one indicates that the colorization preserves the anatomical structure with minimal perceptual distortion.

2) *Order preservation*: To evaluate whether the proposed colorization preserves the relative ordering of vessel intensities, an order-preservation analysis was performed within the vessel region  $\Omega_v$ . Rather than comparing absolute intensity values, this analysis examines whether the sign of the intensity difference between neighboring vessel voxels remains unchanged after colorization. Preserving this ordering is important because local intensity relationships contribute to the visual interpretation of vessel continuity and branching.

For each pair of adjacent vessel voxels  $(x_i, x_{i+1})$ , the direction of the intensity difference in the original grayscale image and the luminance component of the colorized image was computed as

$$\begin{aligned} d_{\text{ref}}(i) &= \text{sign}(V(x_i) - V(x_{i+1})), \\ d_{\text{test}}(i) &= \text{sign}(Y_T(x_i) - Y_T(x_{i+1})), \end{aligned} \quad (13)$$

where  $V(x)$  denotes the normalized grayscale intensity and  $Y_T(x)$  represents the luminance of the colorized representation. Only voxel pairs with non-zero reference differences were considered, i.e.,  $d_{\text{ref}}(i) \neq 0$ , to avoid ambiguous comparisons between equal-intensity voxels.

Order Preservation Ratio (OPR) measures the ratio of voxel pairs whose ordering is maintained after colorization:

$$\text{OPR} = \frac{1}{|I|} \sum_{i \in I} \mathbf{1}[d_{\text{ref}}(i) = d_{\text{test}}(i)]. \quad (14)$$

A value of **OPR** = 1 indicates perfect preservation of local intensity ordering, whereas smaller values indicate increasing disruption of the original ordering.

To quantify the extent of ordering errors, the inversion rate was also computed as the proportion of voxel pairs whose relative ordering is completely reversed:

$$\text{Inversion} = \frac{1}{|I|} \sum_{i \in I} \mathbf{1}[d_{\text{ref}}(i) = -d_{\text{test}}(i)]. \quad (15)$$

A low inversion rate indicates that the colorization introduces few ordering reversals, thereby preserving the local intensity relationships that characterize vascular anatomy.

3) *Perceptual separability*: was evaluated to quantify how clearly the colorized vessel regions can be distinguished from their immediate local background. For this purpose, the RGB representation  $T$  was first converted into the perceptually motivated CIELAB color space:  $(L, a, b) = \text{Lab}(T)$ . Let  $\Omega_v$  denote the vessel region and let  $\Omega_r$  denote the surrounding ring region used as the local background. The mean CIELAB values of the ring region were used as the local background reference. For each vessel voxel  $x \in \Omega_v$ , perceptual difference from surrounding background was computed as

$$\Delta E(x) = \sqrt{(L - \mu_r(L))^2 + (a - \mu_r(a))^2 + (b - \mu_r(b))^2}. \quad (16)$$

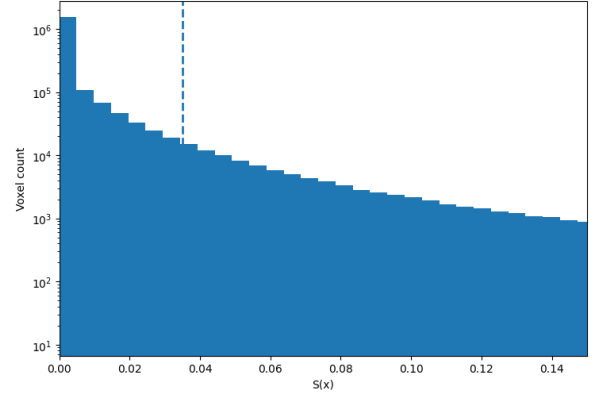


Fig. 2. Frangi vesselness  $S(x)$  distribution inside the central mask.

Average vessel-to-background perceptual separation was obtained by averaging  $\Delta E(x)$  over all voxels in the vessel region:

$$\Delta E_{\text{mean}} = \frac{1}{|\Omega_v|} \sum_{x \in \Omega_v} \Delta E(x). \quad (17)$$

Discriminability was quantified using a  $d'$ -style measure. This metric compares the separation between the vessel and ring distributions relative to their pooled variability:

$$\Delta E d' = \frac{|\mu_v - \mu_r|}{\sqrt{0.5(\sigma_v^2 + \sigma_r^2) + 10^{-8}}}. \quad (18)$$

Here,  $\mu_v$  and  $\sigma_v$  denote the mean and standard deviation of the vessel-region perceptual values, while  $\mu_r$  and  $\sigma_r$  denote the corresponding statistics for the local ring background. A larger  $\Delta E_{\text{mean}}$  or  $\Delta E d'$  indicates stronger perceptual separation between vessels and adjacent tissue.

#### H. Statistical Analysis

A two-sided paired  $t$ -test was utilized to assess changes in mean values under the assumption of approximate normality, and the Wilcoxon signed-rank test was used as a non-parametric alternative. Statistical significance was computed at  $p < 0.05$ .

### IV. RESULT

#### A. Validation of Vessel Mask

The statistical threshold  $\tau = \mu_c + 0.4\sigma_c$  successfully separates low-response background voxels from high-confidence vessel regions, as illustrated in Fig. 2. The dashed vertical line indicates the statistical threshold  $\tau = \mu_c + 0.4\sigma_c$ .

Fig. 3 shows an overlay of vesselness values  $M_{\text{central}}$  and the extracted vessel mask  $M_{\text{vessel}}$ , where the distribution of vesselness values for the extracted mask  $M_{\text{vessel}}$  is clearly shifted toward higher responses compared to the central mask.

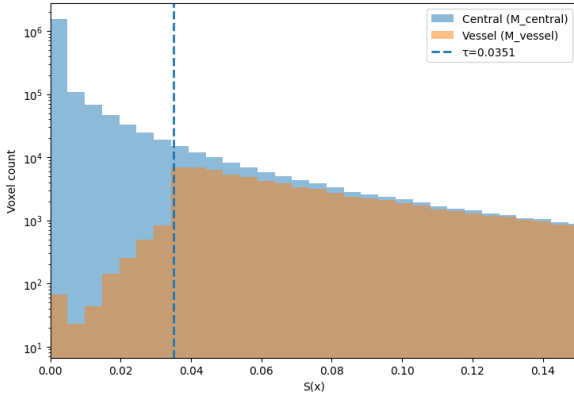


Fig. 3. Central vs. vessel distributions of  $S(x)$ .

### B. Single-Image Quantitative Analysis

The grayscale (baseline), full-volume cosine mapping, and vessel-restricted cosine mapping were evaluated. Table 1 summarizes the quantitative metrics for a single image analysis.

The grayscale image shows  $NMSE = 0.000000$ ,  $SSIM = 1.000000$ ,  $OPR = 0.998405$ , and zero inversion. It also gives baseline perceptual separability, with  $L_{d'} = 0.749028$ , chromatic discriminability = 0.732236, and  $\Delta E_{d'} = 0.224801$ .

A quantifiable structural distortion, with  $NMSE = 0.018082$  and  $SSIM = 0.847326$ , is produced with the use of full-volume cosine mapping. Although global intensity ordering stays mostly preserved ( $OPR = 0.997503$ ), a non-negligible level of inversion rate 0.015662 is introduced. Regarding perceptual characteristics, full cosine mapping lowers chromatic discriminability (Chroma  $d' = 0.601273$ ) and overall perceptual separation ( $\Delta E_{mean} = 8.909666$ ), while only marginally increasing  $\Delta E_{d'} = 0.236465$ . Vessel-restricted cosine mapping preserves anatomical structure with near-perfect fidelity ( $NMSE = 0.000001$ ,  $SSIM = 0.999993$ ), matching the grayscale baseline in both  $OPR$  (0.998405) and inversion (0.000000). Vessel-restricted mapping improves chromatic discriminability (Chroma  $d' = 0.795021$ ) relative to both grayscale and full mapping, while maintaining comparable luminance discriminability ( $L_{d'} = 0.748340$ ) and local contrast (0.083362). It slightly increases perceptual separation ( $\Delta E_{mean} = 10.190250$ ) and achieves  $\Delta E_{d'} = 0.232772$  a comparable result to full mapping but without associated distortions.

### C. Twenty-Images Quantitative Analysis

To assess consistency, the pipeline was applied to all 20 MRA volumes, and the same metrics were used to assess the grayscale, full-volume cosine, and vessel-restricted cosine mappings. Table 2. Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, and Fig. 10 show that the trends observed in the single-image analysis were consistent across the 20-volume subset. The ring is the local background around vessels, used to evaluate how well vessels stand out from nearby tissue.

### D. Statistical Analysis of Quantitative Metrics

For vessel versus grayscale, the difference was highly significant (t-test  $p < 10^{-6}$ , Wilcoxon  $p = 2 \times 10^{-6}$ ). Similarly, vessel versus full-image cosine mapping showed significant improvement (t-test  $p = 10^{-6}$ , Wilcoxon  $p = 2.7 \times 10^{-5}$ ) in perceptual discriminability. No statistically significant difference was noticed between vessel-restricted mapping and grayscale (t-test  $p = 0.0527$ , Wilcoxon  $p = 0.0679$ ). In contrast, vessel-restricted mapping showed a significant improvement over full-image cosine mapping ( $p < 10^{-6}$ ) in OPR. Vessel-restricted mapping exhibited lower inversion compared to full-image cosine mapping, with marginal statistical significance (t-test  $p = 0.0422$ , Wilcoxon  $p = 0.0679$ ).

### E. Qualitative Visualization

To complement the quantitative and statistical analyses, qualitative comparisons of grayscale, full-image cosine, and vessel-restricted cosine representations are shown in Fig. 11. Fig. 11 presents three views: axial slice visualization, maximum intensity projection (MIP), and a zoomed region highlighting local vascular structures.

## V. DISCUSSION

Fig. 2 indicates that vascular structures occupy a sparse and distinct region of the intensity space, while Fig. 3 indicates that the extracted mask lines up with anatomical vessels. Fig. 4 indicates that the vessel mask captures bright vascular structures while suppressing surrounding tissue responses. As shown in Table I, the vessel restricted cosine method preserves the original information. Whereas full-image cosine mapping introduces major structural distortions, as evidenced by lower  $SSIM$  scores and higher  $NMSE$  values. Global intensity ordering remains fairly strong, but inversion artifacts reveal how periodic color modulation disrupts local relationships. Vessel-restricted mapping, however, maintains nearly perfect structural fidelity and minimizes these inversion issues. By confining the transformation to vessels only, it avoids the distortions seen in the full-image approach. Vessel-restricted cosine mapping increases chromatic discriminability without alteration to luminance-based structure.

Table II and Figs. 5–10 show that the patterns revealed in the single case generalize across the multiple images. Vessel-restricted mapping closely matches the grayscale baseline in structural measures while providing improved chromatic separability, while full-image cosine mapping consistently shows degraded structural metrics and reduced perceptual clarity. The lower inversion in vessel-restricted representations, compared to full cosine mapping, further emphasizes the stability of the localized approach. In particular, Figs. 5-6 confirm structure preservation, Figs. 7-8 indicate stable intensity ordering with little inversion, and Figs. 9-10 demonstrate vessel-background perceptual distinction and local contrast behavior.

The statistical analysis shows that substantial improvements in perceptual discriminability ( $\Delta E_{d'}$ ) were noted when comparing vessel-restricted mapping to both grayscale and full-image cosine representations. At the same time, the lack of

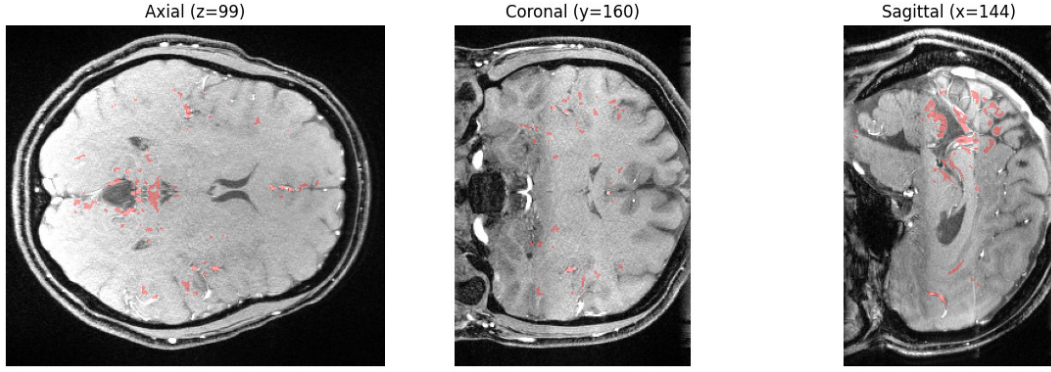


Fig. 4. Extracted vessels overlay.

TABLE I  
QUANTITATIVE COMPARISON OF GRAYSCALE, FULL-VOLUME COSINE MAPPING, AND VESSEL-RESTRICTED COSINE MAPPING.

| Method | NMSE     | SSIM     | OPR      | Inv.     | $L_{d'}$ | $\text{Chroma}_{d'}$ | $\Delta E_{\text{mean}}$ | $\Delta E_{d'}$ | Local C. |
|--------|----------|----------|----------|----------|----------|----------------------|--------------------------|-----------------|----------|
| Gray   | 0.000000 | 1.000000 | 0.998405 | 0.000000 | 0.749028 | 0.732236             | 10.146879                | 0.224801        | 0.083883 |
| Full   | 0.018082 | 0.847326 | 0.997503 | 0.015662 | 0.750136 | 0.601273             | 8.909666                 | 0.236465        | 0.072914 |
| Vessel | 0.000001 | 0.999993 | 0.998405 | 0.000000 | 0.748340 | 0.795021             | 10.190250                | 0.232772        | 0.083362 |

TABLE II  
QUANTITATIVE COMPARISON ACROSS METHODS (MEAN  $\pm$  STANDARD DEVIATION).

| Method | NMSE                | SSIM                | OPR                 | Inv.                | $L_{d'}$            | $\text{Chr}_{d'}$   | $\Delta E_m$         | $\Delta E_{d'}$     | Loc. C.             |
|--------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|----------------------|---------------------|---------------------|
| Gray   | 0.0000 $\pm$ 0.000  | 1.0000 $\pm$ 0.000  | 0.9971 $\pm$ 0.002  | 0.0000 $\pm$ 0.000  | 0.7512 $\pm$ 0.061  | 0.7365 $\pm$ 0.0596 | 10.9066 $\pm$ 0.6462 | 0.1945 $\pm$ 0.0728 | 0.0901 $\pm$ 0.0085 |
| Full   | 0.0176 $\pm$ 0.0017 | 0.8500 $\pm$ 0.0181 | 0.9956 $\pm$ 0.0024 | 0.0030 $\pm$ 0.0060 | 0.7515 $\pm$ 0.0609 | 0.6476 $\pm$ 0.0629 | 9.5393 $\pm$ 0.5529  | 0.2065 $\pm$ 0.0733 | 0.0781 $\pm$ 0.0073 |
| Vessel | 0.0000 $\pm$ 0.0000 | 1.0000 $\pm$ 0.0000 | 0.9971 $\pm$ 0.0016 | 0.0000 $\pm$ 0.0000 | 0.7504 $\pm$ 0.0604 | 0.8306 $\pm$ 0.0847 | 10.9527 $\pm$ 0.6471 | 0.2018 $\pm$ 0.0714 | 0.0896 $\pm$ 0.0085 |

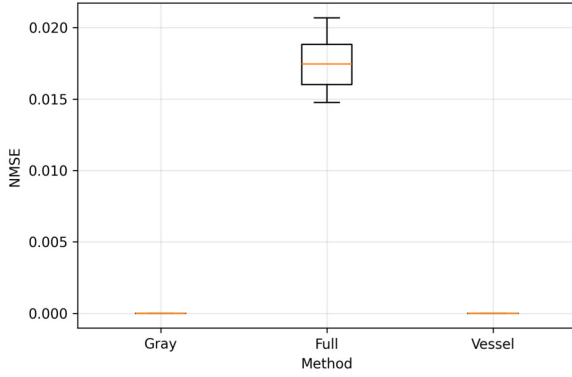


Fig. 5. Boxplot comparison of NMSE.

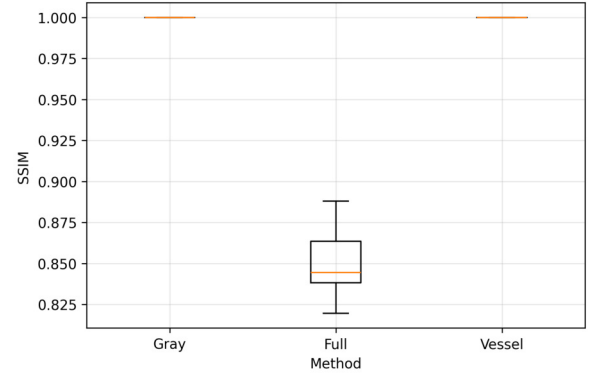


Fig. 6. Boxplot comparison of SSIM.

significant differences in object preservation between vessel-restricted mapping and grayscale suggests that the vessel cosine mapping maintains structural reliability while enhancing perceptual discriminability.

From Fig. 11, the grayscale appearance of surrounding tissue is preserved as in the case of vessel-restricted cosine mapping while confining chromatic variations to vessel regions. Whereas global color distortions and intensity inconsistencies are introduced with the use of full-image cosine mapping, which affects both vascular and non-vascular regions. The result from Fig. 11 shows that vessel-restricted cosine mapping

provides localized perceptual enhancement while preserving anatomical fidelity.

This study is limited by the use of 20 healthy MRA volumes from a 61-volume corpus. The cosine parameters were fixed, standard sequential or diverging colormaps were not compared, and luminance analysis used Rec. 709 convention.

## VI. CONCLUSION

The effectiveness of cosine mapping is highly dependent on its spatial application. Full-image cosine mapping introduced measurable structural distortion, with reduced structural sim-

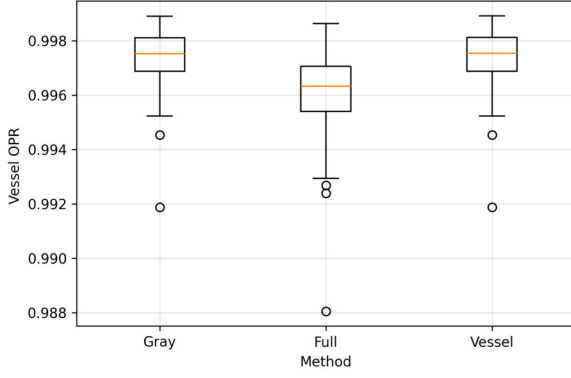


Fig. 7. Boxplot comparison of OPR.

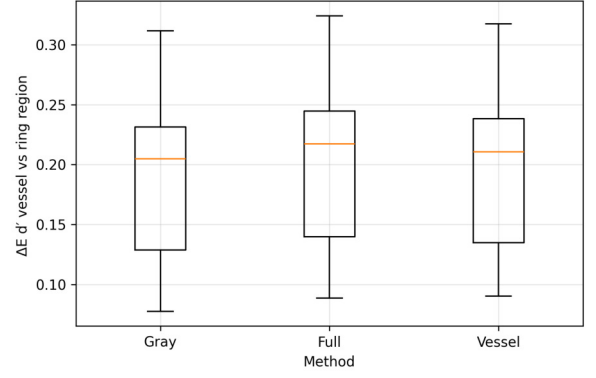


Fig. 9. Boxplot comparison of vessel ring metric.

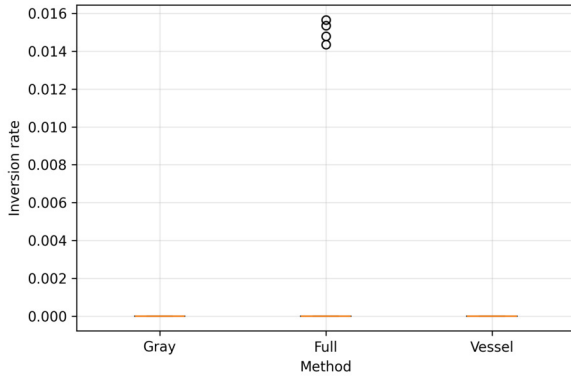


Fig. 8. Boxplot comparison of inversion rate.

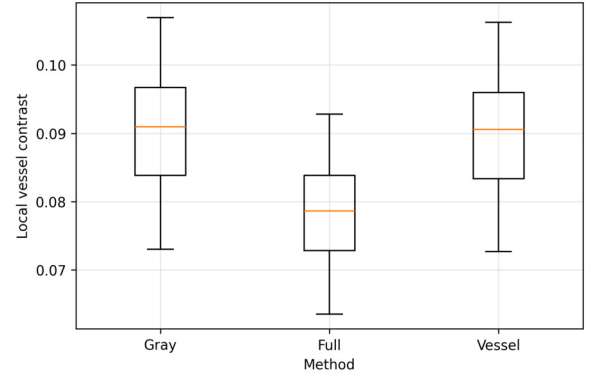


Fig. 10. Boxplot comparison of local vessel contrast.

ilarity  $\approx 0.85$  and increased reconstruction error of  $\approx 0.018$ , along with non-zero inversion artifacts of  $\approx 0.003$ . Although global intensity ordering remained relatively high with a value of  $\approx 0.9950.8$ , full-image mapping reduced chromatic discriminability of  $\approx 0.65$ . In contrast, vessel-restricted cosine mapping preserved anatomical structure with a near-perfect fidelity with an error of  $\approx 0$  and a structural similarity value of  $\approx 1.0$ , and eliminated inversion artifacts of  $0.000$ , while maintaining object preservation comparable to grayscale of  $\approx 0.997$ . Importantly, it improved chromatic discriminability  $\approx 0.83$  relative to both grayscale  $\approx 0.74$  and full-image cosine mapping and achieved slightly improved perceptual separation of  $\Delta E_{\text{mean}} of \approx 10.95$ . These improvements were statistically significant,  $\Delta E_d$  showing strong significance when comparing vessel-restricted mapping to grayscale and full-image cosine (paired t-test  $p < 10^{-6}$ , Wilcoxon  $p < 10^{-5}$ ).

Qualitative analysis suggested that vessel-restricted mapping provides localized chromatic enhancement of vascular structures while preserving the grayscale appearance of surrounding tissue, avoiding the global distortions observed in full-image cosine mapping.

The proposed vessel-restricted approach therefore shows a favorable balance between enhancement and fidelity, suggesting its potential as a visualization-oriented enhancement

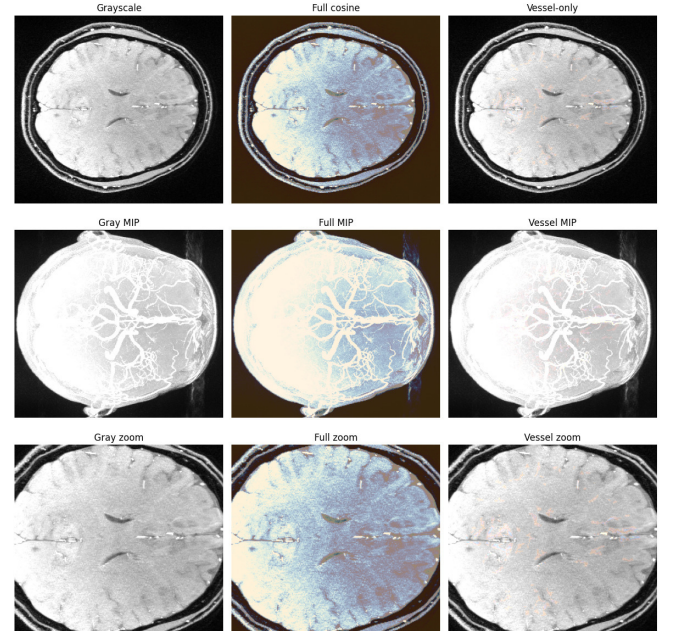


Fig. 11. Qualitative comparison of grayscale, full-image cosine, and vessel-restricted cosine representations across slice, MIP, and zoomed views.

method for MRA. A promising direction for future work is to explore optimization strategies to select cosine mapping parameters for maximizing perceptual discriminability while preserving structural constraints. Additionally, utilizing global perceptual metrics and task-based clinical evaluation will be important to assess the broader impact of the method on diagnostic performance.

#### DATA AND CODE AVAILABILITY

Data underlying the results presented in this paper are publicly available at <https://www.nitrc.org/projects/icbmmra/>. The source code used for preprocessing, cosine colorization, vessel-restricted masking, and metric evaluation is available upon reasonable request from the corresponding author.

#### ACKNOWLEDGMENT

The study is funded by the Lithuanian Research Council (LMT) under project grant number S-PD-25–40. During the preparation of this work, the authors used OpenAI tools to help refine and improve the code for the experiments. They also used Grammarly to review and improve the manuscripts' grammar. The authors take full responsibility for the content of this publication.

#### REFERENCES

- [1] Gadhave DG, Sugandhi VV, Jha SK, Nangare SN, Gupta G, Singh SK, Dua K, Cho H, Hansbro PM, Paudel KR. Neurodegenerative disorders: Mechanisms of degeneration and therapeutic approaches with their clinical relevance. *Ageing research reviews*. 2024 Aug 1;99:102357, <https://doi.org/10.1016/j.arr.2024.102357>.
- [2] Rezaei M, Mohammadikhavah S, Faraji H, Ardalani R, Rezaei M, Shirazinodeh A. Early diagnosis of Alzheimer's disease based on brain morphological changes: A comprehensive approach combining voxel-based morphometry and deep learning. *NeuroImage: Reports*. 2026 Mar 1;6(1):100315, <https://doi.org/10.1016/j.ynirp.2025.100315>.
- [3] BAlsbati SM, Al-Gburi RM, Mustafa A, AHMED SK, Issa AM, Al-Naimi TM, AlSaad R, Elhenidy AM. Advances in Deep Learning for Multimodal Brain Imaging: A Comprehensive Survey. *Neuroscience Informatics*. 2025 Dec 19:100252, <https://doi.org/10.1016/j.neuri.2025.100252>.
- [4] Karthikeyan S, Muthu Kumar B, Kiran ML, Srivatsan K. Computational Brain Imaging Framework for Neurological Mapping and Disorder Classification Using Multimodal Image Processing. *International Journal of Computational Intelligence Systems*. 2025 May 19;18(1):121, <https://doi.org/10.1007/s44196-025-00852-1>.
- [5] Ricci M, Cimini A, Chiaravalloti A, Filippi L, Schillaci O. Positron emission tomography (PET) and neuroimaging in the personalized approach to neurodegenerative causes of dementia. *International Journal of Molecular Sciences*. 2020 Oct 11;21(20):7481, <https://doi.org/10.3390/ijms21207481>.
- [6] Ansari AS, Mohammadi MS, Cattani C, Tassaddiq A. An advanced multimodal image fusion model for accurate detection of Alzheimer's disease using MRI and PET. *Frontiers in Medical Technology*. 2025 Dec 2;7:1699821, <https://doi.org/10.3389/fmedt.2025.1699821>.
- [7] Bhattacharya S, Prusty S, Pande SP, Gulhane M, Lavate SH, Rakesh N, Veerasamy S. Integration of multimodal imaging data with machine learning for improved diagnosis and prognosis in neuroimaging. *Frontiers in Human Neuroscience*. 2025 Mar 21;19:1552178, <https://doi.org/10.3389/fnhum.2025.1552178>.
- [8] Zubair M, Hussain M, Al-Bashrawi MA, Bendechache M, Owais M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. *Computer Methods and Programs in Biomedicine*. 2025 Sep 9:109014, <https://doi.org/10.1016/j.cmpb.2025.109014>.
- [9] Calhoun VD, Sui J. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*. 2016 May 1;1(3):230-44, <https://doi.org/10.1016/j.bpsc.2015.12.005>.
- [10] Mathur AN, Khattar A, Sharma O. 2D to 3D medical image colorization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 2021 (pp. 2847-2856), <https://doi.org/10.1109/WACV48630.2021.00289>.
- [11] Odusami M, Damasevicius R, Milieskaite-Belousoviene E, Maskeliunas R. Multimodal neuroimaging fusion for Alzheimer's disease: an image colorization approach with mobile vision transformer. *International Journal of Imaging Systems and Technology*. 2024 Sep;34(5):e23158, <https://doi.org/10.1002/ima.23158>.
- [12] Pei E. *Ambient Environment Effect on Color Perception*. Rochester Institute of Technology; 2025.
- [13] Costanzo L, Failla G, Aluigi L, Baroncelli TA, Bua C, De Marchi S, Diaco E, Di Paola F, Di Pino FL, Mannello F, Martinelli O. Operative Procedures for Ultrasound Assessment of Extracranial Artery Disease: A Narrative Review by the Italian Society for Vascular Investigation (SIDV). *Journal of Clinical Medicine*. 2025 Oct 6;14(19):7050, <https://doi.org/10.3390/jcm14197050>.
- [14] Alfadeel H. Impact of CT Stroke Window Settings on Acute Stroke Detection. *Scientific Journal of University of Benghazi*. 2025 Jun 29;38(1):193-214, <https://doi.org/10.37376/sjuob.v38i1.7328>.
- [15] D'Angelo T, Mastrodicasa D, Lanzafame LR, Yel I, Koch V, Gruenewald LD, Sharma SP, Ascenti V, Micari A, Blandino A, Vogl TJ. Optimization of window settings for coronary artery assessment using spectral CT-derived virtual monoenergetic imaging. *La radiologia medica*. 2024 Jul;129(7):999-1007, <https://doi.org/10.1007/s11547-024-01835-6>.
- [16] Mohammad NI. Hierarchical Spatial Algorithms for High-Resolution Image Quantization and Feature Extraction. *arXiv preprint arXiv:2510.08449*. 2025 Oct 9, <https://doi.org/10.1049/iet-ipr.2019.0921>.
- [17] Singh N, Bhandari AK. Image contrast enhancement with brightness preservation using an optimal gamma and logarithmic approach. *IET Image Process* 14 (4): 794–805 [Internet]. 2019
- [18] Sarkar K, Halder TK, Mandal A. Adaptive power-law and cdf-based geometric transformation for low-contrast image enhancement. *Multimedia Tools and Applications*. 2021 Feb;80(4):6329-53, <https://doi.org/10.1007/s11042-020-10004-6>.
- [19] Khan MA, AlGhamdi MA. An intelligent and fast system for detection of grape diseases in RGB, grayscale, YCbCr, HSV, and L\*a\*b\* color spaces. *Multimedia Tools and Applications*. 2024 May;83(17):50381-99, <https://doi.org/10.1007/s11042-023-17446-8>.
- [20] Mookiah S, Parasuraman K, Kumar Chandar S. Color image segmentation based on improved sine cosine optimization algorithm. *Soft Computing*. 2022 Dec;26(23):13193-203, <https://doi.org/10.1007/s00500-022-07133-5>.
- [21] Reinhold JC, Dewey BE, Carass A, Prince JL. Evaluating the impact of intensity normalization on MR image synthesis. In *Proceedings of SPIE—the International Society for Optical Engineering*. 2019 Mar (Vol. 10949, p. 109493H), <https://doi.org/10.1117/12.2513089>.
- [22] Johnson R. A review of three-dimensional medical image visualization. *Health Data Science*. 2022(1):83-101, <https://doi.org/10.34133/2022/9840519>.
- [23] Longo A, Morscher S, Najafabadi JM, Jüstel D, Zakian C, Ntzachristos V. Assessment of a Hessian-based Frangi vesselness filter in optoacoustic imaging. *Photoacoustics*. 2020 Dec 1;20:100200, <https://doi.org/10.1016/j.pacs.2020.100200>.
- [24] Sheethal. M. S, Amudha P. Enhanced Brain Tumor Detection via Multi-scale Frangi Gaussian Matrix and Efficient Convolutional Networks on MRI. *J. Inf. Hiding Multim. Signal Process*. 2024 Sep;15(4):236-70.
- [25] Qin B, Jin M, Hao D, Lv Y, Liu Q, Zhu Y, Ding S, Zhao J, Fei B. Accurate vessel extraction via tensor completion of background layer in X-ray coronary angiograms. *Pattern recognition*. 2019 Mar 1;87:38-54, <https://doi.org/10.1016/j.patcog.2018.09.015>.
- [26] Zhang M, Wang J, Cao X, Xu X, Zhou J, Chen H. An integrated global and local thresholding method for segmenting blood vessels in angiography. *Heliyon*. 2024 Nov 30;10(22), <https://doi.org/10.1016/j.heliyon.2024.e38579>.
- [27] Recommendation IT. *Colour conversion from Recommendation ITU-R BT. 709 to Recommendation ITU-R BT*. 2020. Geneva: ITU-R. 2015.

## Data collection, aggregation, and analysis in the Internet of Everything environment – a case study

Andrzej Gołabek  
0000-0001-7973-9534  
InfoSoftware Polska Sp. z o. o.  
ul. Przemysłowa 5A  
38-457 Szczepańcowa, Poland  
Email:  
agolabek@infosoftware.pl

Piotr Sowa,  
Mateusz Kozik  
InfoSoftware Polska Sp. z o. o.  
ul. Przemysłowa 5A  
38-457 Szczepańcowa, Poland  
Email:  
psowa@infosoftware.pl,  
mkozik@infosoftware.pl

Andrzej Paszkiewicz,  
Marek Bolanowski  
0000-0001-7573-3856  
0000-0003-4645-967X  
Rzeszow University of  
Technology,  
Al. Powstańców Warszawy 12,  
35-029 Rzeszów, Poland  
Email:  
apaszkiewicz@infosoftware.pl,  
mbolanowski@infosoftware.pl

**Abstract**— This article presents a practical approach to topics related to the Internet of Everything (IoE) environment, specifically describing the main mechanisms and challenges associated with the collection, aggregation, and analysis of data from multiple sources. In distributed IoE environments, data is transmitted using various communication protocols and at different frequencies. Both network packets and logs transmitted by IoE devices are specific to a given device type and often vary depending on the manufacturer. In large-scale IoE environments, the very large volume of data transmitted in various formats requires the implementation of effective mechanisms to aggregate it in near real time, especially when considering its analysis in the context of cybersecurity. This article presents a general operational diagram of a system enabling the analysis of data from a distributed IoE environment, incorporating mechanisms for data acquisition from multiple sources, caching, scaling, and multi-stage data analysis that machine learning and artificial intelligence support.

### I. INTRODUCTION

INTERNET of Everything (IoE) environments represent an extension of the traditional concept of the Internet of Things (IoT), encompassing not only end devices but also users, business processes, applications, analytical systems and network infrastructure [1, 2, 3]. In an IoE environment, data is not generated solely by individual sensors or telemetry devices, but by an entire ecosystem of interconnected elements. In practice, IoE environments generate vast amounts of data from many heterogeneous sources [4, 5]. This data includes, amongst other things, system logs, telemetry data, network packets, diagnostic messages, security events, alerts from IDS/IPS systems, syslog data, NetFlow, and information from industrial devices, OT networks and cloud infrastructure. This data is transmitted using various mechanisms and

communication protocols, such as MQTT, Modbus RTU, Syslog, HTTP/HTTPS, UDP, RTP, REST API, CoAP and NetFlow, amongst others. Furthermore, many manufacturers of the same types of devices use different formats for transmitted data and messages, apply varying levels of log detail, and employ different data transmission frequencies, which necessitates the use of additional mechanisms to normalize this data prior to further analysis [6, 7].

From the perspective of customers and organizations implementing IoE solutions, it is essential to ensure an effective, consistent and, above all, secure method of collecting and aggregating data from numerous dispersed locations. Once this data has been appropriately filtered, normalized, classified and, in many cases, enriched with additional context, it should be accessible via a central analytics system. The collected and aggregated data form the basis for further analysis, particularly about the correct operation and performance of the entire IoE environment, as well as cybersecurity. Therefore, in IoE environments, it is essential to implement scalable systems that enable the collection, aggregation and analysis of data in near real time. Furthermore, the systems implemented must be equipped with mechanisms to mitigate the risk of data loss and must ensure business continuity even under increased load [8, 9].

Despite the growing number of solutions designed to monitor and secure IoE environments, there remains significant scope for research and implementation regarding their practical application in small and medium-sized enterprises and local government bodies. The SIEM, SOAR, IDS/IPS systems and stream analysis platforms available on the market are designed for large organizations that possess technical and financial resources, as well as specialized staff. For small organizations, implementing such solutions is often impossible

due to their high costs and the lack of access to specialized staff with experience in solutions integrating multiple heterogeneous data sources [10, 11]. At the same time, these organizations use various IT systems (often distributed), IoT/IoE devices, cloud applications and security tools from different suppliers in their day-to-day operations, which creates a need for flexible analytical solutions tailored to the organization's actual capabilities. A particular challenge remains the acquisition, normalization and correlation of data generated across the entire IoE environment implemented within the organization. The diversity of communication protocols, data formats and transmission frequencies means that small and medium-sized enterprises require simplified, modular and scalable systems supporting near-real-time data analysis [12].

The innovative aspect of the approach presented in this article lies in the development of a system concept for the IoE environment, combining mechanisms for data aggregation, queuing, normalization, stream analysis, anomaly detection and artificial intelligence support in cybersecurity processes [13, 14]. A key element of the proposed approach's innovation is the use of a Feedback Loop mechanism, which creates feedback between the analytical layer and the data pre-processing layer. This mechanism allows for the automatic updating of correlation rules, the modification of stream analysis parameters, and the adjustment of feature vectors used in anomaly detection and cybersecurity analysis. As a result, the system can respond not only to detected incidents but also to changes in traffic characteristics, the location of data sources, load levels, and threat profiles. This approach transforms the IoE environment from a passive data aggregation and monitoring system into an adaptive architecture capable of self-adjusting to changing operational conditions.

From the perspective of small and medium-sized enterprises, modular architecture allows the system's functional scope to be tailored to the organization's needs and capabilities, without the need to build costly infrastructure. This approach enables the gradual implementation of monitoring, event correlation, anomaly detection and automated incident response functions, which significantly lowers the implementation barrier and enhances their level of cybersecurity [15].

## II. CUSTOMER REQUIREMENTS, LOG CHARACTERISTICS, AND SECURITY MECHANISMS IN IOE ENVIRONMENTS

The main requirements that customers place on IoE environments are, above all, high system availability and scalability, reliable data transmission, the ability to centrally monitor infrastructure, and ease of integration with existing IT solutions. Scalability is of particular importance, as the number of devices, sensors, applications and data-generating systems is growing rapidly as organizations expand. Another key requirement is the ability to analyze data in near real time, particularly in environments with limited hardware resources, which is what small and medium-sized enterprises can afford. It is important to customers that such solutions enable the rapid detection of failures, anomalies, configuration errors and potential security incidents. Automation is crucial, and

more specifically the implementation of artificial intelligence and machine learning mechanisms that will support customers through the automatic classification of events, the detection of unusual communication patterns, and the reduction of false alarms.

The information logged by IoE devices is highly varied. It includes telemetry data, system events, diagnostic information, network traffic data, error messages, security alerts and information on the status of end devices. The log formats, transmission frequency and level of detail depend not only on the type of device, but also on the manufacturers. Some IoE devices transmit data periodically, whilst others, such as sensors, only send logs after a change in status or following a specific event. Therefore, one of the key stages in processing data from such a diverse environment is its normalization and standardization. Standardizing the data enables its subsequent analysis and correlation. This process includes, amongst other things, parsing messages, assigning timestamps, classifying event types, and adding context such as the device's location, type, IP address or severity level.

In IoE environments under development, data security is of paramount importance alongside effective data collection and aggregation mechanisms, and customers are increasingly drawing attention to this. The basic security measures that must be implemented in such environments include data transmission encryption, network segmentation, access control, multi-factor authentication (2FA/MFA), device authentication, and monitoring of the integrity of transmitted information. The principles of Zero Trust architecture are also frequently applied, in which no element of the IoE infrastructure can be considered trusted by default. Every device, user and application should be verified before being granted access to system resources. SIEM and SOAR systems also play a significant role, enabling centralized log collection, event correlation, and the identification of and automatic response to potential threats. The combination of effective and secure mechanisms for log collection and aggregation, correlation rules, AI-powered analysis, and SOAR playbooks create an IoE environment with a high level of cybersecurity.

## III. CASE STUDY AND KEY COMPONENTS OF THE IOE ENVIRONMENT

A sample Internet of Everything (IoE) environment designed to collect, aggregate and analyze data from multiple distributed data sources is presented as a case study. A diagram of the proposed environment is shown in Fig. 1. This diagram includes customer requirements relating to scalability, high availability, fault tolerance, near-real-time data analysis and security-related elements.

The first element shown in the diagram is the data sources, which include IoE devices, telemetry systems, network infrastructure and security components such as EDR, IDS/IPS and UTM. The data is transmitted using various communication protocols, such as MQTT, Syslog, HTTPS, NetFlow and Modbus, to so-called data collectors and data producers.

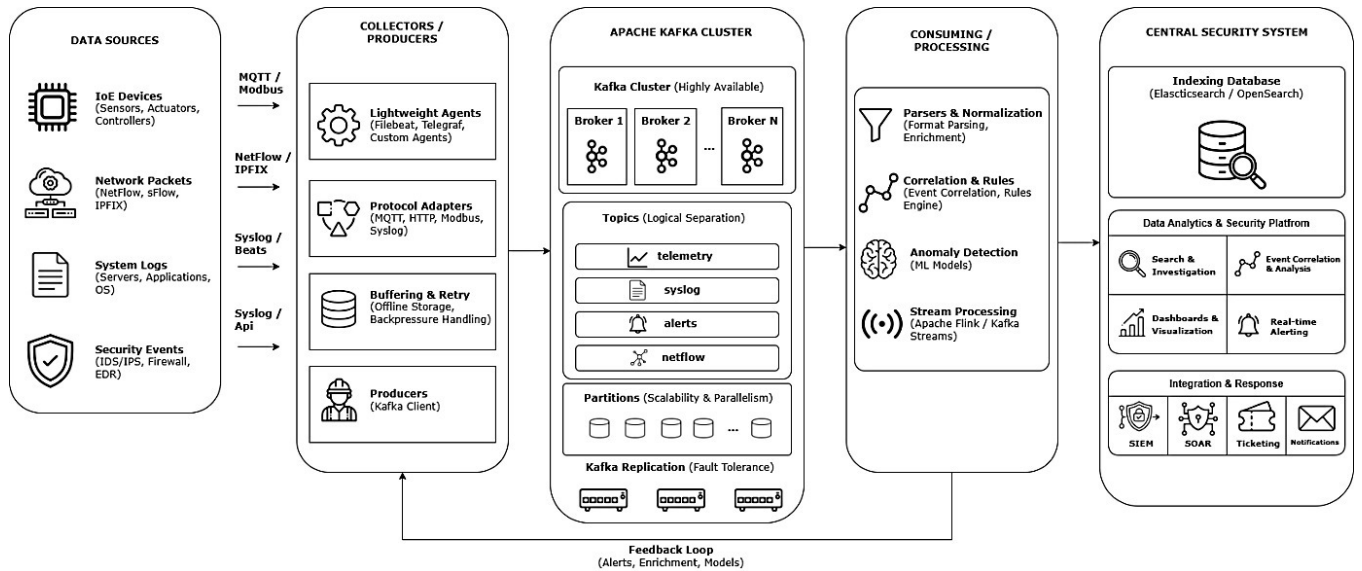


Fig. 1. Diagram of an example environment for collecting, aggregating, and analyzing data from multiple IoE sources.

The collector and producer layer is primarily responsible for collecting data, performing initial filtering, and normalizing it. This component of the environment includes, amongst other things, log-collecting agents, communication protocol adapters that enable the integration of devices from different manufacturers, and data producers such as the Kafka Client.

Another component of this environment is the Apache Kafka cluster, which acts as a distributed queuing and data transport system. The data transmitted is logically divided into topics relating to telemetry, syslog data, alerts and Net-Flow. This layer utilizes partitioning and replication mechanisms that ensure high performance, fault tolerance and the ability to process very large volumes of data in parallel in near real time.

The fourth element shown in the diagram is the data processing and analysis layer. It is primarily responsible for parsing incoming logs, event correlation, stream analysis and anomaly detection. Machine learning algorithms have been implemented within this layer to support the initial identification of security incidents and unusual communication patterns.

The final layer of this environment is a central security system that utilizes an indexed database, enabling rapid searching, analysis and correlation of events. This system integrates SIEM and SOAR solutions, as well as a ticketing system, to support the SOC team in responding to security incidents. A key feature of this system is the ability to map events to the tactics and techniques described in the MITRE ATT&CK matrix. This allows detected incidents to be classified not only as individual alerts, but as elements of a potential attack chain. The system also utilizes Indicators of Compromise (IoC), such as suspicious IP addresses, domains, file extensions, unusual process IDs, malware signatures or characteristic patterns of network communication, to automatically enrich logs and compare current events against databases of known

threats. In addition, the central security system collects data on software versions and device types and compares this with

CVE (Common Vulnerabilities and Exposures) databases to identify potential attack vectors. This layer also incorporates artificial intelligence mechanisms in the form of machine learning models that analyze data from a behavioral perspective (UEBA) and an AI assistant that supports the SOC team in analyzing incidents, issues and detected anomalies. The implemented artificial intelligence significantly streamlines the response process to detected anomalies and security incidents and provides suggestions regarding correlation rule settings and SOAR scenarios.

A key component of this environment is the Feedback Loop mechanism, which is responsible for transmitting feedback between the analytical layer and the data pre-processing layer (collectors). This mechanism provides feedback that enables the system to dynamically adapt its operation based on the results of ongoing analysis of logs and security events. It allows for the automatic updating of correlation rules both in time and space, the modification of stream analysis parameters, and the adjustment of feature vectors for more accurate cybersecurity analysis. Adaptation over time means that the system can respond to the changing dynamics of the IoE environment, such as an increase in the number of messages, the emergence of anomalies, periodic overloads, changes in communication patterns, or the occurrence of a security incident. From this perspective, correlation rules, detection thresholds and stream analysis parameters are not static, but can be adjusted based on the current state of the system and observed events. Spatial adaptation, on the other hand, refers to the ability to vary the method of data analysis and processing depending on the source's location, network segment, device type, the role of a given component in the architecture, and the criticality level of the monitored area. Different rules may apply to control systems, different ones to stream sources, and yet

others to application data sources or telemetry sensors. As a result, the IoE environment is not treated as a uniform communication space, but as a collection of areas with varying levels of risk, differing operational significance, and distinct requirements regarding security and performance.

#### IV. SUMMARY

The rapid growth of Internet of Everything environments is leading to a steady increase in the number of devices, applications and services that generate data continuously and in a distributed manner. In practice, this means there is a need to build modern, scalable systems capable of efficiently collecting, aggregating and analyzing large volumes of data from heterogeneous sources in near real time, whilst operating with severely limited hardware resources.

The Feedback Loop mechanism used is one of the key elements of the innovative architecture proposed in this article. This mechanism enables the transfer of analysis results from the analytical layer to the layer of collectors and data pre-processing components, allowing for the dynamic modification of correlation rules, stream analysis parameters, and feature sets used in detection processes. In practice, this means the system can be continuously adapted to the current threat profile, infrastructure load, and the significance of individual data sources.

The proposed solution is not limited to the passive collection and analysis of information but creates an adaptive mechanism that supports the automation of cybersecurity, the reduction of false alarms, and a more efficient use of IoE infrastructure resources. The introduction of feedback between the analysis and data acquisition layers increases the resilience and scalability of the entire environment. It also allows for the further development of the system towards self-learning architecture, in which decisions regarding prioritization, filtering and correlation of data are gradually optimized based on observed events, anomalies and security incidents.

Of particular importance is the ability to adapt the system's operation in both temporal and spatial dimensions. The temporal dimension refers to the system's ability to respond to changes occurring over time, such as an increase in traffic volume, the occurrence of anomalies, changes in communication patterns, infrastructure overload, or the emergence of a new attack pattern. This allows detection and correlation rules to be updated not only periodically, but also in response to current events observed in the environment. The spatial dimension, on the other hand, concerns the ability to differentiate analysis depending on the location of data within the IoE architecture. The system can apply different detection

thresholds, correlation rules and levels of analytical detail for different network segments, physical locations, device classes and areas with varying levels of criticality. This avoids treating the entire infrastructure uniformly and allows analytical resources to be concentrated where the risk or operational significance of the data is greatest.

#### REFERENCES

- [1] E. Siow, T. Tiropanis, W. Hall, "Analytics for the Internet of Things: A Survey", *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, 2018. DOI: 10.1145/3204947.
- [2] M. Ge, H. Bangui, B. Buhnova, "Big Data for Internet of Things: A Survey", *Future Generation Computer Systems*, vol. 87, pp. 601–614, 2018. DOI: 10.1016/j.future.2018.04.053.
- [3] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, A. Hussain, "Big Data and IoT-Based Applications in Smart Environments: A Systematic Review", *Computer Science Review*, vol. 39, 100318, 2021. DOI: 10.1016/j.cosrev.2020.100318.
- [4] D. Fawzy, S. M. Moussa, N. L. Badr, "The Internet of Things and Architectures of Big Data Analytics: Challenges of Intersection at Different Domains", *IEEE Access*, vol. 10, pp. 4969–4992, 2022. DOI: 10.1109/ACCESS.2022.3140409.
- [5] A. Naghib, N. Jafari Navimipour, M. Hosseinzadeh, A. Sharifi, "A Comprehensive and Systematic Literature Review on the Big Data Management Techniques in the Internet of Things", *Wireless Networks*, vol. 29, pp. 1085–1144, 2023. DOI: 10.1007/s11276-022-03177-5.
- [6] I. Ali, S. Sabir, E. Khan, "Privacy-Preserving Data Aggregation in Resource-Constrained Sensor Nodes in Internet of Things: A Review", *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 41–50, 2018. DOI: 10.1016/j.fcij.2017.11.
- [7] J. Hou, L. Qu, W. Shi, "A Survey on Internet of Things Security from Data Perspectives", *Computer Networks*, vol. 148, pp. 295–306, 2019. DOI: 10.1016/j.comnet.2018.11.026.
- [8] G. Calderon, G. del Campo, E. Saavedra, A. Santamaria, "Monitoring Framework for the Performance Evaluation of an IoT Platform with Elasticsearch and Apache Kafka", *Information Systems Frontiers*, vol. 26, pp. 2373–2389, 2024. DOI: 10.1007/s10796-023-10409-2.
- [9] T. P. Raptis, A. Passarella, "A Survey on Networked Data Streaming with Apache Kafka", *IEEE Access*, vol. 11, pp. 85333–85350, 2023. DOI: 10.1109/ACCESS.2023.3303810.
- [10] A. Emer, M. Unterhofer, E. Rauch, "A Cybersecurity Assessment Model for Small and Medium-Sized Enterprises", *IEEE Engineering Management Review*, vol. 49, no. 2, pp. 98–109, 2021. DOI: 10.1109/EMR.2021.3078077.
- [11] N. Rawindaran, A. Jayal, E. Prakash, "Machine Learning Cybersecurity Adoption in Small and Medium Enterprises in Developed Countries", *Computers*, vol. 10, no. 11, 150, 2021. DOI:10.3390/computers10110150.
- [12] M. Ulrich, F. Frank, A. S. Klug, "An AI Adoption Model for SMEs: A Conceptual Framework", *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 702–708, 2021. DOI: 10.1016/j.ifacol.2021.08.082.
- [13] A. Chatterjee, B. S. Ahmed, "IoT Anomaly Detection Methods and Applications: A Survey", *Internet of Things*, vol. 19, 100568, 2022. DOI: 10.1016/j.iot.2022.100568.
- [14] K. DeMedeiros, A. Hendawi, M. Alvarez, "A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks", *Sensors*, vol. 23, no. 3, 1352, 2023. DOI: 10.3390/s23031352.
- [15] J. Kinyua, L. Awuah, "AI/ML in Security Orchestration, Automation and Response: Future Research Directions", *Intelligent Automation & Soft Computing*, vol. 28, no. 2, pp. 527–545, 2021. DOI: 10.32604/iasc.2021.016240.

# Parsing Hierarchical Document Structure with Graph Neural Networks

Thomas Reiser, Veronika Vasileva, Volker Riediger, Jan Jürjens

0009-0007-5452-7800

0009-0006-2024-4653

0000-0003-1461-7442

0000-0002-8938-0470

University of Koblenz

Email: {treiser, vvasileva, riediger, juerjens}@uni-koblenz.de

Alexander Esser, Petra Steiner

0000-0002-5974-2637

Bundesinstitut für Berufsbildung,

Email: {alexander.esser, steiner}@bibb.de

**Abstract**—Document hierarchy parsing is used to identify hierarchical relationships between layout elements in documents, such as headings, paragraphs, and list items. This information makes documents that are not born-digital easier to navigate for human readers and can support automated processing of these documents. Similar to existing approaches, we insert a sequence of nodes that represent layout elements into a layout tree to incrementally reconstruct the original document hierarchy. Different to existing research, our approach uses a GraphSAGE-based feature extraction network and a deep neural network based scorer to predict parent relationships between layout element pairs using text and layout features. Our data set contains 50 historical German documents and we apply k-fold cross validation with five splits to validate the model's ability to generalize from a small data set to unseen data. The model was able to obtain an average F1-score of over 91% on the selected data set.

## I. INTRODUCTION

**D**OCUMENT HIERARCHY PARSING describes the identification of hierarchical relationships between layout elements like headings and paragraphs in documents [1]. The document hierarchy thus organizes elements like sections and subsections in full documents. An example of such hierarchical relationships in a single-paged document is shown in Figure 1.

A document hierarchy that is known to machines and researchers can support at many tasks, e.g., table of content reconstruction, better document navigation for human readers, or better chunking for retrieval augmented generation with LLMs with a limited number of tokens.

In this paper, we aim to identify the document hierarchy in historical documents given text and layout features that are obtained by a self-trained document layout analysis (DLA) model based on DocLayout-YOLO [2] and machine recognized text in the detected bounding boxes using the Tesseract OCR engine.

Our data set is based on an excerpt of a set of historical vocational education and training (VET) regulations maintained in an occupational archive maintained by the Federal Institute for Vocational Education and Training (BIBB). Because legal documents in particular are frequently organized in sections, subsections and paragraphs, these document corpora represent

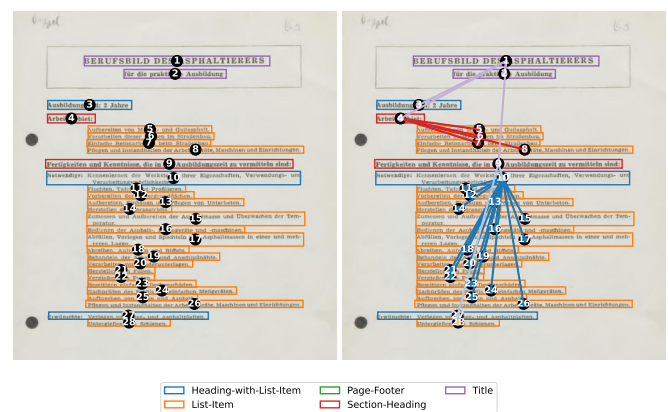


Fig. 1. Example of hierarchical structure in a document layout graph. Outgoing edges are the parent of the element with the ingoing edge. Ingoing edges are visualized by a dot in the center of the bounding box.

a good foundation of hierarchy parsing methods as they often also follow deeper document hierarchies.

A common approach for parsing document hierarchy is the insertion of a node sequence into a layout tree, where pairs of nodes are scored to decide how likely a node in the tree is a parent node of the new inserted node [3, 4]. Our approach also follows this idea but uses GraphSAGE [5] to pass information of neighboring nodes to the parent candidate node in order to help the scorer decide which node is the parent of the next inserted node.

Due to the underlying graph-structure of layout trees, we aim to investigate the effectiveness of graph neural networks in parsing document hierarchy by answering the following research questions:

RQ1 How can graph-based neighborhood aggregation be used to predict document hierarchy?

RQ2 Which data set characteristics impact the prediction quality of document hierarchy parsing the most?

Compared to other studies, this does not only utilize node features of candidate parents and siblings, but also their respective parent nodes through a single feature extraction network based on aggregation of neighboring node features

with GraphSAGE rather than a combination of different scores. Another advantage coming from using graph neural networks like GraphSAGE is that nodes up to the number of additional GraphSAGE layers away from the candidate parent can be incorporated without adding further feature extractors for, e.g., grandparents and grandchildren of the candidate parent.

In the example of Figure 1, the order in which unconnected nodes are inserted into the generated tree is given by the reading order, that is visualized by the number in the center of the nodes on the left side of the plot. For each node in the sequence, a new edge is added from the selected parent node to the newly added node, until all nodes are added and a full layout hierarchy tree as seen in the right side of the plot is built.

To decide where nodes are added, a GraphSAGE feature extraction network is used to generate features for nodes that have already been added to the graph and a scorer that scores pairs of candidate parents and new nodes until the tree is completed like seen on the right side of the plot.

The remaining paper is structured as follows: Section II discusses related work and the current status of methods that address the problem of document hierarchy parsing. Section III presents the proposed methodology and the used model architecture. In Section IV, the results are presented and discussed using F1-score and describing common errors found in the predictions of our model. Finally, Section V summarizes our findings, followed by a critical discussion of our methods and an outlook for future research.

Our contributions are a new model architecture to extract features from document layout graphs and a new scorer architecture for scoring pairs of layout elements to decide where new nodes are added.

## II. RELATED WORK

Document understanding is a broad research field that revolves around methods like document layout analysis [6, 7, 8] which describes methods that identify and classify layout elements in document images, named entity recognition [9, 10], aiming to extract information about, e.g., names and places in texts, optical character recognition (OCR) [11] to extract text from images, and form understanding [12, 13], used for question answering in forms like receipts or questionnaires.

Hierarchical structure analysis on the other hand focuses more on the logical and hierarchical structure of documents that have been previously processed with, for example, document layout analysis and OCR and frequently builds on top of these results. In [3], the Hierarchy Extraction from Long Document (HELD) model was introduced, where nodes are sequentially inserted into a tree structure. The decision where a new node is inserted is done by a put-or-skip module which computes the probability of the new node being inserted on a parent node of the graph under construction.

In [4], the Detect-Order-Construct approach is introduced that combines document layout analysis with reading order reconstruction followed by the construction of the document hierarchy tree. Similar to [3], they restrict candidate parents



Fig. 2. Features of the layout elements

to the rightmost path of the constructed graph. But instead of using a put-or-skip module, they utilize a Construct module which organizes a sequence of identified section headings. A relation prediction head scores how likely a new node is either a parent or a sibling of a candidate node in the table of contents tree. These scores are then used for a multi-class classification problem with softmax cross-entropy to select the parent node with highest score.

DocHieNet [1] is a large document hierarchy parsing data set that includes a variety of layouts rather than being reduced to a small specific domain like, e.g., PubLayNet [6] which focuses only on medical papers. To test their data set, they adapted the HELD approach with the put-or-skip module to compute probabilities where a newly added node must be inserted into the hierarchy tree.

Document hierarchy is also relevant for form understanding tasks. DocStruct [14] for example uses a multimodal method for form understanding to recognize the organizational structure of forms. Similar to parent-child relations in common document hierarchy parsing tasks, they define key-value pairs in forms. They use text, layout and image features to then compute a score that represents the probability of two layout elements having a key-value relationship.

The usage of graph neural networks (GNNs) for document understanding has also already been investigated. Paragraph2Graph [15] is a framework that uses GNNs for different document understanding tasks which is able to encode features at line-level but also encode recognized information in a graph using task-specific layers while not relying on language-specific encoders.

## III. METHOD

In our approach, we encode page number, reading order, text, bounding box information and class features of layout elements and incrementally build a hierarchical document structure tree as seen in Figure 1 by inserting a sequence of nodes in their respective reading order. Figure 2 shows how the positional and textual features are represented. The green box around the text illustrates the bounding box that was previously detected by a layout analysis model. The  $x$  and  $y$  values represent the top left position of the input bounding box, whereas the  $h$  and  $w$  values represent height and width of the bounding box.

For each node, a set of candidate parents is selected, i.e., the rightmost path of the document tree at the stage where the next node is inserted. Our method assumes there is one layout element that can be considered to be the root element

for the document and all other layout elements have only one parent element. Similar to other works [3, 4], we only allow insertion into the rightmost path of the existing tree starting with a root node which is assumed to be the title element of the document. This ensures that nodes cannot be inserted into a section of the document that has already been completed in the tree.

The process can be summarized in four steps:

- 1) Select candidate parents for new node, i.e., the rightmost path
- 2) Extract features from candidate nodes and new node
- 3) Score pairs of candidate parent nodes and new node
- 4) Select parent with highest scoring pair

Due to size limitations in our data set, which consists of only 50 documents, we use k-fold cross validation with five splits and shuffling. This means, our data set is shuffled to reorder the documents in it and then divided into five equally-sized splits. Afterwards, our model is trained five times where four of these splits are used for training and the remaining one for testing. In the end, model performance is assessed using F1-score on predicted and true edges.

#### A. Data Set

For the training and evaluation of our method, a data set was manually annotated. The data set consists of 50 documents with two to 15 pages containing between 23 and 218 layout elements. Each document is structured in a graph structure with a set of nodes and edges. Nodes are represented by layout elements and edges which define the relationships between these layout elements. The left plot in Figure 3 shows the distribution of documents by the number of pages in our data set, the middle plot shows the distribution of layout elements by the number of pages, and the right plot the frequency of the deepest hierarchy level in each document. Here, the deepest hierarchy is defined as the length of the longest path from the root element to any leaf element.

The layout elements consist of page number, reading order position, bounding box, their text, and a class which was assigned by a document layout analysis model. Edges on the other hand are directed and each outgoing edge represents a parent relationship to the target node of that edge.

The bounding boxes are defined by horizontal and vertical position of their top left corner as well as width and height. These are normalized to a range between 0 and 1, representing relative position in the original document image. The same normalization was applied to page number and reading order position. Layout classes are one of 13 classes assigned during document layout analysis on given documents. These classes include section headings, paragraphs, list items, page headers, and several other classes. However, our approach is not limited to these classes and designed to allow different classes in future research. Nevertheless, the selection of layout classes will impact the document structure result because section headings, for example, are likely to be followed by a new paragraph. Floating text layout elements that are on the page images, but not part of the main text, like page headers, page

numbers, etc. are ignored for the hierarchy construction task and filtered out before training and inference, using the classes that were previously assigned by the DLA model.

A visualization of used features can be found in Figure 2.

#### B. Model Architecture

Our model consists of three components: A BERT encoder, based on XLM-RoBERTa [16] for text feature generation, a graph neural network based on GraphSAGE [5] for feature generation on the existing graph where new nodes are inserted, and a scorer that decides which node is most likely the parent node of a newly inserted node.

The original text is extracted from the document images using Tesseract OCR [17]. The XLM-RoBERTa encoder was selected because of its ability to handle German language due to its multilingual support.

The positional features are created using the bounding box coordinates from a document layout analysis (DLA) model that predicts the positioning on the original image. The layout classes are also predicted by the same DLA model but treated as a separate feature.

For the text feature encoder, we keep the pretrained weights. The feature extraction network and the pair scorer are described in the following sections.

1) *Feature Extraction Network*: The feature extraction network is visualized in Figure 4.

In the first step of the feature extraction network, textual and positional features are both processed by a linear layer, followed by a layer norm and a ReLU activation function. The class features are processed by an embedding layer and a layer norm. Embedding layers are particularly useful for categorical information to learn similarity between classes. For the given documents for example, paragraphs and list items are more similar than paragraphs and section headings because paragraphs are less likely to start a new section than section headings. For the non-categorical features, a ReLU activation function is applied before the resulting feature vectors are stacked in the fusion layer.

After feature fusion where the extracted features are concatenated, messages from neighboring nodes are passed through a GraphSAGE layer with an LSTM aggregator.

GraphSAGE is a special kind of graph neural network (GNN) that is able to learn graph structures inductively [5]. This enables these networks to learn the structure of unseen graphs, making them particularly useful in tasks like document hierarchy parsing where new graphs are built incrementally. In these networks, features of neighboring nodes are aggregated by an aggregation function that is learned during model training.

For this paper, an LSTM aggregator has been selected over mean or other convolutional aggregators because it can better learn relationships between different neighboring nodes, which can be particularly helpful for document hierarchy trees where relationships between sibling nodes can also exist. However, for GraphSAGE, the order of nodes is not known because nodes in a graph are not ordered by default. Therefore, a

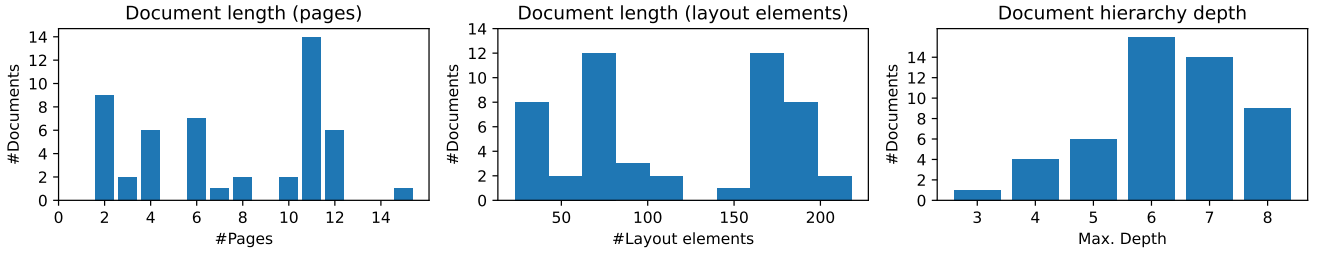


Fig. 3. Quantitative description of the used data set: Document length in number of pages, number of layout elements (nodes) and maximum hierarchy depths.

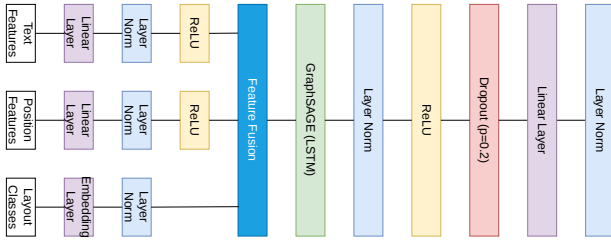


Fig. 4. Architecture of the proposed feature extraction network.

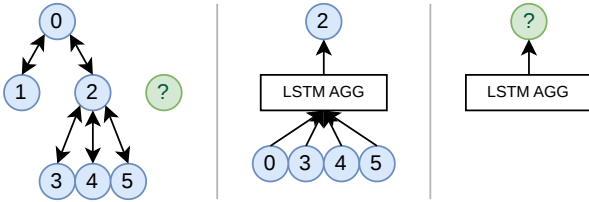


Fig. 5. Message passing on an example graph (left) for node 2 (middle) and node ? (right) in the GraphSAGE layer.

random sequence of nodes is used by the LSTM aggregator. The GraphSAGE layer passes information from the candidate's parent and children nodes to the candidate parent, as shown in Figure 5. In the given example, the disconnected green node has no messages to pass as it has no ingoing edges.

A single GraphSAGE layer was selected in particular to pass information from all children that were already added to a node and should help the pair scorer decide where the next node is inserted. In the rightmost tree of Figure 7, messages from nodes 0 and 3 would be passed to node 2 when  $score(2, ?)$  is calculated.

This layer is then followed by another Layer Norm and ReLU layer before a dropout layer that is used to avoid overfitting. The dropout layer deactivates 20% of all neurons in the network randomly for the upcoming computation to avoid reliance on a few single nodes when making predictions.

In the last step, another linear layer is added, followed by a layer norm. Because cross entropy loss was used which applies softmax activation anyways, no other activation is applied afterwards.

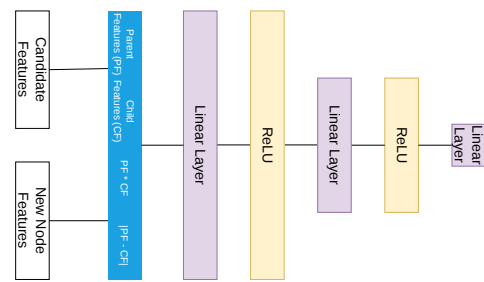


Fig. 6. Architecture of the pair scorer. The candidate features and new node features are both generated from the feature extraction network.

2) *Pair Scorer*: The pair scorer evaluates pairs of nodes to decide which node in the already existing layout tree is the parent of the next node in the insertion sequence. It takes the candidate parent's features, extracted by the feature extraction network, and the features of the next node. The node with highest score is then selected to be the parent of the new node.

Architecturally, the pair scorer is a deep neural network with three linear layers followed by a ReLU activation function each. The input of this network is the concatenation of the extracted parent and child features, as well as their product and the absolute value of their difference. This combination was selected to also include relative information between parent and child features.

Therefore, the first linear layer has four times the number of the feature extraction network's output dimension neurons. In the second layer, this is cut in half to two times the number of output neurons. In the final layer, this dimension is further reduced to a single scalar value.

The overall network architecture for the pair scorer is sketched in Figure 6.

### C. Training and Inference

To simulate the incremental construction of a graph at training time, for each graph in the annotated data set, all prefix graphs are computed. In particular, the training data set consists of all subtrees of the document trees in the annotated data set where all nodes up to a certain reading order position are added. Because this problem is trivial for prefix graphs

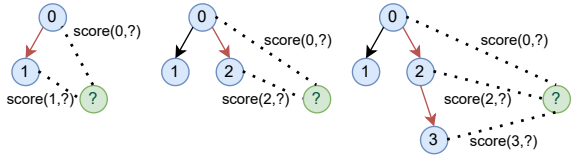


Fig. 7. Visualization of the training and prediction process. Red arrows mark the rightmost path containing the candidate parents. Blue nodes have already been assigned to the document tree. Green nodes with question mark are newly assigned to the document tree. Dashed lines mark pairs that are scored for classification.

containing only one node, those cases were removed from the training data set.

Examples of the prefix graph and new nodes that are added in each larger graph are visualized in Figure 7. During training, the GraphSAGE network learns which features to generate, whereas the scorer learns how to combine these features to generate higher scores for the correct parent.

Although edges are directed when it comes to sorting nodes in reading order, during training and inference, they are considered as undirected. This is particularly important for the message passing in the GraphSAGE network, so children that are already connected to the parent pass their information to it when a new node is added, ensuring they impact the decision for newly added nodes.

Only nodes on the rightmost path of the already constructed layout tree are considered as candidate nodes because we assume that once a new section at the same level started, all previous sections are completed. The rightmost path restriction models this assumption. To encode features on both the existing graph and the unseen node, the new node is added to the graph without any edges. Then, the pair scorer evaluates pairs of the nodes on the rightmost path. Cross entropy loss is used during training because the model is optimizing a multi-class classification problem with only one single class being the correct one for the new node. The candidate nodes correspond to the possible classes to which the new node can be assigned. The highest-scoring pair then indicates the two nodes between which the new edge is added.

On prediction time, the layout tree starts with the first node in the insertion sequence and then adds nodes sequentially, using the parent assignment of the trained models. A new link between each of the models is added after the parent node has been selected using the pair scorer.

#### IV. RESULTS

We applied k-fold cross validation with five splits. Each split contained 40 training documents and 10 test documents that the model has not seen during training.

In each test run, the number of edges that were both in ground truth and predictions (true positives), the number of edges that were in the predictions but not in the ground truth (false positives), the number of edges that were in the ground truth but not in the predictions (false negatives) and the number of edges that were neither in the predictions nor the ground

TABLE I  
MACRO (PER-TREE AVERAGE) AND MICRO (PER-EDGE) F1-SCORES OF EACH SPLIT SPLIT (IN %).

| Split            | 1            | 2            | 3            | 4            | 5            | Mean         |
|------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| <b>Micro</b>     |              |              |              |              |              |              |
| <b>GraphSAGE</b> | 92.13        | <b>94.34</b> | <b>90.21</b> | <b>87.95</b> | <b>92.82</b> | <b>91.49</b> |
| <b>Linear</b>    | 91.78        | 91.78        | 88.43        | 87.66        | 88.45        | 89.62        |
| <b>Ablation</b>  | <b>93.09</b> | 89.43        | 86.98        | 82.52        | 88.05        | 88.05        |
| <b>Macro</b>     |              |              |              |              |              |              |
| <b>GraphSAGE</b> | 94.26        | <b>91.86</b> | <b>90.85</b> | 87.63        | <b>92.02</b> | <b>91.32</b> |
| <b>Linear</b>    | 93.88        | 89.33        | 88.15        | <b>88.11</b> | 85.29        | 88.95        |
| <b>Ablation</b>  | <b>94.94</b> | 88.08        | 88.56        | 82.16        | 87.95        | 87.95        |

truth (true negatives) is counted to then compute F1-score on each of the test splits.

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Precision depends on the number of true positives, recall on the number of false negatives. In our case, each false positive also implies a false negative since every edge has only one chance to be detected by the model, so that precision = recall in our particular approach. Therefore, we report results using the F1-score, the harmonic mean of precision and recall because it provides a standard metric that reflects both precision and recall which are equal in our context.

For better comparability, the experiment was conducted in 2 additional setups: We replaced the GraphSAGE layer by a linear layer of same dimensionality (Linear setup) and removed it entirely (Ablation setup). The results are shown in Table I. Bold printed numbers indicate the best result in each split and on average. The micro results refer to the evaluation on all edges in the split, the macro results refer to the average per-tree F1-score in the split.

While each of these other two architectures was able to outperform the GraphSAGE variant in singular splits, the GraphSAGE architecture still has the highest average micro and macro F1-score by roughly 1.8% for micro and 2.3% on macro, showing its better ability to generalize from the training data. It remains to be investigated how this difference changes on larger and more heterogeneous data sets.

However, since measuring only the recognition of all edges in a split does not necessarily give enough attention to the quality of individual document trees (edges that are added early are arguably easier to add), true and false positives and negatives are also reported. Both evaluations are combined because a single short document with low F1-score can heavily decrease the average of the F1-scores, even if all other trees were predicted successfully. The results of average F1-scores per tree in each split are also shown in Table I.

The average F1-score per tree in each test split is 91.32%. When counting all edges in each test split, the average F1-

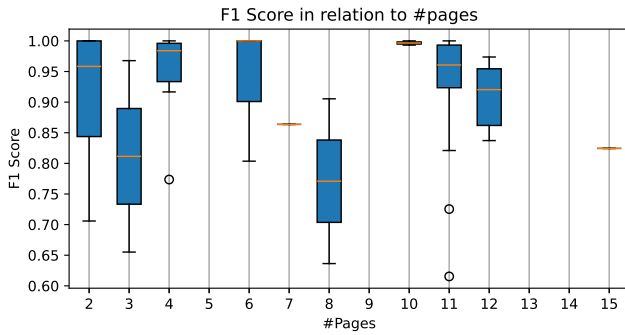


Fig. 8. Distribution of F1-scores for documents with different number of pages.

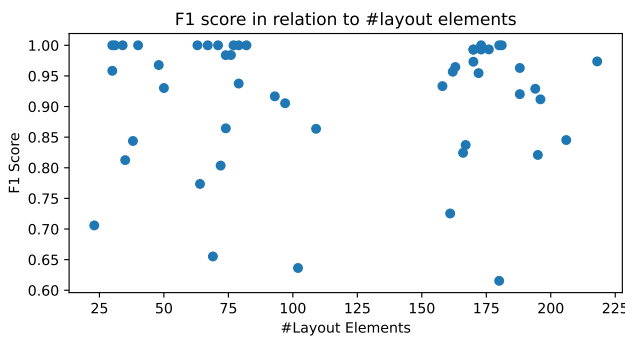


Fig. 9. Relationship between number of layout elements and F1-score.

score is 91.49%. This indicates that the high performance is not due to a few large trees with many correctly predicted edges, but that the model generally reconstructs the layout hierarchy trees well. Also, the model successfully learned to build layout trees on unseen data with overall high reliability. However, there were overall high fluctuations in the results, as seen for example in Figure 9 where each tree-wise F1-score is plotted on the y-axis and the document length measured in number of layout element is shown on the x-axis.

This plot also shows that the overall prediction quality is not directly influenced by document length, as also seen in the boxplot in Figure 8 which shows the distribution of F1-scores in relation to the number of pages in a document. While there are strong fluctuations in the average F1-score between different document lengths, these are expected due to the limited data set size.

There are only few predictions with low F1-score for both long and short documents and there seems to be no obvious relation between number of nodes and F1-score or number of pages and F1-score. Across all test splits, all but six of the 50 predicted trees achieved an F1-score of more than 80%, as seen in Figure 9. Even the document with the most nodes achieved a high F1-score of over 95%.

The relationship between depth of the document hierarchy and the F1-score is plotted in Figure 10. The single document that had a maximum hierarchy depth of three had an F1-

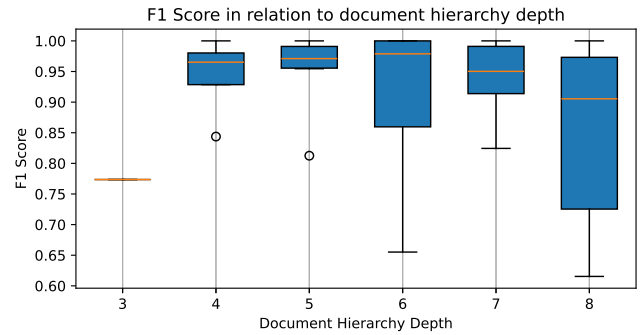


Fig. 10. F1-scores for all predictions per maximum hierarchy depth.

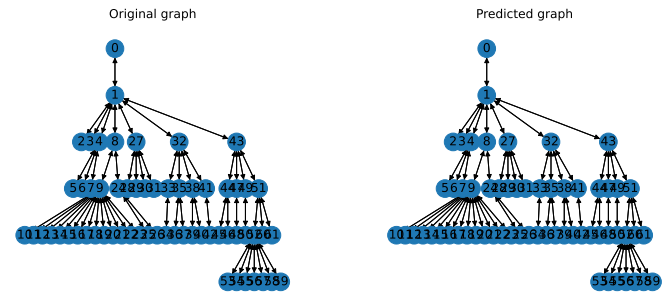


Fig. 11. Example of a tree in the ground truth compared to the predictions by the developed model (F1-score: 100%).

score of only 77%. Reason for this might be the lack of examples in the training data set and the model not being able to generalize from documents with deep hierarchies to flat ones. For documents with four to seven hierarchy levels, the average score was over 90%. While even the documents with deepest hierarchy have an average score of around 85%, there was a lot of inconsistency in the result of these trees compared to the less deep trees.

Other than the overall document length, it can be observed that deeper hierarchy causes more inconsistencies, as seen in the box plot. Because deeper hierarchies imply more possible classes for the given classification problem, F1-scores also more frequently decrease for these.

An example of a successfully predicted tree compared to its ground truth is shown in Figure 11. In this example, the basic document structure is captured, with the same nodes introducing new sections in both the ground truth and the prediction.

In example of Figure 12, the model failed to capture the actual layout tree depth. While many of the central nodes that introduce a new layer have been recognized correctly in the upper layers of the layout tree, errors like this can either propagate after a few wrong insertions or are just not fully learned by the model. This tree is also one of the challenging examples in the data set, where a paragraph is the parent node of following list items, which rarely occurs in the data set and is part of a data imbalance problem. This issue can likely be addressed by more similar examples in a larger data set.



## REFERENCES

- [1] H. Xing, C. Cheng, F. Gao, Z. Shao, Z. Yu, J. Bu, Q. Zheng, and C. Yao, "DocHieNet: A large and diverse dataset for document hierarchy parsing," 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.65 pp. 1129–1142. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.65/>
- [2] Z. Zhao, H. Kang, B. Wang, and C. He, "Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception," 2024. [Online]. Available: <https://arxiv.org/abs/2410.12628>
- [3] R.-Y. Cao, Y.-X. Cao, G.-B. Zhou, and P. Luo, "Extracting variable-depth logical document hierarchy from long documents: Method, evaluation, and application," *Journal of Computer Science and Technology*, vol. 37, no. 3, pp. 699–718, 2022. doi: 10.1007/s11390-021-1076-7. [Online]. Available: <https://doi.org/10.1007/s11390-021-1076-7>
- [4] J. Wang, K. Hu, Z. Zhong, L. Sun, and Q. Huo, "Detect-order-construct: A tree construction based approach for hierarchical document structure analysis," *Pattern Recognition*, vol. 156, p. 110836, 2024. doi: 10.1016/j.patcog.2024.110836. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320324005879>
- [5] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2018. [Online]. Available: <https://arxiv.org/abs/1706.02216>
- [6] X. Zhong, J. Tang, and A. Jimeno Yepes, "Publaynet: Largest dataset ever for document layout analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. doi: 10.1109/ICDAR.2019.00166 pp. 1015–1022.
- [7] B. Pfizmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, "Doclaynet: A large human-annotated dataset for document-layout segmentation," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539043. ISBN 9781450393850 p. 3743–3751. [Online]. Available: <https://doi.org/10.1145/3534678.3539043>
- [8] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A benchmark dataset for document layout analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020. doi: 10.18653/v1/2020.coling-main.82 pp. 949–960. [Online]. Available: <https://aclanthology.org/2020.coling-main.82/>
- [9] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022. doi: 10.1109/TKDE.2020.2981314
- [10] Y. Wang, H. Tong, Z. Zhu, and Y. Li, "Nested named entity recognition: A survey," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 6, Jul. 2022. doi: 10.1145/3522593. [Online]. Available: <https://doi.org/10.1145/3522593>
- [11] R. Avyodri, S. Lukas, and H. Tjahyadi, "Optical character recognition (ocr) for text recognition and its post-processing method: A literature review," in *2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, 2022. doi: 10.1109/ICTIIA54654.2022.9935961 pp. 1–6.
- [12] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," 2019. [Online]. Available: <https://arxiv.org/abs/1905.13538>
- [13] A. Abdallah, D. Eberharter, Z. Pfister, and A. Jatowt, "Transformers and language models in form understanding: A comprehensive review of scanned document analysis," 2024. [Online]. Available: <https://arxiv.org/abs/2403.04080>
- [14] Z. Wang, M. Zhan, X. Liu, and D. Liang, "DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020. doi: 10.18653/v1/2020.findings-emnlp.80 pp. 898–908. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.80/>
- [15] S. Wei and N. Xu, "Paragraph2graph: A gnn-based framework for layout paragraph analysis," 2023. [Online]. Available: <https://arxiv.org/abs/2304.11810>
- [16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. doi: 10.18653/v1/2020.acl-main.747 pp. 8440–8451.
- [17] R. Smith, "An overview of the tesseract OCR engine," in *International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, 2007. doi: 10.1109/ICDAR.2007.4376991 pp. 629–633.
- [18] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," 2022. [Online]. Available: <https://arxiv.org/abs/2204.08387>
- [19] Text Encoding Initiative Consortium, "TEI P5: Guidelines for Electronic Text Encoding and Interchange," <https://tei-c.org/guidelines/>, 2023, version P5.

# Still Waiting for the Shock: AI's Limited Impact on Early-Career Vacancies, Skills and Tasks

Stefan Speckesser  
0000-0002-2442-7194  
Brighton Business School  
University of Brighton  
Mithras House  
Lewes Road  
Brighton, BN2 4AT  
United Kingdom  
Email: S.S.Speckesser@brighton.ac.uk

Lei Xu  
0000-0002-6590-0435  
Loughborough Business School  
Sir Richard Morris Building  
Loughborough University  
Epinal Way  
Loughborough, LE11 3TU  
United Kingdom  
Email: l.xu2@lboro.ac.uk

**Abstract**—This analysis of 620,000 UK apprenticeship vacancies suggests that AI is not the primary cause of declining opportunities, which instead stems from structural labour market weaknesses. The research utilized DistilRoBERTa to map digital skills from unstructured text and Google Gemini 1.5 Flash to dynamically generate an occupation-specific task taxonomy. Econometric results show no significant impact on overall vacancy volumes or technical task density. However, a modest increase in digital skill requirements was found for higher-level roles, suggesting AI complements advanced qualifications rather than displacing them. While intermediate roles have collapsed, findings attribute this to policy shifts like the Apprenticeship Levy rather than automation.

**Index Terms**—Apprenticeships, digital skills, artificial intelligence

## I. INTRODUCTION

OVER the past seven years, the United Kingdom youth labour market has experienced profound structural volatility, exacerbated by the scarring effects of the COVID-19 pandemic, the cost-of-living crisis, and the rapid integration of generative artificial intelligence. As of early 2026, approximately 957,000 young people aged 16 to 24 are classified as not in education, employment, or training, driving the youth unemployment rate to 16.0 percent (Office for National Statistics, 2026). Concurrently, the apprenticeship system has undergone a radical transformation, systematically squeezing out the traditional youth entry-level market.

Moving beyond speculative discourse, this study utilizes a novel dataset of 619,629 government apprenticeship vacancies to empirically track the evolution of entry-level job roles before and after artificial intelligence integration. The methodology integrates advanced natural language processing and large language model techniques to map occupations to the European Skills, Competences, Qualifications and Occupations framework. This allows for a granular analysis of task composition and the identification of digital skill intensity within specific roles.

The empirical analysis is structured around three central hypotheses designed to capture different dimensions of arti-

ficial intelligence induced labor disruption. The first hypothesis posits that the disruption caused by artificial intelligence impacts the total volume of job adverts, specifically measured by a reduced growth rate of vacancies in apprenticeship standards requiring extensive digital skills. The null hypothesis posits no such macroeconomic impact on vacancy volumes. The second hypothesis evaluates skill composition, suggesting that artificial intelligence disruption increases the digital skills expected in roles, with this shift expected to be more significant for higher apprenticeships. The null hypothesis posits no impact on skill composition. The third hypothesis examines task shifts, suggesting that artificial intelligence leads to task reductions or significant shifts within standards defined by extensive digital skills. The null hypothesis asserts no significant change in the task composition of these roles.

By employing a rigorous difference-in-differences design with placebo tests, the research ensures the observed trends are statistically robust, providing a precise technical evaluation of how artificial intelligence is reshaping the demand for early-career skills. Preliminary findings indicate that the null hypotheses for overall vacancy volumes and technical task shifts cannot be rejected. However, digital skill requirements have measurably increased for higher-level educational roles.

## II. LITERATURE REVIEW AND POLICY BACKGROUND

### A. The Youth Labour Market and Apprenticeships

Historically, entry-level roles served as a vital training ground where young workers performed routine cognitive tasks to acquire domain expertise. Following the introduction of the Apprenticeship Levy in 2017, the volume of apprenticeship starts contracted sharply, stabilizing at volumes approximately 30 percent below their pre-Levy peaks (Department for Education, 2026). The defining trend within this contraction is the collapse of intermediate Level 2 apprenticeships, which historically served as the primary onboarding route into basic administrative and technical roles for school leavers. These intermediate roles fell from 43 percent of all starts in the 2017 to 2018 academic year to just 18.6 percent in 2024 to 2025.

Conversely, employer demand has migrated toward higher and degree apprenticeships, which now account for nearly 40 percent of all starts. This pivot reflects a broader training paradox wherein employers increasingly utilize levy funds to upskill existing older staff rather than hiring new entry-level trainees. In the 2024 to 2025 academic year, 51.3 percent of all apprenticeship starts were undertaken by individuals aged 25 and over (Department for Education, 2026). As routine cognitive tasks become commoditized by algorithms, employers are redirecting investments toward higher-level cognitive management and specialized digital skills, leaving a widening gap at the entry level.

### *B. Task Shifts and the Expected Impact of Artificial Intelligence*

To understand the distributional consequences of artificial intelligence, economic scholarship relies on the task-based approach. Breaking from canonical models of skill-biased technological change, which viewed technology merely as factor-augmenting, the task-based framework separates workers' skills from the discrete tasks they perform [7]. In the routine-biased technological change era of the late twentieth century, computer capital acted as a direct substitute for human workers in performing codifiable, routine tasks, effectively hollowing out middle-skill administrative and manufacturing jobs while complementing high-skill abstract reasoning [3].

Generative artificial intelligence, however, disrupts this historical pattern. Exhibiting the traits of a general-purpose technology, large language models excel at pattern recognition, implicit rule learning, and contextual synthesis, allowing them to automate non-routine cognitive tasks such as ideation, legal analysis, and content creation. Consequently, generative artificial intelligence poses the greatest task-exposure risk to high-income, highly educated knowledge workers.

Despite this high theoretical exposure, empirical evidence suggests artificial intelligence currently functions primarily as a labour-augmenting tool. By automating specific sub-tasks, artificial intelligence drives skill compression, disseminating tacit expertise to novice workers and forcing a reallocation of human labour toward higher-order strategic, decision-making, and interpersonal functions [8]. Under the theoretical framework of Acemoglu and Restrepo [4], [5], the aggregate macroeconomic outcomes of this technology depend on the balance between the displacement effect of automation and the reinstatement effect, which requires the continuous generation of new, labour-intensive tasks in which humans hold a comparative advantage.

As cognitive tasks become commoditized, educational and employment pathways are migrating toward the human premium, encompassing roles requiring high-touch empathy, specialized producing industries relying on physical craft, and ethical governance requiring human-in-command oversight [10].

## III. DATA AND LLM MODELLING

### *A. Data Sources and the ESCO Digital Framework*

This study utilizes vacancy data from the Department for Education Find an apprenticeship service. The dataset captures 619,629 adverts, corresponding to 942,798 individual apprenticeship jobs, posted between August 2018 and October 2025. The analysis restricts this universe to 492,818 adverts regulated under current Apprenticeship Standards, enabling the integration of complete regulatory data regarding occupational roles and educational levels.

To systematically quantify digital skill demand within these unstructured vacancies, the methodology integrates the European Skills, Competences, Qualifications and Occupations classification (ESCO). Specifically, the study extracts the digital skills domain, which is grounded in the European Digital Competence Framework for Citizens. By utilizing this poly-hierarchical ontology comprising 25 distinct digital competences rather than an arbitrary keyword list, the methodology ensures the extraction of statistically validated concepts, minimizing hallucinatory matching and correcting for occupational biases.

### *B. LLM Modelling: Mapping ESCO and Vacancies*

To evaluate occupational exposure to artificial intelligence, unstructured vacancy texts were mapped to the ESCO framework to identify digital skill intensity. A constraint-based domain association strategy was engineered using Distil-RoBERTa, a knowledge-distilled encoder model optimizing the balance between deep linguistic comprehension and computational speed [12], [13].

The raw vacancy descriptions underwent a rigorous natural language processing pipeline involving HTML stripping, punctuation removal, and Porter stemming. To minimize false positives, the semantic anchor of each ESCO skill was restricted exclusively to its alternate, technical labels. An exhaustive cosine similarity comparison was then executed between the vector representations of every vacancy and skill signature. A stringent threshold of 0.45 was applied, converting continuous similarity scores into binary digital skill intensity metrics for econometric modelling.

### *C. Generative AI and Task Extraction Modelling*

To evaluate task-level impacts, Google Gemini 1.5 Flash was utilized to dynamically generate occupation-specific taxonomies, producing 30 distinct, three-word technical tasks for each of the 463 Apprenticeship Standards. Due to severe computational constraints and API rate limits associated with processing massive text corpora via cloud-based generative models, the methodology pivoted to local, deterministic natural language processing for the extraction phase. The final extraction algorithm utilized set-based keyword proximity matching to verify if the discrete words of each generated task appeared within the aggregated annual vacancy corpus for specific occupations.

TABLE I  
DIGITAL SKILLS FOUND IN APPRENTICESHIP VACANCY ADVERTS

| Digital Skill                                                   | Adverts | Vacancies | Mean | Std. dev. | Min  | Max  |
|-----------------------------------------------------------------|---------|-----------|------|-----------|------|------|
| Solve technical problems                                        | 491,148 | 745,334   | 0.00 | 0.06      | 0.00 | 1.00 |
| Use e-services                                                  | 491,148 | 745,334   | 0.21 | 0.41      | 0.00 | 1.00 |
| Creatively use digital technologies                             | 491,148 | 745,334   | 0.13 | 0.34      | 0.00 | 1.00 |
| Evaluate data, information and digital content                  | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 0.00 |
| Computer programming                                            | 491,148 | 745,334   | 0.00 | 0.04      | 0.00 | 1.00 |
| Identify digital competence gaps                                | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 1.00 |
| Collaborate through digital technologies                        | 491,148 | 745,334   | 0.00 | 0.01      | 0.00 | 1.00 |
| Protect personal data and privacy                               | 491,148 | 745,334   | 0.00 | 0.03      | 0.00 | 1.00 |
| Share through digital technologies                              | 491,148 | 745,334   | 0.02 | 0.13      | 0.00 | 1.00 |
| ICT safety                                                      | 491,148 | 745,334   | 0.22 | 0.41      | 0.00 | 1.00 |
| Identify technological needs                                    | 491,148 | 745,334   | 0.45 | 0.50      | 0.00 | 1.00 |
| Manage data, information and digital content                    | 491,148 | 745,334   | 0.00 | 0.01      | 0.00 | 1.00 |
| Digital communication and collaboration                         | 491,148 | 745,334   | 0.00 | 0.01      | 0.00 | 1.00 |
| Integrate and re-elaborate digital content                      | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 1.00 |
| Develop digital content                                         | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 0.00 |
| Digital data processing                                         | 491,148 | 745,334   | 0.00 | 0.01      | 0.00 | 1.00 |
| Problem-solving with digital tools                              | 491,148 | 745,334   | 0.16 | 0.37      | 0.00 | 1.00 |
| Interact through digital technologies                           | 491,148 | 745,334   | 0.09 | 0.28      | 0.00 | 1.00 |
| Intellectual property regime of digital content                 | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 1.00 |
| Manage digital identity                                         | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 0.00 |
| Protect the environment from the impact of digital technologies | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 0.00 |
| Protect health and well-being while using digital technologies  | 491,148 | 745,334   | 0.08 | 0.27      | 0.00 | 1.00 |
| Protect ICT devices                                             | 491,148 | 745,334   | 0.02 | 0.13      | 0.00 | 1.00 |
| Browse, search and filter data, information and digital content | 491,148 | 745,334   | 0.00 | 0.00      | 0.00 | 0.00 |
| Digital content creation                                        | 491,148 | 745,334   | 0.00 | 0.01      | 0.00 | 1.00 |

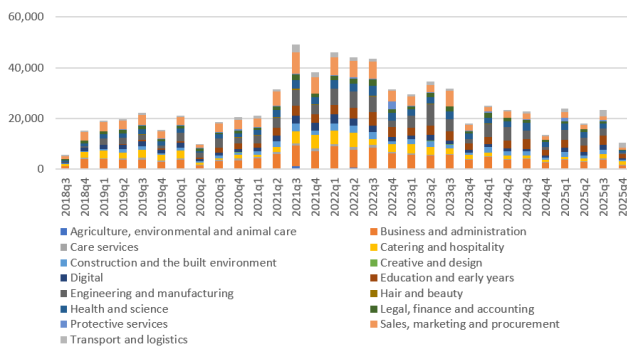


Fig. 1. Vacancies advertised on Find an apprenticeship, by quarter of closure date

#### D. Data Description

Overall vacancy volumes experienced severe contraction during the 2020 pandemic onset, followed by robust recovery peaking in 2022, and a subsequent gradual cooling through late 2025 (Fig. 1). Intermediate apprenticeships and business administration roles dominate traditional pathways (Fig. 2). Demand for the 25 digital skills is highly concentrated (Table I). Identifying technological needs appears in 45 percent of all vacancies, while information and communication tech-

TABLE II  
TOTAL DIGITAL SKILLS AND COMPETENCES PER ADVERT BY ROUTE

| Average of digital skills mentioned        | 2020 | 2022 | 2024 |
|--------------------------------------------|------|------|------|
| Agriculture, environmental and animal care | 0.55 | 1.16 | 0.91 |
| Business and administration                | 2.05 | 2.35 | 2.63 |
| Care services                              | 0.95 | 0.85 | 1.13 |
| Catering and hospitality                   | 0.24 | 0.25 | 0.24 |
| Construction and the built environment     | 1.04 | 1.30 | 1.56 |
| Creative and design                        | 1.40 | 1.29 | 0.86 |
| Digital                                    | 2.92 | 3.06 | 3.44 |
| Education and early years                  | 0.23 | 0.26 | 0.29 |
| Engineering and manufacturing              | 1.26 | 1.42 | 1.58 |
| Hair and beauty                            | 0.25 | 0.21 | 0.12 |
| Health and science                         | 1.10 | 0.71 | 0.53 |
| Legal, finance and accounting              | 1.71 | 2.06 | 2.16 |
| Protective services                        | 2.06 | 2.14 | 2.65 |
| Sales, marketing and procurement           | 1.61 | 2.06 | 2.39 |
| Transport and logistics                    | 1.34 | 1.44 | 1.69 |

nology safety is explicitly requested in 22 percent. Abstract digital creation skills exhibit near-zero presence. Digital skill intensity is compounding in technical sectors, with the digital route peaking at an average of 3.44 skills per advert in 2024 (Table II). The top five percent of apprenticeship standards most exposed to digital skills are heavily concentrated at the advanced, higher, and degree levels (Table III).

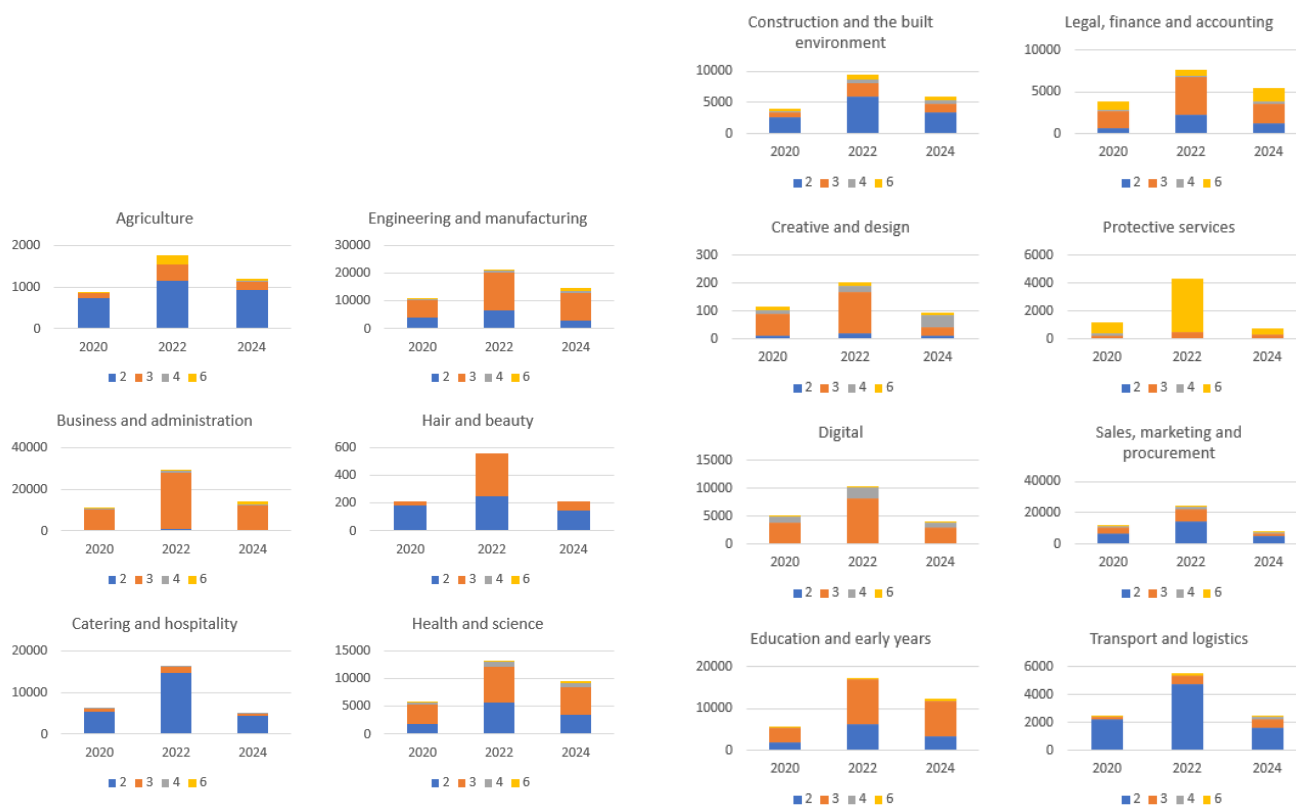


Fig. 2. Annual total vacancies by routes, levels and years

TABLE III  
TOP FIVE-PERCENTILE DIGITAL SKILLS APPRENTICESHIPS

| Standard Name                                     | Level        | Digital skills mentioned |
|---------------------------------------------------|--------------|--------------------------|
| Artificial intelligence (AI) data specialist      | Degree       | 3.00                     |
| Assistant technical director (visual effects)     | Higher       | 3.50                     |
| Broadcast and media systems technician            | Higher       | 4.92                     |
| Conveyancing technician                           | Higher       | 3.27                     |
| Cyber security technician                         | Advanced     | 2.97                     |
| Cyber security technologist                       | Higher       | 3.21                     |
| Cyber security technologist (2021)                | Higher       | 3.22                     |
| DevOps engineer                                   | Higher       | 2.95                     |
| Education technician                              | Advanced     | 5.33                     |
| Electrical power protection and plant c. engineer | Higher       | 3.67                     |
| Electro-mechanical engineer                       | Degree       | 2.85                     |
| Geospatial survey technician                      | Advanced     | 3.00                     |
| IT solutions technician                           | Advanced     | 3.53                     |
| Information communications technician             | Advanced     | 3.16                     |
| Infrastructure technician                         | Advanced     | 3.16                     |
| Investment operations administrator               | Intermediate | 6.00                     |
| Junior management consultant                      | Higher       | 3.22                     |
| Licensed conveyancer                              | Degree       | 6.00                     |
| Mammography associate                             | Higher       | 3.00                     |
| Post graduate engineer                            | Degree       | 3.00                     |
| Project controls professional                     | Degree       | 4.00                     |
| Revenues and welfare benefits practitioner        | Higher       | 2.88                     |
| Through life engineering services specialist      | Degree       | 3.00                     |
| Unified communications trouble shooter            | Higher       | 3.67                     |

#### IV. ECONOMETRIC MODELLING

This study investigates the empirical relationship between the widespread adoption of generative artificial intelligence and three key entry-level labour market outcomes: aggregate apprenticeship vacancy volumes, the intensity of required digital competencies, and shifts in specific occupational tasks. Utilizing a difference-in-differences strategy, the analysis defines the post-treatment period as 2023 and onwards, aligning with the market integration of advanced large language models. The primary treated group is strictly defined by sectors demanding a high magnitude of cognitive and digital skills, specifically the business, digital, protective services, and legal, finance and accounting routes. Conversely, the control group comprises vacancies in non-tangible human premium sectors reliant on physical craft or high-touch care, which are insulated from immediate algorithmic automation.

The econometric model testing hypotheses 1 and 3 at the level of occupation is specified as:

$$Y_{it} = \beta_0 + \beta_1(Post\_AI_t \times Exposed_i) + \gamma X_{it} + \delta_i + \theta_t + \epsilon_{it} \quad (1)$$

where  $Y_{it}$  represents job growth or tasks for occupation  $i$  at time  $t$ . The interaction term  $Post\_AI_t \times Exposed_i$  serves as the difference-in-differences estimator, capturing the shift in outcomes post-artificial intelligence introduction. The model includes fixed effects for the standard  $\delta_i$  and time  $\theta_t$ , further including specific route fixed effects  $X_{it}$ .

The econometric model evaluating Hypothesis 2 (digital skill intensity) is specified as:

$$Y_{it} = \beta_0 + \beta_1(Post\_AI_t \times Exposed_i) + \delta_i + \theta_t + \epsilon_{it} \quad (2)$$

where  $Y_{it}$  denotes the total number of digital competencies specified per individual advertisement  $i$  at time  $t$ , weighted by the number of available positions to reflect the true skill demand per job. To control for unobserved heterogeneity, the model incorporates time fixed effects ( $\theta_t$ ) alongside occupation fixed effects ( $\delta_i$ ). Estimations are conducted across the aggregate sample and subsequently stratified by educational level.

To validate the baseline estimates, two rigorous placebo tests are incorporated: a pre-artificial intelligence parallel trends test restricting data to 2021 and earlier, and a randomized sector assignment test which removes truly exposed sectors and randomly assigns treatment status to the remaining control sectors.

#### V. ARTIFICIAL INTELLIGENCE IMPACTS

##### A. Impact on Vacancy Volumes (Hypothesis 1)

The initial hypothesis posits that artificial intelligence disruption negatively impacts the total volume of job adverts in highly exposed sectors. Consistent with evidence that organizational friction and the delayed redesign of task structures inhibit immediate macroeconomic productivity shocks, the empirical results demonstrate no significant aggregate disruption to entry-level vacancy volumes.

The difference-in-differences estimator for the aggregate sample is negative but statistically indistinguishable from zero (Table IV). When stratified by educational level, the magnitude of the coefficient decreases as qualification requirements rise, yet none of the estimates achieve statistical significance. Consequently, the null hypothesis of no macroeconomic impact on the aggregate volume of apprenticeships cannot be rejected.

##### B. Impact on Digital Skill Requirements (Hypothesis 2)

The second hypothesis evaluates whether artificial intelligence shifts the intensity of digital competencies expected in entry-level roles. The primary model indicates that artificial intelligence exposed sectors inherently demand a higher baseline of digital competencies, and the post-treatment interaction reveals a statistically significant aggregate increase in total digital skill requirements (Table V).

However, robustness testing highlights critical nuances. Pre-period parallel trends tests indicate that intermediate Level 2 and advanced Level 3 apprenticeships exhibited significant divergence in skill trends well before the 2023 treatment period, invalidating causal attribution to generative artificial intelligence for lower tiers. In contrast, placebo models confirm the validity of the difference-in-differences estimator for Level 4, Level 5, and Degree Level 6 apprenticeships. The combined evidence credibly rejects the null hypothesis for higher educational tiers, demonstrating that artificial intelligence has driven a genuine, modest rise in digital skill requirements for advanced roles, reflecting skill complementarity rather than a hollowing-out effect.

##### C. Impact on Occupational Tasks (Hypothesis 3)

The third hypothesis asserts that artificial intelligence induces a significant shift or reduction in technical task density within highly exposed roles. Utilizing the deterministically extracted 30-task matrix, the econometric results fail to find pervasive evidence of task shifts. Across the majority of specified models, the interaction term for artificial intelligence exposure remains statistically insignificant (Table X and Table XI).

While isolated point estimates suggest a task shift occurring within higher Level 4 and 5 apprenticeships, scrutiny via pre-period robustness tests reveals that these specific occupational standards fail the parallel trends assumption (Table XII). These sectors exhibited significant differences in task trends prior to the advent of generative artificial intelligence, indicating that observed changes are manifestations of ongoing, long-term structural digitalization rather than an immediate artificial intelligence shock. Given the absence of robust coefficients, the null hypothesis regarding task shifts cannot be rejected.

#### VI. DISCUSSION AND POLICY IMPLICATIONS

The empirical analysis confirms that generative artificial intelligence has not precipitated an aggregate macroeconomic shock within the United Kingdom youth entry-level labour market. The observable decline in apprenticeship volumes

TABLE IV  
IMPACT ON GROWTH OF APPRENTICESHIP VACANCIES (LOGLINEAR)

|                              | (1)<br>All<br>(Sectors) | (2)<br>Level 2<br>Intermediate | (3)<br>Level 3<br>Advanced | (4)<br>Level 4-5<br>Higher | (5)<br>Level 6<br>Degree | (6)<br>All<br>(High ICT) |
|------------------------------|-------------------------|--------------------------------|----------------------------|----------------------------|--------------------------|--------------------------|
| Post-AI (time)               | -0.136<br>(0.103)       | -0.437*<br>(0.185)             | -0.0940<br>(0.169)         | 0.169<br>(0.213)           | 0.155<br>(0.232)         | -0.0817<br>(0.0938)      |
| AI Exposed Sector (affected) | 0.103<br>(0.166)        | 1.584*<br>(0.740)              | 0.973***<br>(0.281)        | 0.272<br>(0.258)           | 0.409<br>(0.325)         | -0.426<br>(0.346)        |
| DiD (time_affect)            | -0.108<br>(0.218)       | -1.177<br>(0.942)              | -0.398<br>(0.372)          | -0.312<br>(0.339)          | 0.0380<br>(0.424)        | 0.314<br>(0.468)         |
| Constant                     | 3.643***<br>(0.0781)    | 4.325***<br>(0.140)            | 3.899***<br>(0.128)        | 2.688***<br>(0.162)        | 2.639***<br>(0.180)      | 3.764***<br>(0.0707)     |
| N                            | 2071                    | 528                            | 735                        | 454                        | 354                      | 1937                     |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE V  
IMPACT ON NUMBER OF DIGITAL SKILLS MENTIONED PER ADVERT

|                   | (1)<br>All<br>(Sectors) | (2)<br>All<br>(Sectors) | (3)<br>Level 2<br>Intermediate | (4)<br>Level 2<br>Intermediate | (5)<br>Level 3<br>Advanced | (6)<br>Level 3<br>Advanced | (7)<br>Level 4-5<br>Higher | (8)<br>Level 4-5<br>Higher | (9)<br>Level 6<br>Degree | (10)<br>Level 6<br>Degree |
|-------------------|-------------------------|-------------------------|--------------------------------|--------------------------------|----------------------------|----------------------------|----------------------------|----------------------------|--------------------------|---------------------------|
| Post-AI (time)    | -0.0198***<br>(0.00512) | 0.0743***<br>(0.00480)  | 0.0251***<br>(0.00666)         | 0.0827***<br>(0.00593)         | -0.106***<br>(0.00837)     | 0.0559***<br>(0.00817)     | 0.0934***<br>(0.0245)      | 0.193***<br>(0.0226)       | -0.129***<br>(0.0228)    | 0.00808<br>(0.0233)       |
| AI Exposed Sector | 1.143***<br>(0.00642)   | -5.075<br>(4562.7)      | 0.886***<br>(0.0195)           | 1.888***<br>(0.226)            | 1.272***<br>(0.00881)      | 14.52<br>(8678.2)          | 0.650***<br>(0.0277)       | 0.0170<br>(1.730)          | -0.371***<br>(0.0234)    | 1.679<br>(1.409)          |
| DiD (time_affect) | 0.434***<br>(0.00946)   | 0.348***<br>(0.00876)   | 0.488***<br>(0.0302)           | 0.381***<br>(0.0264)           | 0.598***<br>(0.0127)       | 0.369***<br>(0.0122)       | 0.392***<br>(0.0386)       | 0.290***<br>(0.0363)       | 0.398***<br>(0.0302)     | 0.312***<br>(0.0307)      |
| Constant          | 0.998***<br>(0.00353)   | 0.729**<br>(0.237)      | 0.944***<br>(0.00430)          | 0.728***<br>(0.211)            | 1.044***<br>(0.00610)      | -8.518<br>(8678.2)         | 1.189***<br>(0.0183)       | 3.49e-11<br>(1.412)        | 1.263***<br>(0.0181)     | -1.20e-09<br>(1.150)      |
| Standard FE       |                         | Included                |                                | Included                       |                            | Included                   |                            | Included                   |                          | Included                  |
| N                 | 582824                  | 582824                  | 220125                         | 220125                         | 303918                     | 303918                     | 28935                      | 28935                      | 29846                    | 29846                     |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE VI  
ROBUSTNESS TEST: PRE-AI PARALLEL TRENDS

|                   | (1)<br>All              | (2)<br>All             | (3)<br>Level 2        | (4)<br>Level 2         | (5)<br>Level 3        | (6)<br>Level 3         | (7)<br>Level 4        | (8)<br>Level 4       | (9)<br>Level 6        | (10)<br>Level 6      |
|-------------------|-------------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|-----------------------|----------------------|-----------------------|----------------------|
| Post-AI (time)    | -0.0726***<br>(0.00755) | 0.0318***<br>(0.00737) | -0.00392<br>(0.00902) | 0.0897***<br>(0.00855) | -0.240***<br>(0.0132) | -0.0714***<br>(0.0132) | 0.412***<br>(0.0536)  | 0.135*<br>(0.0553)   | 0.121*<br>(0.0510)    | 0.179***<br>(0.0511) |
| AI Exposed Sector | 0.892***<br>(0.0107)    | -0.943<br>(1.478)      | 0.272***<br>(0.0399)  | 0.420<br>(0.267)       | 0.856***<br>(0.0151)  | -1.001<br>(10033.6)    | 1.005***<br>(0.0557)  | 1.610<br>(1.581)     | -0.350***<br>(0.0524) | -2.939<br>(1.779)    |
| DiD (time_affect) | 0.142***<br>(0.0131)    | 0.000564<br>(0.0127)   | 0.501***<br>(0.0463)  | 0.150**<br>(0.0477)    | 0.266***<br>(0.0183)  | 0.0821***<br>(0.0183)  | -0.453***<br>(0.0641) | -0.0778<br>(0.0649)  | 0.153*<br>(0.0611)    | -0.0614<br>(0.0601)  |
| Constant          | 1.044***<br>(0.00633)   | 0.910***<br>(0.274)    | 0.918***<br>(0.00748) | 0.877***<br>(0.246)    | 1.267***<br>(0.0111)  | 3.700<br>(15032.7)     | 0.749***<br>(0.0488)  | -3.82e-10<br>(1.369) | 1.098***<br>(0.0453)  | 3.821**<br>(1.258)   |
| Standard FE       |                         | Included               |                       | Included               |                       | Included               |                       | Included             |                       | Included             |
| N                 | 307320                  | 307320                 | 128534                | 128534                 | 153787                | 153787                 | 13122                 | 13122                | 11877                 | 11877                |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE VII  
ROBUSTNESS TEST 2: RANDOMIZED SECTORS

|                   | (1)<br>All             | (2)<br>All             | (3)<br>Level 2        | (4)<br>Level 2         | (5)<br>Level 3          | (6)<br>Level 3         | (7)<br>Level 4       | (8)<br>Level 4       | (9)<br>Level 6        | (10)<br>Level 6     |
|-------------------|------------------------|------------------------|-----------------------|------------------------|-------------------------|------------------------|----------------------|----------------------|-----------------------|---------------------|
| Post-AI (time)    | -0.0139**<br>(0.00537) | 0.0682***<br>(0.00480) | 0.0232**<br>(0.00784) | 0.0615***<br>(0.00686) | -0.0933***<br>(0.00798) | 0.0657***<br>(0.00720) | 0.125***<br>(0.0271) | 0.190***<br>(0.0241) | -0.105***<br>(0.0305) | 0.0182<br>(0.0293)  |
| AI Exposed Sector | 0.0399***<br>(0.00679) | 0.0183**<br>(0.00586)  | 0.0285**<br>(0.00928) | 0.00301<br>(0.00797)   | 0.0507***<br>(0.0107)   | 0.0399***<br>(0.00926) | 0.0736*<br>(0.0366)  | -0.00661<br>(0.0317) | 0.105*<br>(0.0442)    | 0.0897*<br>(0.0391) |
| DiD (time_affect) | -0.0190<br>(0.00987)   | 0.0216*<br>(0.00853)   | 0.00838<br>(0.0146)   | 0.0753***<br>(0.0125)  | -0.0418**<br>(0.0146)   | -0.0327*<br>(0.0127)   | -0.103*<br>(0.0488)  | 0.00858<br>(0.0423)  | -0.0803<br>(0.0556)   | -0.0333<br>(0.0495) |
| Constant          | 0.986***<br>(0.00372)  | 0.730***<br>(0.202)    | 0.936***<br>(0.00510) | 0.730***<br>(0.207)    | 1.029***<br>(0.00581)   | -0.0399<br>(1.197)     | 1.167***<br>(0.0203) | 1.007<br>(1.263)     | 1.231***<br>(0.0243)  | -0.0745<br>(1.244)  |
| Standard FE       |                        | Included               |                       | Included               |                         | Included               |                      | Included             |                       | Included            |
| N                 | 411683                 | 411683                 | 209434                | 209434                 | 171403                  | 171403                 | 17441                | 17441                | 13405                 | 13405               |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE VIII  
ROBUSTNESS TEST: PRE-AI PARALLEL TRENDS

|                   | (1)<br>All              | (2)<br>All             | (3)<br>Level 2        | (4)<br>Level 2         | (5)<br>Level 3        | (6)<br>Level 3         | (7)<br>Level 4        | (8)<br>Level 4       | (9)<br>Level 6        | (10)<br>Level 6      |
|-------------------|-------------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|-----------------------|----------------------|-----------------------|----------------------|
| Post-AI (time)    | -0.0726***<br>(0.00755) | 0.0318***<br>(0.00737) | -0.00392<br>(0.00902) | 0.0897***<br>(0.00855) | -0.240***<br>(0.0132) | -0.0714***<br>(0.0132) | 0.412***<br>(0.0536)  | 0.135*<br>(0.0553)   | 0.121*<br>(0.0510)    | 0.179***<br>(0.0511) |
| AI Exposed Sector | 0.892***<br>(0.0107)    | -0.943<br>(1.478)      | 0.272***<br>(0.0399)  | 0.420<br>(0.267)       | 0.856***<br>(0.0151)  | -1.001<br>(10033.6)    | 1.005***<br>(0.0557)  | 1.610<br>(1.581)     | -0.350***<br>(0.0524) | -2.939<br>(1.779)    |
| DiD (time_affect) | 0.142***<br>(0.0131)    | 0.000564<br>(0.0127)   | 0.501***<br>(0.0463)  | 0.150**<br>(0.0477)    | 0.266***<br>(0.0183)  | 0.0821***<br>(0.0183)  | -0.453***<br>(0.0641) | -0.0778<br>(0.0649)  | 0.153*<br>(0.0611)    | -0.0614<br>(0.0601)  |
| Constant          | 1.044***<br>(0.00633)   | 0.910***<br>(0.274)    | 0.918***<br>(0.00748) | 0.877***<br>(0.246)    | 1.267***<br>(0.0111)  | 3.700<br>(15032.7)     | 0.749***<br>(0.0488)  | -3.82e-10<br>(1.369) | 1.098***<br>(0.0453)  | 3.821**<br>(1.258)   |
| Standard FE       |                         | Included               |                       | Included               |                       | Included               |                       | Included             |                       | Included             |
| N                 | 307320                  | 307320                 | 128534                | 128534                 | 153787                | 153787                 | 13122                 | 13122                | 11877                 | 11877                |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE IX  
ROBUSTNESS TEST 2: RANDOMIZED SECTORS

|                   | (1)<br>All             | (2)<br>All             | (3)<br>Level 2        | (4)<br>Level 2         | (5)<br>Level 3          | (6)<br>Level 3         | (7)<br>Level 4       | (8)<br>Level 4       | (9)<br>Level 6        | (10)<br>Level 6     |
|-------------------|------------------------|------------------------|-----------------------|------------------------|-------------------------|------------------------|----------------------|----------------------|-----------------------|---------------------|
| Post-AI (time)    | -0.0139**<br>(0.00537) | 0.0682***<br>(0.00480) | 0.0232**<br>(0.00784) | 0.0615***<br>(0.00686) | -0.0933***<br>(0.00798) | 0.0657***<br>(0.00720) | 0.125***<br>(0.0271) | 0.190***<br>(0.0241) | -0.105***<br>(0.0305) | 0.0182<br>(0.0293)  |
| AI Exposed Sector | 0.0399***<br>(0.00679) | 0.0183**<br>(0.00586)  | 0.0285**<br>(0.00928) | 0.00301<br>(0.00797)   | 0.0507***<br>(0.0107)   | 0.0399***<br>(0.00926) | 0.0736*<br>(0.0366)  | -0.00661<br>(0.0317) | 0.105*<br>(0.0442)    | 0.0897*<br>(0.0391) |
| DiD (time_affect) | -0.0190<br>(0.00987)   | 0.0216*<br>(0.00853)   | 0.00838<br>(0.0146)   | 0.0753***<br>(0.0125)  | -0.0418**<br>(0.0146)   | -0.0327*<br>(0.0127)   | -0.103*<br>(0.0488)  | 0.00858<br>(0.0423)  | -0.0803<br>(0.0556)   | -0.0333<br>(0.0495) |
| Constant          | 0.986***<br>(0.00372)  | 0.730***<br>(0.202)    | 0.936***<br>(0.00510) | 0.730***<br>(0.207)    | 1.029***<br>(0.00581)   | -0.0399<br>(1.197)     | 1.167***<br>(0.0203) | 1.007<br>(1.263)     | 1.231***<br>(0.0243)  | -0.0745<br>(1.244)  |
| Standard FE       |                        | Included               |                       | Included               |                         | Included               |                      | Included             |                       | Included            |
| N                 | 411683                 | 411683                 | 209434                | 209434                 | 171403                  | 171403                 | 17441                | 17441                | 13405                 | 13405               |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE X  
IMPACT ON TOTAL NUMBER OF JOB TASKS FOUND IN ADVERTS (MODELLED WITH ROUTES AFFECTED)

|                   | (1)<br>All          | (2)<br>All           | (3)<br>Level 2<br>Inter. | (4)<br>Level 2<br>Inter. | (5)<br>Level 3<br>Adv. | (6)<br>Level 3<br>Adv. | (7)<br>Level 4-5<br>Higher | (8)<br>Level 4-5<br>Higher | (9)<br>Level 6<br>Degree | (10)<br>Level 6<br>Degree |
|-------------------|---------------------|----------------------|--------------------------|--------------------------|------------------------|------------------------|----------------------------|----------------------------|--------------------------|---------------------------|
| Post-AI (time)    | -1.015*<br>(0.416)  | -1.030***<br>(0.156) | -1.267<br>(0.714)        | -2.177***<br>(0.248)     | -0.997<br>(0.679)      | -0.868***<br>(0.251)   | -0.266<br>(1.002)          | 0.113<br>(0.434)           | 0.349<br>(0.783)         | 0.839*<br>(0.387)         |
| AI Exposed Sector | 5.608***<br>(0.616) | 8.698*<br>(3.455)    | 7.026***<br>(2.074)      | 1.667<br>(2.643)         | 8.183***<br>(0.972)    | 0.646<br>(3.664)       | 5.275***<br>(1.097)        | -1.738<br>(3.313)          | 4.125***<br>(1.029)      | -1.098<br>(2.873)         |
| DiD (time_affect) | 0.757<br>(0.886)    | 0.332<br>(0.331)     | 4.351<br>(3.267)         | 1.249<br>(1.144)         | -0.148<br>(1.399)      | -0.646<br>(0.513)      | 0.166<br>(1.532)           | -0.375<br>(0.662)          | 1.183<br>(1.450)         | 0.0984<br>(0.703)         |
| Constant          | 7.244***<br>(0.293) | -1.44e-10<br>(2.434) | 8.390***<br>(0.488)      | -1.30e-12<br>(2.289)     | 7.668***<br>(0.478)    | 0.868<br>(2.577)       | 5.365***<br>(0.730)        | 5.000*<br>(2.316)          | 3.259***<br>(0.587)      | 1.161<br>(2.007)          |
| Standard FE       |                     | Included             |                          | Included                 |                        | Included               |                            | Included                   |                          | Included                  |
| N                 | 1413                | 1413                 | 404                      | 404                      | 602                    | 602                    | 234                        | 234                        | 173                      | 173                       |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE XI  
IMPACT ON TOTAL NUMBER OF JOB TASKS FOUND IN ADVERTS (MODELLED AS STANDARDS AFFECTED)

|                   | (1)<br>All          | (2)<br>All          | (3)<br>Level 2<br>Inter. | (4)<br>Level 2<br>Inter. | (5)<br>Level 3<br>Adv. | (6)<br>Level 3<br>Adv. | (7)<br>Level 4-5<br>Higher | (8)<br>Level 4-5<br>Higher | (9)<br>Level 6<br>Degree | (10)<br>Level 6<br>Degree |
|-------------------|---------------------|---------------------|--------------------------|--------------------------|------------------------|------------------------|----------------------------|----------------------------|--------------------------|---------------------------|
| Post-AI (time)    | -0.849*<br>(0.371)  | -0.800*<br>(0.353)  | -1.059<br>(0.698)        | -1.190<br>(0.616)        | -1.160<br>(0.599)      | -0.957<br>(0.553)      | -0.0413<br>(0.774)         | 0.137<br>(0.669)           | 0.756<br>(0.669)         | 0.592<br>(0.638)          |
| AI Exposed Sector | 5.974***<br>(0.445) | 7.941***<br>(0.976) | 8.780***<br>(1.604)      | 19.08***<br>(2.875)      | 7.990***<br>(0.703)    | 12.44***<br>(1.791)    | 5.483***<br>(0.775)        | 4.477<br>(3.666)           | 4.677***<br>(0.730)      | -3.704<br>(4.425)         |
| DiD (time_affect) | 0.0238<br>(1.340)   | -0.889<br>(1.299)   | 0<br>(.)                 | 0<br>(.)                 | 3.118<br>(2.157)       | -0.582<br>(2.086)      | -1.818<br>(1.936)          | -3.639*<br>(1.688)         | -1.003<br>(1.833)        | -0.0506<br>(1.810)        |
| Constant          | 7.161***<br>(0.277) | 4.358***<br>(0.726) | 8.293***<br>(0.483)      | 5.599***<br>(0.867)      | 7.715***<br>(0.449)    | 1.837<br>(1.480)       | 5.273***<br>(0.643)        | 0.932<br>(3.503)           | 3.080***<br>(0.538)      | 3.704*<br>(1.698)         |
| Route FE          |                     | Included            |                          | Included                 |                        | Included               |                            | Included                   |                          | Included                  |
| N                 | 1413                | 1413                | 404                      | 404                      | 602                    | 602                    | 234                        | 234                        | 173                      | 173                       |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

TABLE XII  
ROBUSTNESS TEST FOR STANDARDS AFFECTED: PRE-AI PARALLEL TRENDS

|                   | (1)<br>All          | (2)<br>All          | (3)<br>Level 2<br>Inter. | (4)<br>Level 2<br>Inter. | (5)<br>Level 3<br>Adv. | (6)<br>Level 3<br>Adv. | (7)<br>Level 4-5<br>Higher | (8)<br>Level 4-5<br>Higher | (9)<br>Level 6<br>Degree | (10)<br>Level 6<br>Degree |
|-------------------|---------------------|---------------------|--------------------------|--------------------------|------------------------|------------------------|----------------------------|----------------------------|--------------------------|---------------------------|
| Post-AI (time)    | 0.829<br>(0.599)    | 0.965<br>(0.570)    | 0.353<br>(1.121)         | 0.825<br>(1.001)         | 1.280<br>(0.928)       | 1.424<br>(0.861)       | 1.237<br>(1.262)           | 1.724<br>(1.058)           | 0.445<br>(1.037)         | 0.108<br>(1.032)          |
| AI Exposed Sector | 5.616***<br>(0.660) | 7.497*<br>(3.534)   | 7.049**<br>(2.181)       | 19.56***<br>(3.948)      | 7.616***<br>(1.023)    | 12.97***<br>(2.517)    | 6.096***<br>(1.146)        | 6.445<br>(4.867)           | 4.088***<br>(0.990)      | 4.833<br>(2.824)          |
| DiD (time_affect) | -0.131<br>(1.728)   | -1.127<br>(1.740)   | 0<br>(.)                 | 0<br>(.)                 | 5.346<br>(2.761)       | 5.002<br>(2.952)       | -4.906*<br>(2.129)         | -7.947***<br>(1.800)       | -0.917<br>(2.984)        | -0.120<br>(3.001)         |
| Constant          | 6.646***<br>(0.527) | 3.780***<br>(1.084) | 8.132***<br>(0.967)      | 5.222***<br>(1.381)      | 6.758***<br>(0.817)    | 0.730<br>(2.111)       | 4.462***<br>(1.169)        | -1.724<br>(4.755)          | 2.972**<br>(0.933)       | 2.595<br>(2.407)          |
| Route FE          |                     | Included            |                          | Included                 |                        | Included               |                            | Included                   |                          | Included                  |
| N                 | 716                 | 716                 | 217                      | 217                      | 306                    | 306                    | 113                        | 113                        | 80                       | 80                        |

Note: Standard errors in parentheses. Significance levels: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

stems from compounding structural factors, including shifting minimum wage frameworks, economic stagnation, and the systemic redirection of levy funding toward incumbent workers [14]. Rather than evidencing technological displacement or technical task reduction, the data reveals a modest increase in digital competency requirements exclusively for higher-tier apprenticeships. This indicates that advanced digital proficiency operates as a strong complement to artificial intelligence. Consequently, policy interventions should prioritize repairing collapsed intermediate training pipelines to equip young entrants with requisite digital literacies, rather than mitigating immediate algorithmic job destruction.

These null findings align directly with the macroeconomic productivity paradox associated with artificial intelligence [1], [2]. While micro-level experiments demonstrate performance gains in codifiable functions, translating these isolated efficiencies into aggregate macroeconomic shocks requires profound organizational restructuring. Because algorithms execute specific functions rapidly while organizations adapt slowly, this inherent friction significantly delays observable entry-level labour displacement. Furthermore, transitioning from automating routine functions to hard-to-learn tasks involving tacit knowledge yields diminishing marginal returns for total factor productivity [1]. To ensure inclusive economic growth, institutional incentives must steer technological development toward the reinstatement effect [5]. By prioritizing the creation of new, human-complementary tasks that demand sophisticated digital competencies, policymakers can ensure artificial intelligence functions as a mechanism for robust labour augmentation rather than widespread occupational substitution [6].

#### ACKNOWLEDGMENT

Corresponding author: Dr Lei Xu, Email: l.xu2@lboro.ac.uk. Our research makes use of apprenticeship vacancies published by the UK Department for Education (DfE), which are subject to Crown Copyright, and have been made available for public use under Open Government Licence v3.0 (OGL).

#### GENERATIVE AI DISCLOSURE STATEMENT

During the preparation of this work, the authors used Google Gemini 1.5 Flash / Gemini 3.5 Flash to support the development of Python code for data processing and LLM implementation. Gemini and NotebookLM Pro were also used for editing and reducing the draft write-up, which did not affect the content. The authors take full responsibility for the content of this publication.

#### REFERENCES

- [1] D. Acemoglu, "The Simple Macroeconomics of AI," *Massachusetts Institute of Technology*, Working Paper, 2024.
- [2] D. Acemoglu, "The Simple Macroeconomics of AI," *Economic Policy*, vol. 40, no. 121, 2025.
- [3] D. Acemoglu and D. Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," *Handbook of Labor Economics*, vol. 4, pp. 1043–1171, 2011.
- [4] D. Acemoglu and P. Restrepo, "Artificial Intelligence, Automation, and Work," *The Economics of Artificial Intelligence: An Agenda*, NBER, pp. 197–236, 2018.
- [5] D. Acemoglu and P. Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, vol. 33, no. 2, pp. 3–30, 2019.
- [6] D. Autor, "Applying AI to Rebuild Middle Class Jobs," *National Bureau of Economic Research*, Working Paper 32140, 2024.
- [7] D. H. Autor, F. Levy, and R. J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, vol. 118, no. 4, pp. 1279–1333, 2003.
- [8] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at Work," *National Bureau of Economic Research*, Working Paper 31161, 2023.
- [9] Department for Education, "Apprenticeships and traineeships: Academic year 2025/26," Explore Education Statistics, HM Government, 2026.
- [10] A. Korinek, "Economic policy challenges for the age of AI," *National Bureau of Economic Research*, Working Paper 32980, 2024.
- [11] Office for National Statistics, "Young people not in education, employment, and training (NEET), UK," HM Government, 2026.
- [12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *Proceedings of EMNLP*, 2019.
- [13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [14] The Sutton Trust, *The apprenticeship levy and social mobility: Who benefits from the shift to higher-level training?*, The Sutton Trust, 2024.



# Iterative Inductive Learning for Generalized Modeling of Technical Systems

Balázs Ádám Toldi, András Földvári, and András Pataricza  
0009-0007-4751-5043, 0000-0002-4559-6990, 0000-0002-6516-129X  
Budapest University of Technology and Economics  
Budapest, Hungary

Email: balazs.toldi@edu.bme.hu, {foldvari.andras, pataricza.andras}@vik.bme.hu

Zsigmond Pap  
0009-0004-4980-8075  
Ericsson Hungary  
Budapest, Hungary

Email: zsigmond.pap@ericsson.com

**Abstract**—Designing critical IT infrastructures to meet service-level agreements requires evaluating a combinatorial space of deployment parameters and workload scenarios that cannot be exhaustively tested. We present an integrated framework that operates on qualitative abstractions of measurement data to produce deterministic, interpretable classification rules transferable across deployments with different hardware dimensioning. The framework combines Rough Set Theory (RST) with Inductive Logic Programming (ILP): RST separates certain from ambiguous measurement outcomes and directly constrains the ILP rule learner through an asymmetric cost function reflecting safety requirements. A monotonicity check validates resource-scaling assumptions and classifies violations, and when violations indicate architectural bottlenecks, a structural system model with causal annotations traces root causes and recommends targeted measurement extensions. We evaluate the framework on a 5G UPF scaling testbed.

## I. INTRODUCTION

MODERN IT infrastructures must meet user needs while efficiently using system resources. To this end, modular and scalable construction is prevalent, with functional and performance compliance increasingly obtained through modular composition at installation time. This allows an extremely large number of final systems to be built from a single basic set, but the resulting combinatorial explosion prevents comprehensive testing of the entire system.

For instance, a containerized network service, such as a 5G Core deployment, is dimensioned using numerous parameters (e.g., the number and capacity of processing instances, CPU and memory provisioning, network configuration) that must meet strict service-level agreements (SLAs) under diverse designated workloads (varying traffic types and client counts). Even a modest number of parameters with a few qualitative levels yields hundreds or thousands of distinct configurations.

Since exhaustive testing is infeasible, the central problem is *prediction by extrapolation from limited benchmark measurements*: from a small number of tested configurations, reliably predict whether an untested combination of deployment parameters will satisfy SLAs under a given workload. In safety-critical infrastructures, the answer must be conservative: given the severe impact of service failures, it is preferable to flag a viable configuration as potentially failing than to declare a failing configuration healthy.

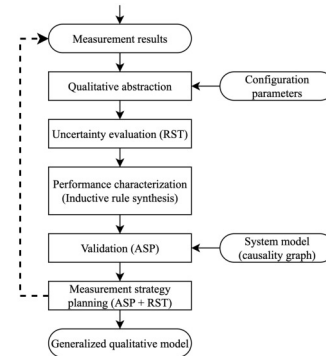


Fig. 1. Workflow

Standard machine-learning approaches fall short on three counts. They are probabilistic classifiers producing confidence scores rather than the deterministic pass/fail verdicts that strict SLAs require; their black-box nature prevents operators from inspecting the reason behind a classification, hindering root-cause diagnosis; and models trained on absolute metric values (e.g., CPU utilization) are tied to a specific hardware dimensioning and are not portable across deployments.

We propose an integrated framework that relies on *qualitative abstraction* to derive a discrete model that preserves the main phenomena influencing the SLA while eliminating the scalability and portability problems of the continuous domain. On this representation it combines *Rough Set Theory* (RST), which reasons about the uncertainty and inconsistency introduced by abstraction and noise without distributional assumptions, with *Inductive Logic Programming* (ILP), which produces human-readable deterministic rules, to extract a generalized, interpretable, and transferable model for configurable systems (Section II gives the necessary background).

### A. Workflow

The framework operates as a closed-loop pipeline (Fig. 1): *measurement campaigns* produce benchmark data; RST-based *uncertainty evaluation* separates certain from ambiguous outcomes; *performance characterization* synthesizes deterministic rules under RST-derived constraints; a *validation* stage checks physical assumptions (e.g., monotonicity under re-

source scaling) and, where available, a structural system model to trace inconsistencies; and *measurement strategy planning* produces a prioritized list of configurations for the next round. By expressing the entire pipeline in *Answer Set Programming* (ASP), uncertainty evaluation, rule synthesis, and measurement planning reside in a single declarative framework, enabling formal verification and seamless integration of domain knowledge.

The main contributions are:

- 1) An *integrated framework* in which RST analysis directly constrains and parameterizes inductive rule learning: lower approximations define hard coverage/exclusion constraints, while the boundary region and outcome frequencies drive an asymmetric cost function for conservative classification under noise.
- 2) An *ASP-encoded checking module* that detects physical constraints (dominance violations under the same workload) and classifies their confidence via RST boundary membership (distinguishing genuine non-monotonicity from noise-induced inconsistency).
- 3) A targeted *measurement strategy* derived from these analyses, replacing exhaustive or random testing with a prioritized selection that maximally reduces uncertainty and validates the assumptions behind the learned rules.

The remainder of the paper reviews background (Section II), presents the framework and data preparation (Section III), formalizes the RST uncertainty evaluation (Section IV) and the one-sided rule learning with an asymmetric cost function (Section V), details validation (monotonicity checking, RST integration, system-model root-cause analysis, and the measurement strategy) (Section VI) and reports a case study (Section VII).

### B. Running Example

Throughout the paper we use, as a running example, the scaling of a distributed packet-processing microservice deployed in a Kubernetes cluster, treating it as a representative data-intensive service rather than dwelling on protocol details. A custom traffic generator drives a configurable number of client devices through one or more service instances and measures end-to-end performance, forming a closed measurement loop. Each experiment is defined by a *workload* (a traffic profile and a client count) and a *deployment configuration* (CPU allocation and instance count), and is summarized by application-level metrics (CPU and memory utilization, packet rates, throughput, packet drops) from which the HEALTHY/UNHEALTHY outcome is derived. The concrete testbed and parameter ranges are detailed in the case study (Section VII).

### C. Related work

#### a) Performance modeling of configurable systems:

Performance-influence models [1] learn the effect of configuration options via stepwise linear regression, with later work reducing sample size through sampling strategies [2] or reusing models across environments via transfer learning [3].

These approaches produce continuous regression outputs (confidence scores or numerical estimates) rather than deterministic pass/fail verdicts, are opaque about *why* a configuration fails, and are tied to a specific hardware dimensioning because they are trained on absolute metric values. Our framework addresses all three: qualitative abstraction maps measurements to relative levels (LOW/MEDIUM/HIGH), making rules hardware-independent by construction; ILP produces deterministic, human-readable rules; and RST-based uncertainty quantification drives an asymmetric cost function that guarantees conservative classification.

b) *Rough set methods in cloud and IT systems*: Recent work applies rough sets and their derivatives to operational cloud decisions: three-way decision fusion for elastic scaling [4] (whose learnability is formally established [5]), decision-theoretic rough sets for task classification [6], and RST discernibility [7] for QoS dimensionality reduction before metaheuristic service composition [8]. These use RST as a one-shot classifier or a pre-processing reducer. We instead treat the boundary region as the central analytical object that (i) drives an asymmetric cost function for conservative rule learning, (ii) classifies the confidence of monotonicity violations, and (iii) iteratively guides the measurement strategy. Moreover, we couple RST with ILP to produce deterministic, transferable rules, which none of the above do.

## II. BACKGROUND

### A. Qualitative Modeling

Qualitative modeling [9] represents continuous aspects of systems symbolically: instead of precise numerical values, it abstracts quantities into discrete categories capturing the operationally relevant distinctions. For example, CPU allocations of 500 and 1000 millicore can be represented as LOW and HIGH relative to the workload's demand, without committing to absolute thresholds that differ across hardware. The abstraction is governed by three principles [9]: *discretization* partitions a continuous domain into qualitative values defined by context-meaningful thresholds; *relevance* requires the chosen categories to capture the distinctions that matter (here, whether a configuration meets its SLA); and *ambiguity* is the inherent cost of abstraction, since values near a threshold may behave differently despite sharing a qualitative label, producing inconsistent observations.

In our framework, qualitative modeling serves two purposes. It enables *scale-independent* rules that transfer across deployments with different hardware dimensioning, provided thresholds are defined relative to capacity; and the discretization yields a finite attribute space directly compatible with RST's indiscernibility relation, where the abstraction's ambiguity is precisely what the boundary region captures.

### B. Rough Set Theory

Rough Set Theory (RST) [10] reasons about uncertainty and inconsistency in data without requiring prior assumptions about data distributions, relying solely on the observed data to separate certain knowledge from ambiguous cases.

An *information system*  $(U, A)$  consists of a finite set of objects  $U$  and attributes  $A$ . A *decision table*  $(U, A \cup \{d\})$  adds a decision attribute; in our setting,  $d \in \{\text{HEALTHY}, \text{UNHEALTHY}\}$ . Any subset  $B \subseteq A$  induces an *indiscernibility relation* that groups objects sharing the same values on all attributes in  $B$  into equivalence classes.

Given a decision class  $X \subseteq U$ , the *lower approximation*  $\underline{B}(X)$  contains equivalence classes entirely within  $X$  (certain membership), the *upper approximation*  $\overline{B}(X)$  contains classes with any overlap with  $X$  (possible membership), and the *boundary region*  $\text{BND}_B(X) = \overline{B}(X) \setminus \underline{B}(X)$  contains classes that cannot be deterministically classified. The *approximation quality*  $\gamma_B(d)$  measures the fraction of objects outside the boundary region;  $\gamma = 1$  means fully deterministic classification.

### C. Answer Set Programming

Answer Set Programming (ASP) [11], [12] is a declarative formalism for modeling and solving search and optimization problems, based on the stable model semantics [13].

An ASP program consists of rules of the form:

$h :- a_1, \dots, a_n, \text{not } b_1, \dots, \text{not } b_m.$

where  $h$  is the *head* and the conjunction of literals forms the *body*: the head is derived when all positive body literals can be proven and none of the negated ones can. The operator `not` implements *negation as failure* (NAF): `not b` holds when there is no derivation for  $b$ , expressing absence of evidence rather than explicit falsity and giving ASP a non-monotonic, *closed-world* semantics. Solutions are *answer sets*: self-consistent atom sets satisfying all rules. Solvers additionally support aggregates (`#count`, `#sum`) and optimization (`#minimize`, `#maximize`), making ASP well-suited as the unifying formalism across our framework's uncertainty analysis, rule learning, and measurement strategy.

### D. Inductive learning

Inductive Logic Programming (ILP) [14] aims to find a set of logical rules, called a *hypothesis*, that together with existing background knowledge explains a set of examples. In this work, we employ ILP in the context of Answer Set Programming, using the ILASP framework [15] and its scalable variant FastLAS [16].

A learning task is a tuple  $T = \langle B, M, E \rangle$ , where  $B$  is background knowledge (an ASP program),  $M$  is a *mode bias* defining the rule space, and  $E$  is a set of examples. Each example carries a *penalty* (the cost of not covering it), a partial interpretation (inclusions/exclusions), and a per-instance ASP *context*; examples with infinite penalty must be covered, while finite-penalty ones may be left uncovered at their cost. The learner minimizes a score combining hypothesis length with the penalties of uncovered examples. FastLAS [16] generalizes this with *domain-specific scoring functions* that decompose over individual rules, improving scalability. This fits our setting in two ways: finite penalties tolerate noisy examples proportionally to their frequency, and customizable scoring lets us encode domain-specific preferences (e.g.,

biasing toward conservative classifications) directly into the optimization criterion.

## III. FRAMEWORK OVERVIEW

This section defines the variable groups arising in benchmark campaigns, describes how raw data is prepared, and outlines how the pipeline stages interact; each stage is then detailed in its own section.

### A. Variable Structure

The variables of a benchmark campaign fall into three groups with distinct roles. (1) *Measured metrics* (CPU and memory utilization, throughput, packets per second, dropped packets, etc.) serve exclusively to derive each run's binary HEALTHY/UNHEALTHY outcome against predefined service-level thresholds. (2) *Workload requirements* (traffic type, user equipment count, and further demand-side parameters) characterize the service demand. (3) *Deployment configuration* (UPF count, core strength, memory capacity, etc.) describes the provisioned resources. The learning task operates on groups (2) and (3): given a workload requirement and a candidate deployment configuration, the learned rules predict whether the configuration can support the workload. Group (1) is consumed during data preparation but does not appear as an attribute in the learned rules.

### B. Data Preparation

Each campaign run produces time-series telemetry from which the measured metrics (and hence the binary outcome) are derived, while the workload and deployment variables are represented as qualitative levels. The same configuration-workload pair may be measured multiple times; repeated runs are aggregated per equivalence class while retaining the frequency of each outcome, which feeds both the uncertainty evaluation and the cost function of the learning stage (Section IV details the construction).

### C. Pipeline Stages

The pipeline comprises four stages (Fig. 1), each detailed later: *uncertainty evaluation* (Sec. IV) applies RST to the aggregated decision table, separating the lower approximations from the boundary region and reporting  $\gamma$ ; *performance characterization* (Sec. V) synthesizes deterministic rules for one decision class under the RST-derived constraints and cost function, with the complementary class as the NAF default; *validation* (Sec. VI) runs an ASP-encoded monotonicity check and, when a system model is available, traces persistent inconsistencies to candidate unmeasured variables; and *measurement strategy planning* prioritizes the next campaign (boundary resolution, monotonicity testing, coverage gaps), closing the loop with analysis-driven selection.

## IV. UNCERTAINTY EVALUATION

### A. Decision Table Construction

We now detail how the three variable groups enter the decision table. The measured metrics derive the binary decision

attribute only. Let  $M = \{m_1, \dots, m_k\}$  be the metrics of a run, each with a service-level threshold  $\tau_i$  and compliance condition (e.g.,  $m_i \leq \tau_i$  for drop rates,  $m_i \geq \tau_i$  for throughput). A run is UNHEALTHY if *any* metric violates its requirement:

$$d = \begin{cases} \text{HEALTHY} & \text{if all metrics satisfy their thresholds} \\ \text{UNHEALTHY} & \text{otherwise} \end{cases} \quad (1)$$

The workload and deployment variables form the attribute set. Categorical variables (e.g., traffic type) enter directly; numerical ones (e.g., UPF count, core strength) are mapped to qualitative levels for scale independence. For a numerical variable  $v$  with value  $x$  and thresholds  $\theta_{\text{low}}, \theta_{\text{high}}$ :

$$q(v, x) = \begin{cases} \text{LOW} & \text{if } x \leq \theta_{\text{low}} \\ \text{MEDIUM} & \text{if } \theta_{\text{low}} < x \leq \theta_{\text{high}} \\ \text{HIGH} & \text{if } x > \theta_{\text{high}} \end{cases} \quad (2)$$

with thresholds defined relative to the deployment's capacity range.

The placement of these landmarks directly governs the boundary region: wider levels merge numerically distinct configurations into the same class, enlarging the boundary and lowering  $\gamma$ , while tighter levels shrink it but leave fewer repeated observations per class. The thresholds are therefore an *engineering modeling decision* (whether derived from a capacity model or from automated clustering) and must reflect the underlying system model and analysis goal for the abstraction to remain meaningful.

Each run  $r_i$  produces a row in the decision table  $\mathcal{T} = (U, A \cup \{d\})$ , where  $U = \{r_1, \dots, r_n\}$  is the set of runs,  $A$  the qualitative attributes, and  $d$  the decision. The same configuration–workload pair may be executed *multiple times*, yielding rows with identical attribute vectors but potentially different decisions; this is the source of inconsistency that RST captures through the boundary region.

Consider two deployment variables, Core Strength and UPF Count  $\in \{L, M, H\}$ , and one workload variable, Traffic  $\in \{\text{mIoT}, \text{eMBB}, \text{std}\}$ . Suppose 8 runs cover 6 distinct configurations, one of which (Core=L, Count=M, eMBB) is measured twice with contradictory outcomes.

### B. Lower Approximations and Boundary Region

Applying the indiscernibility relation  $\text{IND}(A)$  groups repeated runs of the same configuration into shared equivalence classes. Let  $X_U = \{r \in U \mid d(r) = \text{UNHEALTHY}\}$ ; the resulting classes and their classification (with UH:H outcome counts) are shown in Table I.

The lower approximation of UNHEALTHY contains classes where *all* runs are unhealthy:  $\underline{A}(X_U) = E_1 \cup E_4$  (*certainly failing*). Symmetrically,  $\underline{A}(X_H) = E_3 \cup E_5 \cup E_6$  are certainly healthy, and the boundary region  $\text{BND} = \{E_2\}$  holds the single mixed-outcome configuration. The approximation quality at the class level is

$$\gamma_A(d) = \frac{|\{E_1, E_3, E_4, E_5, E_6\}|}{|\{E_1, \dots, E_6\}|} = \frac{5}{6} \approx 0.833, \quad (3)$$

TABLE I  
EQUIVALENCE CLASSES AND THEIR RST CLASSIFICATION

| Class | Core | Count | Traf. | UH:H | Region               |
|-------|------|-------|-------|------|----------------------|
| $E_1$ | L    | L     | mIoT  | 2:0  | $\underline{A}(X_U)$ |
| $E_2$ | L    | M     | eMBB  | 1:1  | BND                  |
| $E_3$ | M    | M     | std   | 0:1  | $\underline{A}(X_H)$ |
| $E_4$ | M    | L     | mIoT  | 1:0  | $\underline{A}(X_U)$ |
| $E_5$ | H    | H     | eMBB  | 0:1  | $\underline{A}(X_H)$ |
| $E_6$ | H    | L     | std   | 0:1  | $\underline{A}(X_H)$ |

i.e., one of six configurations yields inconsistent results across repeated runs.

These results directly constrain and parameterize rule learning. *Coverage*: rules for UNHEALTHY must fire on all configurations in  $\underline{A}(X_U)$ ; failing to do so is a *false negative*, a critical error in safety-relevant systems. *Exclusion*: rules must *not* fire on  $\underline{A}(X_H)$ ; firing here is a *false positive* (overly conservative but not dangerous). *Cost-driven decision*: configurations in BND are classified by an asymmetric cost function penalizing false negatives (weight  $\alpha$ ) more than false positives (weight  $\beta$ ), with the outcome frequencies (Table I) letting the cost reflect empirical evidence, where a 3:1 UH:H boundary class warrants a different penalty than a 1:1 split.

### C. Interpreting the Boundary Region

The boundary region is not merely a nuisance but an informative diagnostic signal with three possible causes: (i) *environmental noise*: behavior is non-deterministic at the chosen abstraction level, an irreducible inconsistency the cost function must tolerate; (ii) *too coarse quantization*: the thresholds of Eq. 2 group meaningfully different values, so refining them may resolve the inconsistency; and (iii) *missing variable*: an unobserved variable drives the differing outcomes, in which case a structural system model can identify candidate variables architecturally connected to the inconsistent component (Section VI). A large boundary region with low  $\gamma$  thus guides the evolution of the measurement campaign.

## V. PERFORMANCE CHARACTERIZATION

This section describes how inductive learning in ASP synthesizes deterministic classification rules from the decision table.

### A. One-Sided Learning with NAF Default

The framework learns rules for one decision class while the complementary class serves as the default via NAF: any configuration not explicitly classified by a learned rule is assigned the default class. The choice depends on safety requirements. The *conservative* (safety-first) option learns rules for HEALTHY with default UNHEALTHY, so a configuration is acceptable only if a rule explicitly supports it (unclassified configurations are treated as potentially failing). The *permissive* option is symmetric (learning UNHEALTHY with default HEALTHY), appropriate when over-provisioning costs outweigh under-provisioning risk. In both cases ASP's closed-world assumption makes the classification deterministic and complete. We present the conservative formulation (learning

HEALTHY), matching our 5G case study; the permissive case is symmetric.

### B. Mapping to an ILP Task

The decision table is mapped to an ILP task  $T = \langle B, M, E \rangle$ . The background knowledge  $B$  encodes the domain: the observed equivalence classes (as node identifiers) and any structural information constraining the rule space. The mode bias  $M$  defines the hypothesis space: the head mode declares the target predicate (e.g., `healthy(V)`) and the body modes declare the qualitative attribute predicates, both positive and negated, allowing rules of the form:

```
healthy(ID) :- attr_1(ID, val_1),
              not attr_2(ID, val_2),
              ... .
```

Negated body literals are essential: they let rules express conditions such as “healthy if core strength is *not* low,” which are more compact and general than enumerating positive combinations.

Each equivalence class produces weighted examples  $E$ . For a class  $E_i$  with  $n_h$  healthy and  $n_u$  unhealthy runs, the conservative formulation (learning HEALTHY) generates: if  $n_h > 0$ , a positive example with inclusion  $\{\text{healthy}(i)\}$  and penalty  $\propto n_h \cdot \alpha$  ( $\alpha$  = weight of false negatives, i.e. unnecessary rejection); if  $n_u > 0$ , a negative example with exclusion  $\{\text{healthy}(i)\}$  and penalty  $\propto n_u \cdot \beta$  ( $\beta$  = weight of false positives, i.e. wrongly accepting a failing configuration). The context encodes the class’s qualitative attribute values as ASP facts. The asymmetry  $\beta > \alpha$  encodes the safety requirement (wrongly accepting a failing configuration is penalized more heavily than rejecting a healthy one), and the permissive formulation uses the opposite ratio. Because penalties scale with observation counts, a class observed 100 times carries proportionally more weight than one observed twice.

For lower-approximation classes the asymmetry is irrelevant: a class in  $\underline{A}(X_H)$  has  $n_u = 0$  (only a positive example, which must be covered) and one in  $\underline{A}(X_U)$  has  $n_h = 0$  (only a negative example, which must not be violated). The asymmetric cost therefore operates primarily on the *boundary region*, where both outcomes coexist and the learner decides by the balance of evidence and the safety-driven penalty ratio. Beyond example penalties, the scoring function can add structural preferences via FastLAS rule-level penalties: a fixed penalty per rule (discouraging complex hypotheses), a per-literal penalty (preferring shorter rules), and custom domain-specific penalties.

### C. Connecting RST Constraints to the ILP Task

The RST constraints are translated through the penalty structure. Lower-approximation classes become hard constraints via infinite penalties ( $\infty$ ) on their positive or negative examples, forcing coverage or exclusion; boundary-region classes produce both example types with finite, frequency- and  $\alpha/\beta$ -derived penalties, letting the learner resolve the ambiguity by minimizing total cost. In practice the hard constraints can instead use very high finite penalties that dominate the boundary ones, trading a strict guarantee for robustness under

extreme noise. The approximation quality  $\gamma$  is a global diagnostic: high  $\gamma$  means a tightly constrained task with accurate rules, while low  $\gamma$  shifts weight onto the cost function, making rule quality depend more on the  $\alpha/\beta$  ratio and observation counts.

## VI. VALIDATION AND MEASUREMENT STRATEGY

The NAF default implicitly encodes a *monotonicity hypothesis*: under the same workload requirements, increasing deployment resources should never degrade system health, so if a weaker configuration passes, every stronger one should too. This can be violated for two reasons: *environmental noise* (background load fluctuations or scheduling jitter causing a stronger configuration to occasionally fail) or a *missing variable* that interacts with the resource dimensions and produces genuinely non-monotone behavior. The validation stage distinguishes the two by integrating the violation analysis with the RST boundary region classification.

### A. Monotonicity Checking

Let  $\mathcal{W}$  be the *workload requirement* variables (e.g., traffic type, UE count) and  $\mathcal{R}$  the *deployment configuration* variables representing provisionable resources (e.g., core strength, UPF count), whose qualitative levels are ordered ( $\text{LOW} \prec \text{MEDIUM} \prec \text{HIGH}$ ). Two configurations share the same workload profile,  $c_a \sim_{\mathcal{W}} c_b$ , iff they agree on every workload variable:

$$c_a \sim_{\mathcal{W}} c_b \iff \forall w \in \mathcal{W}: c_a(w) = c_b(w) \quad (4)$$

Given  $c_a \sim_{\mathcal{W}} c_b$ ,  $c_a$  *resource-dominates*  $c_b$  ( $c_a \succ_{\mathcal{R}} c_b$ ) if every resource variable is at least as high and at least one is strictly higher:

$$c_a \succ_{\mathcal{R}} c_b \iff \begin{aligned} &(\forall r \in \mathcal{R}: c_b(r) \preceq c_a(r)) \\ &\wedge (\exists r \in \mathcal{R}: c_b(r) \prec c_a(r)) \end{aligned} \quad (5)$$

A pair with  $c_a \succ_{\mathcal{R}} c_b$  is a *monotonicity violation* if  $c_b$  is HEALTHY and  $c_a$  UNHEALTHY: a stronger configuration fails where a weaker one passes under identical workload.

The check is encoded declaratively in ASP in four layers. **(1) Stratification:** measured configurations are partitioned by workload profile and a componentwise partial order (Eq. 5) is established over resources, so monotonicity is never evaluated across incomparable workloads. **(2) Violation detection:** within each stratum, the encoding finds violating pairs and records the differing resource dimensions; a single differing dimension is a strong candidate for investigation. **(3) Inter-mediate generation:** for a violation between a stronger  $c_s$  and weaker  $c_w$ , it generates the unmeasured lattice points  $\{c_x \mid \forall r: c_w(r) \preceq c_x(r) \preceq c_s(r)\} \setminus \{c_w, c_s\}$  as *suggested measurements* that narrow the gap between known HEALTHY and UNHEALTHY regions. **(4) Boundary classification:** a *direct boundary* separates an UNHEALTHY  $c_a$  and HEALTHY  $c_b$  with  $c_a \succ_{\mathcal{R}} c_b$  and no measured configuration between them; a *minimal boundary* further requires differing in exactly one dimension by one level, the finest granularity at which the transition can be localized.

### B. Integration with Rough Set Analysis

A violation's confidence depends on the RST membership of the two configurations. If both lie in the lower approximation of their decision classes, it is *confirmed non-monotonicity*: every repeated run agreed, so the contradiction cannot be measurement variability and the NAF assumption is invalid in that region. If either lies in the boundary region, the violation is *possibly noise-induced* and needs more measurements before concluding monotonicity fails. In the ASP encoding, the boundary classification predicate from the uncertainty stage is imported and each violation annotated accordingly; boundary-involving violations are flagged and their Layer 3 suggested measurements receive elevated priority.

### C. System Model and Root Cause Analysis

Violations and the boundary region identify *where* the data is inconsistent, not *why*. When violations cluster around a resource variable, the framework can incorporate a structural *system model* encoded in ASP to find the unmeasured factor causally linked to it. The model represents the architecture as ASP facts: *components and containment* (network functions, bridges, routing daemons and their topology); *connections and data paths* (the user-plane path, distinguished from control-plane connections because bottlenecks typically occur on it); and *observable variables* (configurable parameters and metrics per component, annotated with whether they are currently measured; the gap defines the candidate missing variables). Beyond connectivity, the model encodes directed *causal influence* relations recording that a resource/workload variable increases the load on a component, with the mechanism (e.g., more UPF instances cause routing overhead on a BGP pod maintaining iBGP sessions, and more traffic on a shared bridge). These are verifiable engineering assertions, not statistically inferred causal claims. The model marks *shared resources* and derives *scaling-sensitive* components and *potential bottlenecks* (scaling-sensitive, shared, and lacking a measured capacity variable).

The root cause analysis combines violations with the causal model in three steps: *pattern analysis* groups violations by differing resource variable and identifies the *dominant violation variable*; *causal trace* traverses influence relations from that variable to downstream components, flagging scaling-sensitive shared components without a measured capacity variable as high-confidence root causes (medium if not shared); and *aggregate diagnosis* scores each component by the number of violations tracing to it, yielding the top-ranked root cause. This produces two recommendation types: *extend the learning task*, adding a configurable parameter on the bottleneck as a new attribute may separate previously indistinguishable boundary configurations and improve  $\gamma$ ; or *refine the decision attribute*, adding a metric (e.g., routing convergence time alongside packet drops) that distinguishes failure modes yields cleaner labels and may shrink the boundary. The operator decides which interventions are feasible.

### D. Measurement Strategy

The uncertainty evaluation and monotonicity analysis jointly drive a targeted strategy with three priority tiers: *boundary resolution* (additional runs for violation endpoints in the boundary region, with extended telemetry when the system model identifies candidate variables, simultaneously testing the violation and reducing the boundary); *monotonicity testing* (intermediate lattice configurations between known HEALTHY/UNHEALTHY points, prioritizing those near a direct or minimal boundary); and *coverage gaps* (configurations classified solely by NAF, the lowest priority). This replaces random or exhaustive testing with a minimal, analysis-driven selection, closing the loop to the next campaign round.

## VII. CASE STUDY

We evaluate the framework on measurement data from a 5G Core testbed, showing how it guides the expansion of a measurement campaign and diagnoses the sources of uncertainty that emerge as the configuration space grows.

### A. Dataset preparation

The dataset was built from custom traffic-generator logs and Prometheus time-series data. Preprocessing normalized resource measurements and annotated them with traffic and deployment metadata, then categorized observations into qualitative states via K-means clustering combined with rule-based heuristics (e.g., UNHEALTHY when the packet drop count exceeds a threshold).

### B. Baseline Measurements

The initial campaign covers 9 configurations with core strength fixed at LOW, varying UPF count, UE count, and traffic type across all three levels. After aggregation, the dataset has 10 equivalence classes (one configuration appears with both outcomes) over 1 490 runs. The approximation quality is  $\gamma = 0.500$ : five classes in  $\underline{A}(X_U)$ , three in  $\underline{A}(X_H)$ , and two in the boundary region (a single configuration, UPF=HIGH, UE=LOW, eMBB, observed healthy in 102 runs and unhealthy in 47). No monotonicity violations are detected: within each workload stratum, the NAF assumption holds.

The framework produces stable results: eleven of twelve tested penalty configurations yield the same rules:

```
healthy(V) :- traffic_type(V, embb),
               upf_count(V, high).
healthy(V) :- traffic_type(V, standard).
```

These state that with LOW core strength, the system is healthy when the traffic is eMBB with high UPF count, or when the traffic is standard; all mMTC and low-count eMBB configurations default to UNHEALTHY via NAF. The framework identifies 18 unmeasured combinations (Priority 3), including all MEDIUM/HIGH core strength configurations; no boundary-resolution or monotonicity-testing measurements are needed. The clear recommendation is to extend the campaign to stronger deployments.

### C. Expanded Campaign

Following this recommendation, the campaign is extended with 10 additional configurations covering MEDIUM and HIGH core strength. Combined with the baseline, the full dataset has 22 equivalence classes over 3724 runs. The approximation quality *drops* to  $\gamma = 0.455$ . This is not a regression but a diagnostic signal: the new configurations introduce inconsistencies invisible in the LOW-only baseline. Ten classes are now in  $\underline{A}(X_U)$  and four in  $\underline{A}(X_H)$  (up from three), but eight land in the boundary region (up from two). Although the boundary grows from 20% to 36%, the number of certainly classified configurations increases in both classes, enabling more expressive rule learning.

Six monotonicity violations are detected, all classified as possibly noise-induced (at least one endpoint in the boundary region). The dominant violation variable is UPF count (4 of 6); core strength accounts for 3. For example, (medium core, medium count, medium UE, standard) is UNHEALTHY while the weaker (medium core, LOW count, medium UE, standard) is HEALTHY. The learned rules now capture interactions across all four variables:

```
healthy(V) :- core_strength(V, medium),
              upf_count(V, high),
              not traffic_type(V, miot).
healthy(V) :- not core_strength(V, low),
              ue_count(V, medium).
healthy(V) :- core_strength(V, low),
              not ue_count(V, high),
              not traffic_type(V, embb).
```

The third rule is notable: LOW core strength is not inherently UNHEALTHY, it suffices when UE count is not high and traffic is not eMBB, a nuance invisible in the first campaign where core strength was fixed.

The root cause analysis traces the dominant violation variable (UPF count) through the system model, surfacing two high-confidence potential bottlenecks: the *BGP routing pod*, a shared resource servicing all UPF instances whose route-advertisement overhead grows with UPF count (fixed capacity, not in the attribute set); and the *shared OVS bridge* (N6 interface), through which all instances communicate, with fixed PMD core and queue allocation. The framework thus recommends (1) adding OVS bridge capacity or BGP session indicators as new attributes to separate currently indistinguishable boundary configurations, and (2) measuring BGP convergence time and OVS drop rates alongside packet drops to distinguish routing-induced from capacity-induced failures. This diagnosis was independently corroborated by domain-expert analysis, which identified BGP routing overhead as a likely contributor to the non-monotone behavior.

The framework produces a prioritized plan for the next round: (1) re-measure the 4 boundary-region configurations involved in violations with extended BGP/OVS telemetry; (2) no intermediate configurations are needed (all lattice points between violation endpoints are already measured); (3) 36 unmeasured combinations, primarily LOW core strength with uncovered workloads and MEDIUM core strength with mIoT traffic.

TABLE II  
COMPARISON ACROSS THE THREE ANALYSES

|                          | Baseline | Expanded | Insufficient |
|--------------------------|----------|----------|--------------|
| Equivalence classes      | 10       | 22       | 28           |
| Approx. quality $\gamma$ | 0.500    | 0.455    | 0.286        |
| Certainly healthy        | 3        | 4        | 0            |
| Boundary region          | 2        | 8        | 20           |
| Violations               | 0        | 6        | 22           |
| Rule stability           | stable   | stable   | unstable     |
| System model             | n/a      | BGP, OVS | BGP, OVS     |

### D. Contrast: Insufficient Variable Selection

To illustrate behavior when the attribute set is fundamentally insufficient, we analyze a third variable selection from the same data: core strength (LOW/MEDIUM), UPF count (all three levels), and traffic type, with UE count fixed at 1024. This yields 28 equivalence classes over 10727 runs. The approximation quality drops to  $\gamma = 0.286$ : no class is in  $\underline{A}(X_H)$  (every healthy configuration was also observed unhealthy), 20 of 28 classes lie in the boundary, and 22 monotonicity violations are detected (all noise-induced, UPF count dominant in 18). The learned rules are highly unstable across penalty settings. With balanced penalties the learner produces a single rule covering one class:

```
healthy(V) :- core_strength(V, medium),
              upf_count(V, high),
              not traffic_type(V, miot).
```

With aggressive settings ( $\alpha:\beta = 3:1$ ), broader rules emerge that cover most boundary-region classes:

```
healthy(V) :- not core_strength(V, low),
              traffic_type(V, embb).
healthy(V) :- not core_strength(V, low),
              not upf_count(V, low),
              traffic_type(V, standard).
```

Several configurations produce an empty hypothesis: the NAF default of UNHEALTHY is optimal because healthy runs constitute only  $\sim 10\%$  of observations. The framework correctly diagnoses this: the low  $\gamma$  signals that the current variables cannot explain the outcomes, and the system model points to the same bottlenecks (BGP pod, OVS bridge). Without UE count as a differentiating variable, configurations behaving differently under different loads are merged into one class, inflating the boundary.

### E. Discussion

The *iteration* from baseline to expanded campaign shows the measurement strategy driving productive data collection: the baseline's Priority 3 list motivated adding MEDIUM/HIGH core strength, revealing new interactions. The *drop* in  $\gamma$  from 0.500 to 0.455 illustrates that more data does not automatically mean better models: the expansion surfaced inconsistencies, and the root-cause analysis (BGP stack, OVS bridge) gives a concrete path forward. The *contrast* ( $\gamma = 0.286$ ) shows graceful degradation: rather than silently producing unreliable rules, the framework flags the problem and guides the operator toward a better variable selection. Table II summarizes the three analyses.

## VIII. CONCLUSIONS AND FUTURE WORK

We presented an integrated framework for deterministic, interpretable diagnostics of highly configurable systems. Operating on qualitative abstractions, the framework combines RST with ILP in a unified ASP encoding. RST quantifies measurement uncertainty and constrains the rule learner, a monotonicity check validates resource-scaling assumptions with RST-based confidence classification, and a structural system model traces root causes of violations to architectural bottlenecks. The case study on a 5G testbed demonstrated that the framework produces stable rules when the data is informative, diagnoses the sources of uncertainty when it is not, and provides actionable recommendations for the next measurement campaign.

### A. Implications

For theory, the framework shows that the rough-set boundary region can serve not only as an uncertainty indicator but as a first-class parameter shaping learning constraints, confidence classification, and experimental design within a single formal loop. For practice, it offers operators a deterministic, auditable classification that remains valid across hardware changes without retraining, replacing exhaustive testing with a targeted strategy that reduces campaign cost while maintaining conservative safety guarantees.

### B. Limitations and Threats to Validity

The framework is evaluated on a single 5G UPF testbed, so the reported rules and  $\gamma$  values illustrate the methodology rather than constitute general findings. We address this by construction: qualitative landmarks are defined relative to each deployment's capacity, making rules hardware-independent, and the pipeline applies to any configurable system with repeatable benchmarks, though cross-environment transfer remains to be shown empirically. The results depend on the placement of the quantization landmarks (Eq. 2) and the SLA thresholds defining the decision attribute, which is precisely why the framework represents residual noise explicitly rather than assuming deterministic outcomes. The few equivalence classes limit independent train/test evaluation, so we report diagnostic behavior (rule stability,  $\gamma$ , boundary growth) rather than point accuracy.

### C. Future Work

Alternative quantization methods (entropy-based, data- or expert-driven) could improve discretization quality where a large boundary region stems from poorly calibrated thresholds. The system model's causal annotations could be strengthened toward interventional semantics, enabling predictive reasoning about adding variables before measuring them. Finally, a systematic comparison with alternative interpretable classifiers (decision trees, Boolean Rule Column Generation) would further establish the framework's positioning.

## ACKNOWLEDGEMENTS

Project no. 2025-2.1.2-EKÖP-KDP-2025-00005 has been implemented with the support provided by the Ministry of Cultvia Innovation of Hungary from tthat reflectsl Research, Development and Innovation Fund, financed under the EKÖP\_KDP-25-1-BME-27 funding scheme.

During the preparation of this work, the authors used Claude to assist with drafting and refining manuscript text. Gemini and Claude were also used for generating and optimizing Python scripts used in the experimental simulations. The authors take full responsibility for the content of this publication.

## REFERENCES

- [1] N. Siegmund, A. Grebhahn, S. Apel, and C. Kästner, "Performance-influence models for highly configurable systems," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM, 2015, pp. 284–294. [Online]. Available: <https://doi.org/10.1145/2786805.2786845>
- [2] J. Guo, D. Yang, N. Siegmund, S. Apel, A. Sarkar, P. Valov, K. Czarnecki, A. Wasowski, and H. Yu, "Data-efficient performance learning for configurable systems," *Empirical Software Engineering*, vol. 23, no. 3, pp. 1826–1867, 2018. [Online]. Available: <https://doi.org/10.1007/s10664-017-9573-6>
- [3] P. Jamshidi, M. Velez, C. Kästner, and N. Siegmund, "Transfer learning for performance modeling of configurable systems: An exploratory analysis," in *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, 2017, pp. 497–508. [Online]. Available: <https://doi.org/10.1109/ASE.2017.8115661>
- [4] C. Jiang and Y. Duan, "Elasticity unleashed: Fine-grained cloud scaling through distributed three-way decision fusion with multi-head attention," *Information Sciences*, vol. 660, p. 120127, 2024.
- [5] A. Campagner and D. Ciucci, "Three-way learnability: A learning theoretic perspective on three-way decision," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 30, 2022, pp. 243–246.
- [6] S. Liu, X. Ma, Y. Jia, and Y. Liu, "An energy-saving task scheduling model via greedy strategy under cloud environment," *Wireless Communications and Mobile Computing*, vol. 2022, p. 8769674, 2022.
- [7] S. Widz and S. Stawicki, "Generalized majority decision reducts," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)*, ser. Annals of Computer Science and Information Systems, vol. 8, 2016, pp. 165–174.
- [8] H. Naghavipour, M. Y. I. B. Idris, T. K. Soon, R. B. Salleh, and A. Gani, "Hybrid metaheuristics using rough sets for QoS-aware service composition," *IEEE Access*, vol. 10, pp. 112 609–112 626, 2022.
- [9] K. D. Forbus, "Chapter 9 qualitative modeling," in *Handbook of Knowledge Representation*, ser. Foundations of Artificial Intelligence, F. van Harmelen, V. Lifschitz, and B. Porter, Eds. Elsevier, 2008, vol. 3, pp. 361–393. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S157465260703009X>
- [10] Z. Pawlak, "Rough sets," *International journal of computer & information sciences*, vol. 11, no. 5, pp. 341–356, 1982.
- [11] V. Lifschitz, *Answer set programming*. Springer Berlin, 2019.
- [12] F. Calimeri, W. Faber, M. Gebser, G. Ianni, R. Kaminski, T. Krennwallner, N. Leone, M. Maratea, F. Ricca, and T. Schaub, "ASP-core-2 input language format," *Theor. Practice Logic*, vol. 20, no. 2, pp. 294–309, 2019.
- [13] M. Gelfond and V. Lifschitz, "The stable model semantics for logic programming," in *Proceedings of the Fifth International Conference and Symposium on Logic Programming*. MIT Press, 1988, pp. 1070–1080.
- [14] S. Muggleton, "Inductive logic programming," *New generation computing*, vol. 8, no. 4, pp. 295–318, 1991.
- [15] M. Law, A. Russo, and K. Broda, "The ilasp system for inductive learning of answer set programs," *arXiv preprint arXiv:2005.00904*, 2020.
- [16] M. Law, A. Russo, E. Bertino, K. Broda, and J. Lobo, "Fastlas: Scalable inductive logic programming incorporating domain-specific optimisation criteria," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 03, 2020, pp. 2877–2885.

# Acceptance of Generative AI in Higher Education in Software Engineering - A Systematic Literature Review

Veronika Vasileva, Thomas Reiser, Jan Jürjens

0009-0006-2024-4653, 0009-0007-5452-7800, 0000-0002-8938-0470

University of Koblenz, Institute for Software Technology, Universitätsstr. 1, 56070 Koblenz, Germany

Email: vvasilevaltreiserljuerjens@uni-koblenz.de

**Abstract—Objective.** This Systematic Literature Review (SLR) examines how predictors from established technology acceptance models influence the adoption of Generative AI (GenAI) in higher education among students and lecturers. It aims to identify key acceptance factors, with a particular focus on Software Engineering education.

**Methods.** A rigorous literature review process was applied, including systematic search, selection, and quality assessment of studies. Clearly defined inclusion, exclusion, and quality criteria ensured the use of validated findings. Acceptance factors were synthesized and compared for both students and lecturers.

**Results.** 31 studies (2020–2025) were analyzed. Performance Expectancy (PE), Perceived Trust (PT), Habit (HT), and Attitude Toward Using (ATU) show strong positive effects on Behavioral Intention (BI), while BI and HT are the strongest predictors of Use Behavior (UB). Effects of Effort Expectancy (EE), Social Influence (SI), Facilitating Conditions (FC), and Hedonic Motivation (HM) vary across contexts.

**Conclusion.** Tailored implementation strategies for GenAI in higher education are essential. Contextual and individual factors, such as digital competence and perceived ethics, should be emphasized. Research on acceptance of GenAI in Software Engineering education remains limited and requires more specific future studies.

**Index Terms—**Generative AI, Acceptance, Higher Education, Software Engineering, Systematic Literature Review

## I. INTRODUCTION

ARTIFICIAL INTELLIGENCE (AI) is transforming daily work by automating tasks such as writing, summarizing, and reporting. Generative AI (GenAI), a subset of Artificial Intelligence (AI), generates new content, while general AI focuses on analysis and prediction. In this research, we focus on GenAI. Large Language Models (LLMs) as an example of GenAI are used in chatbots like ChatGPT. Since ChatGPT's release in 2022, GenAI has become widely used, alongside other LLMs like DeepSeek and Gemini.

GenAI is increasingly used in higher education. A survey in [1] shows different student usage, with many students using GenAI at least occasionally. Concerns include accuracy, transparency, privacy, ethics, overreliance, job risks, and legal uncertainty [1].

Despite this, perceptions are largely positive and many students intend continued use due to benefits such as personalized learning, content visualization, and support for writing,

brainstorming, and literature analysis [1], [2], [3], [4]. Thus, GenAI is expected to remain important in academia.

Technology acceptance models identify factors influencing GenAI adoption in universities and support effective, responsible use in education [5].

However, Software Engineering (SE) education has not been specifically addressed in prior GenAI acceptance research. Existing studies include IT-related samples but do not differentiate results by discipline [6], [7], [8].

This SLR identifies predictors of GenAI acceptance among students and lecturers in higher education. By synthesizing prior studies across contexts, it highlights general effects and research gaps, supporting future studies, especially in SE education, and informing research and practice.

## II. RESEARCH QUESTIONS

Our SLR addresses the following three research questions:

*RQ1: Which factors and predictors determine the acceptance of GenAI in higher education by students and by lecturers?*

The objective is to collect and consolidate findings from previous acceptance studies on AI in higher education to gain a comprehensive view of the state of the research in this area.

*RQ2: Which acceptance factors for AI in SE education were addressed?*

Based on the results for RQ1, the objective is to report on acceptance factors and predictors of AI and GenAI relevant to the field of SE education.

## III. BACKGROUND AND RELATED WORK

This section summarizes technology acceptance models used in prior studies on GenAI in higher education to support hypothesis development. It introduces key predictors and dependent variables for explaining and predicting technology acceptance. Related studies within the scope of this SLR are summarized, and our study is distinguished from them.

### A. Models for explaining the acceptance and use of information technology

Acceptance describes how individuals or organizations adopt and use technologies to achieve goals [9]. This section outlines key models of technology adoption.

Acceptance models explain and predict intention and usage, with Behavioral Intention (BI) established as a key predictor of use [10], [11].

**Theory of Planned Behavior (TPB)**, based on Theory of Reasoned Action (TRA), predicts behavior in contexts without full volitional control by adding Perceived Behavioral Control (PBC). Its core predictors of *intention* are *Attitude Toward Behavior (ATB)*, *Subjective Norm (SN)*, and *Perceived Behavioral Control (PBC)*, which positively influence behavior [10], [12], [11].

**Technology Acceptance Model (TAM)** explains acceptance using *Perceived Usefulness (PU)* and *Perceived Ease of Use (PEOU)* [13], [14], [15], [16], [17], [18]. PU influences *Behavioral Intention (BI)* and *Attitude Toward Using (ATU)*, while PEOU affects PU and ATU. BI predicts *Actual Usage (AU)* [12], [15], [16], [17], [11], [19].

**Technology Acceptance Model 2 (TAM2)** extends TAM by adding social (*Subjective Norm (SN)*, *Image (IM)*) and instrumental (*Job Relevance (JR)*, *Output Quality (OQ)*, *Result Demonstrability (RD)*) factors<sup>1</sup> and moderators *Experience (EXP)* and *Voluntariness (VO)* [17]. It removes ATU from the original TAM [13], [14], [15], [17].

**Innovation Diffusion Theory (IDT)** explains innovation diffusion in social systems. For acceptance, it includes *Relative Advantage (RA)*, *Compatibility (CB)*, *Complexity (CX)*, *Trialability (TRI)*, and *Observability (OB)* as predictors of *Attitude Toward Using (ATU)*, leading to *Behavioral Intention (BI)* and *Actual Usage (AU)* [20], [21].

**Unified Theory of Acceptance and Use of Technology (UTAUT)** integrates models such as TAM, TPB, and IDT to explain technology use in organizations. It identifies *Performance Expectancy (PE)*, *Effort Expectancy (EE)*, *Social Influence (SI)*, and *Facilitating Conditions (FC)* as predictors of *Behavioral Intention (BI)*, influencing *Use Behavior (UB)* [11]. Moderators are also included [15], [21], [17], [11], [19].

**UTAUT2** and **UTAUT3** extend UTAUT with *Hedonic Motivation (HM)*, *Price Value (PV)*, *Habit (HT)*, and *Personal Innovativeness (PI)*. These influence BI, while HT, FC, and PI affect UB. These models address consumer contexts and retain moderators such as age, gender, EXP, and VO [11], [22], [7], [23].

Some studies also apply **Protection Motivation Theory (PMT)**, focusing on risk and coping appraisal [24], [25].

## B. Related Work

ACOSTA-ENRIQUEZ et al. examined factors influencing acceptance and use of AI applications in universities using Unified Theory of Acceptance and Use of Technology 2 (UTAUT2). They analyzed 50 publications (2018–2023) from databases such as Scopus, Web of Science, ScienceDirect (SCIENCE), and ProQuest following PRISMA guidelines. Results show PE and HM as the strongest predictors of Behavioral Intention (BI). For students, EE and SI are also

relevant, while for faculty, PE and FC dominate. Adoption varies by context: in Science, Technology, Engineering, and Mathematics (STEM), AI adoption is higher, mainly driven by PE, whereas in non-STEM fields, SI and FC are more important. EE is more relevant for undergraduates, while PE is stronger for graduate students. The authors conclude that tailored strategies are needed for different groups [26].

Our study extends this work by including literature up to November 2025 and additional databases such as ACM Digital Library (ACM) and IEEE Xplore (IEEE) [26]. It focuses on LLM-based chatbots for students as a subset of GenAI. Unlike the prior study, which relies on UTAUT2, this work integrates multiple models and synthesizes shared constructs [26]. While limitations of UTAUT2 are acknowledged [26], this study uses UTAUT as a central framework, incorporating models such as TAM, TPB, and IDT [11]. This enables inclusion of additional predictors like PT, PI, and Anxiety (ANX). Furthermore, this study applies quality criteria for hypothesis models to ensure statistical validity, whereas ACOSTA-ENRIQUEZ et al. mainly rely on content and formal criteria [26].

## IV. RESEARCH METHODOLOGY

Our SLR follows guidelines by KITCHENHAM [27] to ensure rigor and reduce bias. The process consists of four steps: (i) initial search, (ii) systematic search with title and abstract screening, (iii) full-text screening, and (iv) backward snowballing.

### A. Search Strategy and Inclusion Criteria

Three databases are selected: IEEE Xplore (IEEE)<sup>2</sup>, ACM Digital Library (ACM)<sup>3</sup>, and ScienceDirect (SCIENCE)<sup>4</sup>. An initial search with the string *Acceptance of AI in higher education* verifies sufficient literature and supports keyword extraction for the systematic search. We selected and combined generic keywords in order not to narrow the systematic search too much. The resulting search string was:

“acceptance” OR “UTAUT” OR “TAM” AND (“Higher Education”) AND (“Artificial Intelligence”) AND (“Structural Equation Modeling”).

We restrict studies to 2020–2025 and English/German to ensure relevance. These limits are applied directly in IEEE, ACM, and SCIENCE. The timeframe also covers the emergence of major LLM-based tools.

Content-related criteria ensure focus on higher education (ICC1), acceptance (ICC2), and relevant participants (ICC3). ICC4.1 targets GenAI including LLM-based tools, while ICC4.2 includes broader AI-based tools for faculty to increase coverage. To ensure comparability, studies must use established acceptance models (ICC5), enabling synthesis across models. UTAUT serves as the central framework, integrating models such as TRA, TPB, TAM, Motivational Model (MM), Combined TAM and TPB (C-TAM-TPB), Model of PC Utilization (MPCU), IDT, and Social Cognitive Theory (SCT)

<sup>2</sup><https://ieeexplore.ieee.org>, all URLs last visited 2026-04-15

<sup>3</sup><https://dl.acm.org/>

<sup>4</sup><https://www.sciencedirect.com/>

<sup>1</sup>The factors IM, JR and RD are not considered in the studies included in this SLR.

[11], including extensions like TAM2, UTAUT2, and Unified Theory of Acceptance and Use of Technology 3 (UTAUT3) [17], [19]. Some models (e.g., MM, C-TAM-TPB, MPCU) were not identified in the final sample.

Outcome constructs such as BI and UB are required to assess acceptance [11]. Data collection is typically survey-based using Likert scales. Finally, Structural Equation Modeling (SEM), particularly Partial Least Squares Structural Equation Modeling (PLS-SEM), is required for hypothesis testing and analysis of relationships between constructs [9].

### B. Quality Assessment

Quality measures include *convergent validity and reliability*, *discriminant validity*, and *predictive power* [28].

Convergent validity and reliability require Composite Reliability (CR), Cronbachs Alpha (CA) > 0.70, Factor Loadings (FLs) > 0.70, and Average Variance Extracted (AVE) > 0.50 [28], [29]. Items (survey statements on GenAI/AI) measure constructs via Likert scales. Discriminant validity ensures constructs remain distinct. It is tested using the Fornell-Larcker criterion, Heterotrait-Monotrait Ratio of Correlations (HTMT) (< 0.85), and cross-loadings, where item loadings must exceed cross-loadings [30], [31], [28], [29]. Predictive power is assessed using  $R^2$  (> 0.10), indicating explanatory strength. Values above 0.75, 0.5, and 0.25 are classified as substantial, moderate, and weak [32], [33]. Studies must meet reliability limits, at least one discriminant validity criterion, and  $R^2$  requirements. Studies with missing results or textual inconsistencies in the results are excluded.

### C. Study Selection Process

The SLR was conducted in November 2025; later studies were not considered.

Using the search string, 876 studies were identified (26 IEEE, 227 ACM, 623 SCIENCE). Limiting SCIENCE and ACM to the first 100 results reduced this to 226.

After selecting only sources where the full text was available and asserting the correct article type (journal, conference, or research paper) and removing duplicates, 126 titles and abstracts were screened using the remaining criteria. 51 studies were excluded, based on the first violated criterion.

The remaining 75 studies were assessed in full text using inclusion, exclusion, and quality criteria. Another 49 studies were excluded, leaving 26 studies for the next step.

Backward snowballing applied the same criteria to references of the 26 studies. Five additional studies from [7] were included.

In total, 31 studies were identified: 26 on students, 2 on lecturers, and 3 on both. Of these 31 studies three were published in 2023, 17 in 2024 and eleven in 2025.

### D. Data Extraction

Relevant data were extracted using a template with six sections: *Abstract*, *Hypotheses*, *Methodology*, *Results*, *Discussion*, and *Limitations*.

The *Abstract* summarizes the study. *Hypotheses* list constructs and concepts, enabling synthesis across studies.

*Methodology* captures timing, sample, and demographics. *Results* include measurement model quality and path analysis.

The *Discussion* ranks predictors by strength and significance, explaining results and implications. *Limitations* report methodological constraints.

Tags categorize studies by target group (students/lecturers), country, and technology (e.g., ChatGPT, DeepSeek), enabling grouping.

## V. RESULTS

Of the 31 studies, 29 examine acceptance from the students' perspective, while three of the 31 studies include both lecturers and students in their sample. Only the study [6] examines acceptance exclusively from the perspective of faculty members working in higher education institutions. The study [29] includes both students and faculty in its sample, but is considered only for faculty because it violates inclusion criterion *ICC4.1*.

Studies were conducted in USA (1), Europe (5), Africa (1), Western Asia (5), and East Asia (23). The total number is higher than the number of included papers, as some studies were conducted in multiple countries. The distribution highlights a lack of research on the acceptance of GenAI outside of Asia.

We observed a terminology mix in various publications: often, the term AI is used while GenAI or LLMs were in focus. In addition, lecturers are sometimes called teachers, instructors, or university staff.

### A. Results for Students

This section summarizes the results for acceptance of GenAI in higher education by students. We synthesized the predictor effects on ATU, BI, and UB. Considering different cultural and contextual settings enables assessing generalizability beyond limitations of sample size and selection [34], [35], [7], [36], [37], [28], [38], [3], [39], [8], [40], [33], [41], [4].

1) *Performance Expectancy (PE)*: PE aggregates performance-related constructs (PU, RA, PE) [12], [11], [15]. Most studies find a significant positive effect on BI, often among the strongest predictors [35], [42], [7], [36], [37], [43], [44], [38], [45], [28], [46], [3], [39], [8], [47], [40], [33], [41], [4], [34]. PE also affects UB [48]. Some studies report no effect [2], [49], [50], and results may vary by country [51]. Effects are significant for both cognitive and affective appraisal [34].

2) *Effort Expectancy (EE)*: EE aggregates ease-of-use constructs [12], [11], [15]. Several studies show a positive effect on BI, often weak [36], [50], [45], [47], [40], [33], [41], [2]. Effects vary by subgroup (e.g., STEM) and also influence ATU [35], [52], [53]. Many studies find no effect [49], [7], [37], [43], [44], [38], [51], [8], [4]. Overall, results are inconsistent and context-dependent.

3) *Social Influence (SI)*: SI captures social influence [12], [11], [17]. Many studies find a positive effect on BI [2], [36], [44], [46], [53], [8], [47], [40], [33], [4], ranging from strong to weak. Others find no effect [49], [7], [37], [43], [50], [38], [51]. Thus, its influence is context-dependent.

4) *Facilitating Conditions (FC)*: FC reflects external support factors [12], [11]. Many studies show a positive effect on BI, sometimes strongest [35], [44], [28], [46], [3]. Others report weak, no, or negative effects [49], [7], [37], [38], [45], [51], [8], [40], [47]. Effects on UB are similarly mixed [44], [38], [40], [4], [7]. Overall, findings are inconclusive.

5) *Hedonic Motivation (HM)*: HM represents enjoyment [11]. Most studies report a positive effect on BI, often strong [7], [43], [44], [50], [40], [33]. Some studies find no effect [49], [37], [38], [46]. Differences may relate to cognitive vs. affective factors [34].

6) *Habit (HT)*: HT (habit) shows a strong positive effect on BI, often the strongest predictor [49], [7], [38], [40]. Only one study reports no effect [50]. Effects on UB are similar but weaker [7], [38], [40].

7) *Price Value (PV)*: Only two studies examine PV, both finding positive effects on BI, from weak to strong [49], [40], likely due to cultural differences.

8) *Personal Innovativeness (PI)*: Most studies find a positive, usually weak effect of PI on BI [43], [38], [8], [40], [33]. Some report stronger effects [8], while few find none [7], [46].

9) *Anxiety (ANX)*: ANX (technology fear) includes related constructs like Perceived Risk (PR) and Perceived Severity (PSEV) [36], [28], [39]. Results range from negative to non-significant effects on BI [4], [3]. Findings vary across regions [39], [36], indicating strong contextual influence.

10) *Perceived Trust (PT)*: PT (trust) includes reliability and privacy [42], [7], [43]. All studies find significant positive effects on BI or ATU, often strong [42], [7], [43], [46], [53].

11) *Attitude Toward Using (ATU)*: ATU is consistently one of the strongest predictors of BI [34], [2], [52], [50], [53], [41], and sometimes affects UB [54], [55]. Both affective and cognitive components matter [34].

12) *Behavioral Intention (BI)*: The majority, i.e., 14 out of 15 studies that research the relationship, approve that BI strongly predicts UB [2], [7], [36], [37], [44], [50], [38], [28], [3], [39], [47], [40], [33], [4]. Only one study finds no effect [52].

### B. Results for Lecturers

For lecturers, we considered any AI tool rather than focusing on GenAI and LLMs, due to the limited number of studies on lecturer acceptance.

The study PANGGABEAN et al. builds on HSU et al. and is regarded a single study only [39], [28]. Four studies combine teacher and student perspectives rather than separating them [28], [51], [39], [29], making the results more reflective of students, who form the majority of participants.

The results of [51], [28], [39] were presented in subsection V-A, so only [6] and [29] are summarized here.

*Performance Expectancy (PE)* and *Effort Expectancy (EE)* both show significant positive effects on BI in [29], [6], though the strength of EE varies from weakest [6] to second-strongest [29].

*Social Influence (SI)* has the weakest significant effect in [29], and no effect in [6].

*Facilitating Conditions (FC)* shows a significant positive effect on UB in [29] and on BI in [6].

[29] additionally examines data protection and PT. Data protection negatively affects BI but positively affects UB. PT shows no effect on UB in both studies, while it positively affects BI in [29] and moderates the BI-UB relationship in [6].

[6] further examines HM, HT, and PI. *Hedonic Motivation (HM)* and *Habit (HT)* positively influence BI, with HM strongest, while PI shows no effect.

Both studies, consistent with subsection V-A, confirm a strong positive effect of *Behavioral Intention (BI)* on UB. In [6], PT is the only additional predictor of UB, with no significant effect.

## VI. DISCUSSION

This section points out and elaborates on the results of our SLR in relation to the research questions (see section II), highlighting implications and inconsistencies.

The results for students' and lecturers' acceptance factors of GenAI tools are discussed jointly to answer **RQ1**.

Performance Expectancy (PE) consistently shows a strong positive effect on BI for both groups, making it one of the most important predictors. Adoption of LLM-based chatbots depends largely on their ability to solve problems and provide accurate information [52], [6], [7], [36], [37], [28], [3], [8], [40], [33], [4].

The effect of Effort Expectancy (EE) on BI is inconsistent and mainly influenced by prior experience and cultural background. Experienced users perceive chatbots as intuitive, reducing the relevance of EE [49], [7], [43], [44], [53]. In contrast, less experienced users value ease of use more, leading to a significant positive effect [6], [36], [50], [45], [51], [29], [47], [41].

Similarly, the influence of SI varies by culture. It is stronger in collectivist contexts (e.g., Saudi-Arabia, China, Indonesia) [2], [44], [46], [47], [4] and weaker in individualistic regions such as Europe or the USA [2].

The effects of Facilitating Conditions (FC) on BI and UB depend largely on infrastructure availability. In less developed regions, access to devices and internet is critical [35], [42], [44], [50], [28], [53], [39], [29], whereas in developed regions its influence is reduced [49], [7], [37], [38], [45], [51], [8], [40], [33]. Training plays a minor role and mainly affects inexperienced users [49], [7], [37], [38], [45], [51], [8], [40], [33].

Hedonic Motivation (HM) generally has a positive effect on BI, as users enjoy human-like interaction [7], [43], [44], [50], [40], [33]. Lecturers also use it to increase engagement [6]. However, this effect may depend on culture [46] and diminish over time as functional benefits become more important [49], [37], [38].

Habit (HT) is among the strongest predictors of both BI and UB [6], [7], [38], [44], [51], [40], [33]. Familiarity from everyday use encourages integration into academic workflows [6], [7], [37], [51], [40], [33].

PV is rarely studied and likely insignificant, as most chatbots offer sufficient free usage plans [49], [40]. Users often do not perceive differences between paid and free versions [37], [44], [33].

Personal Innovativeness (PI) is a weak predictor of BI [43], [38], [8], [40], [33]. Although curiosity exists, it does not strongly drive usage, though early adopters may influence others depending on SI [7], [43], [46], [8].

The effects of Anxiety (ANX) range from negative to insignificant [36], [28], [3], [39], [4]. This variation can be explained by institutional policies reducing uncertainty [36], different adoption stages [39], and the dominance of positive predictors like PE and SI [3], [36].

Perceived Trust (PT) shows a consistently strong positive influence [42], [7], [43], [46], [53]. It includes reliability and data privacy [46], which are particularly important for lecturers evaluating information quality [6], [29]. Trust is therefore a key prerequisite for adoption [42], [6], [43], [29].

Finally, BI strongly predicts UB, and ATU positively influences BI, with BI as the central driver of usage [34], [2], [52], [7], [36], [37], [44], [50], [53], [3], [39], [47], [40], [33], [54], [41], [55]. Enhancing BI is therefore crucial, primarily through PE, HT, PT, HM, and PI, while reducing ANX. Depending on context, SI and FC may also be important.

In summary, the identified factors and predictors for acceptance of GenAI in higher education answer RQ1.

Although we conducted a comprehensive search, our SLR was not able to answer **RQ2** that aimed at specific factors related to acceptance of GenAI in SE education. This is due to the reason that previous studies did not address this specific field, and even the studies with computer science participants only covered general acceptance factors. Together, this encourages future context and task specific acceptance studies in this field.

#### A. Implications

The included studies build on established acceptance models such as TAM, UTAUT, UTAUT2, and TPB, extending them with additional predictors [35], [42], [7], [36], [37], [43], [44], [38], [45], [28], [46], [51], [3], [39], [8], [47], [40], [33], [41], [4]. Integrating multiple models enables linking technological and psychological factors [50], [53], [3]. A key development is combining UTAUT with PMT to explain tensions between benefits and risks [28], [39].

Our SLR confirms that core predictors from UTAUT (PE, EE, SI, FC) and TAM (PU, PEOU) dominate existing research [35], [42], [7], [36], [37], [43], [44], [38], [45], [28], [46], [51], [3], [39], [8], [47], [40], [33], [41], [4]. However, they must be extended with additional variables to fully explain adoption, including PT, ANX, and education-specific factors such as Learning Value (LV), Task Efficiency (TE), Empowerment in Learning (EIL), task characteristics, and further psychological constructs [35], [42], [7], [36], [48], [43], [51], [53], [3].

Future models should incorporate contextual and individual factors such as institutional guidelines, perceived ethics, and digital competence in GenAI use [5], [56], [57], [58], [59].

To better capture contextual differences, multi-group analyses and moderators (e.g., gender, study level, experience) should be applied [35], [60], [37], [38], [33], [54], [41], [4].

Finally, transferability is limited, as effects of EE, SI, FC, and HM vary by context. Less-studied predictors such as ANX and PT require further investigation across diverse settings to improve generalizability [35], [49], [6], [7], [36], [37], [43], [50], [38], [45], [46], [51], [47], [40], [33], [41].

#### B. Limitations of the included studies

In the context of lecturers, most studies are not exclusively focused on this group, limiting generalizability [28], [51], [39], [29]. Additionally, small lecturer sample sizes further restrict broader conclusions [51], [29].

Cultural and sample-specific constraints reduce generalizability across contexts [34], [35], [7], [36], [37], [28], [38], [3], [39], [8], [40], [33], [41], [4]. Limited sample sizes may not fully represent the wider student population, potentially introducing bias [7], [36], [37], [8], [40], [33].

All studies use cross-sectional designs, capturing data at a single point in time [61], [7], [36], [37], [38], [28], [39], [40], [33], [41], [4]. This limits causal interpretation [61], [28], [39], [41] and prevents analysis of long-term behavioral changes during adaptation [7], [36], [4]. Timing is also critical, as rapid developments in LLM-based chatbots like ChatGPT may quickly affect perceptions and usage intentions [7], [37], [38], [40], [33].

Reliance on self-reported survey data introduces risks such as social desirability bias and recall errors [52], [28], [46], [3], [39]. Moreover, the focus on quantitative methods may overlook important qualitative insights into user experiences with generative AI tools [7], [36], [50], [46], [3], [29], [4].

### VII. THREATS TO VALIDITY

Our SLR follows the methodology in section IV to ensure accuracy, objectivity, transparency, and reproducibility. Nevertheless, potential limitations remain relevant for assessing credibility and generalizability [62], [63].

Five validity types are considered: *construct validity*, *reliability*, *internal validity*, *external validity*, and *conclusion validity* [62], [63].

*a) Construct validity:* Construct validity concerns generalizability to the underlying concept [62]. A key risk is omitting relevant studies due to search terms or database selection [63]. To mitigate this, IEEE, ACM, and SCIENCE were selected for broad, peer-reviewed coverage [64], complemented by a structured search string, keyword selection, and explicit inclusion/exclusion criteria.

*b) Reliability:* Reliability refers to repeatability and consistency [63]. This SLR follows KITCHENHAM [27], and all methodological steps and results are documented to ensure reproducibility.

*c) Internal Validity:* Internal validity risks stem from potential bias in search design and study selection [63]. Mitigation includes predefined inclusion, exclusion, and quality criteria and a structured methodology. Keyword selection was

conducted jointly, while screening decisions by one author were guided by objective criteria to reduce subjectivity.

d) *External Validity*: External validity concerns generalizability [62], [63]. The review covers studies from 2020–2025, ensuring timeliness but limiting scope. Results focus on LLM-based chatbots, restricting transferability to broader AI tools, and the exclusion of gray literature may affect representativeness.

e) *Conclusion validity*: Conclusion validity addresses the risk of missing relevant studies. Due to the large number of results, only the first 100 entries from SCIENCE and ACM were considered, which may omit studies. This risk was mitigated through backward snowballing.

## VIII. CONCLUSION AND FUTURE WORK

This SLR identifies key predictors of GenAI acceptance in higher education to answer RQ1. Across studies, factors such as PE, HT, PT and ATU consistently show strong positive effects on BI, while BI is the strongest predictor of UB. In contrast, EE, SI, HM and FC show context-dependent effects influenced by cultural, economic, and experiential factors. These findings highlight that both functional benefits and contextual conditions shape the adoption of LLM-based tools.

As LLM-based chatbots evolve, improvements in quality and risks may change acceptance patterns. Future SLRs should incorporate new studies to maintain up-to-date insights.

Since existing studies are mainly cross-sectional and survey-based, future research should include longitudinal designs and qualitative or mixed methods to capture long-term behavior and deeper user experiences. Additionally, integrating psychological and education-specific factors, e.g., Empowerment in Learning and Learning Value, could extend current models.

Despite a comprehensive literature search was conducted, the limited findings regarding RQ2 show that acceptance of GenAI in SE education has largely not been studied. This underlines the need to conduct acceptance studies with this specific focus among students and lecturers.

A major research gap also exists regarding lecturers, as only five of 31 studies of this SLR deal with this group. In addition, practical implications to increase responsible use of GenAI in higher education are missing. Future work should address these gaps, include diverse cultural contexts, and examine discipline-specific acceptance, to better understand GenAI use in different academic domains.

## GENERATIVE AI DISCLOSURE STATEMENT

During the preparation of this work the authors used ChatGPT in order to improve the language and clarity of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## ACKNOWLEDGEMENTS

This research was supported by the European research project TITAN: Trusted eNvironments for confidenTiAl computiNg and secure data sharing within European Open Science

Cloud initiative of the European Commission, funded by the European Union (Grant Agreement No. 10112982).

## REFERENCES

- [1] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: perceptions, benefits, and challenges in higher education," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 43, Jul. 2023. doi: 10.1186/s41239-023-00411-8. [Online]. Available: <https://doi.org/10.1186/s41239-023-00411-8>
- [2] B. Agyare, J. Asare, A. Kraishan, I. Nkrumah, and D. K. Adjekum, "A cross-national assessment of artificial intelligence (AI) Chatbot user perceptions in collegiate physics education," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100365, 2025. doi: 10.1016/j.caeai.2025.100365. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X25000050>
- [3] M. Nurtanto, S. Nawanksari, V. L. P. Sutrisno, H. Syahrudin, N. Kholifah, D. Rohmanto, I. S. Utami, F. Mutohari, and M. A. Hamid, "Determinants of behavioral intentions and their impact on student performance in the use of AI technology in higher education in Indonesia: A SEM-PLS analysis based on TPB, UTAUT, and TAM frameworks," *Social Sciences & Humanities Open*, vol. 11, p. 101638, 2025. doi: 10.1016/j.ssaho.2025.101638. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2590291125003663>
- [4] Q. Wu, S. Li, S. Xin, Q. Hou, and P. Li, "A study on students' behavioural intention and use behaviour of artificial intelligence-generated content in physical education: Employing an extended the unified theory of acceptance and use of technology model," *Journal of Hospitality, Leisure, Sport & Tourism Education*, vol. 36, p. 100547, 2025. doi: 10.1016/j.jhlste.2025.100547. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1473837625000139>
- [5] B. G. Acosta-Enriquez, O. Huamaní Jordan, J. E. Morales-Angaspilco, G. Campoverde Ventura, J. A. Ruiz Carrillo, L. E. Blanco-García, and L. Huaranga Rivera, "Influence of perceived ethics, prejudice, and teacher concerns on artificial intelligence literacy and implementation: A cross-sectional study using path analysis," *Computers in Human Behavior Reports*, vol. 20, p. 100829, 2025. doi: <https://doi.org/10.1016/j.chbr.2025.100829>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2451958825002441>
- [6] M. A. Bhat, C. K. Tiwari, P. Bhaskar, and S. T. Khan, "Examining ChatGPT adoption among educators in higher educational institutions using extended UTAUT model," *Journal of Information, Communication and Ethics in Society*, vol. 22, no. 3, pp. 331–353, 2024. doi: 10.1108/JICES-03-2024-0033. [Online]. Available: <https://doi.org/10.1108/JICES-03-2024-0033>
- [7] N. Bouebdallah and W. A. Ben Youssef, "Assessing students' intention to adopt generative artificial intelligence," *Journal of Accounting Education*, vol. 72, p. 100984, 2025. doi: 10.1016/j.jaccedu.2025.100984. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0748571525000351>
- [8] H. N. Sabeh, "What Drives IT Students Toward ChatGPT? Analyzing the Factors Influencing Students' Intention to Use ChatGPT for Educational Purposes," in *2024 21st International Multi-Conference on Systems, Signals & Devices (SSD)*, 2024. doi: 10.1109/SSD61670.2024.10548826. ISSN 2474-0446 pp. 533–539, iSSN: 2474-0446. [Online]. Available: <https://ieeexplore.ieee.org/document/10548826>
- [9] A. Saihi, M. Ben-Daya, M. Hariga, and R. As'ad, "A Structural equation modeling analysis of generative AI chatbots adoption among students and educators in higher education," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100274, 2024. doi: 10.1016/j.caeai.2024.100274. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X24000778>
- [10] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, 1991. doi: 10.1016/0749-5978(91)90020-T. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S074959789190020T>
- [11] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003. doi: 10.2307/30036540. [Online]. Available: <https://www.jstor.org/stable/30036540>
- [12] I. Ajzen, "The theory of planned behavior: Frequently asked questions," *Human Behavior and Emerging Technologies*, vol. 2, no. 4, pp. 314–324, 2020. doi: 10.1002/hbe2.195. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.195>

- [13] F. D. Davis, "A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results," Ph.D. dissertation, Massachusetts Institute of Technology, 1986. [Online]. Available: <https://dspace.mit.edu/bitstream/handle/1721.1/15192/14927137-MIT.pdf>
- [14] —, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989. doi: 10.2307/249008. [Online]. Available: <https://www.jstor.org/stable/249008>
- [15] F. Davis, R. Bagozzi, and P. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science*, vol. 35, pp. 982–1003, 1989. doi: 10.1287/mnsc.35.8.982
- [16] Y. Lee, K. A. Kozar, and K. R. T. Larsen, "The Technology Acceptance Model: Past, Present, and Future," *Communications of the Association for Information Systems*, vol. 12, no. 50, 2003. doi: 10.17705/1CAIS.01250. [Online]. Available: <https://aisel.aisnet.org/cais/vol12/iss1/50>
- [17] 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. doi: 10.1287/mnsc.46.2.186.11926. [Online]. Available: <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.46.2.186.11926>
- [18] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions," *Decision Sciences*, vol. 39, no. 2, pp. 273–315, 2008. doi: 10.1111/j.1540-5915.2008.00192.x. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5915.2008.00192.x>
- [19] 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, pp. 157–178, 2012. doi: 10.2307/41410412. [Online]. Available: <https://www.jstor.org/stable/41410412>
- [20] V. Karnowski and A. S. Kümpe, "Diffusion of Innovations," in *Schlüsselwerke der Medienwirkungsforschung*, M. Potthoff, Ed. Springer Fachmedien, 2016, pp. 97–107. ISBN 9783658099237. [Online]. Available: [https://doi.org/10.1007/978-3-658-09923-7\\_9](https://doi.org/10.1007/978-3-658-09923-7_9)
- [21] G. C. Moore and I. Benbasat, "Integrating Diffusion of Innovations and Theory of Reasoned Action models to predict utilization of information technology by end-users," in *Diffusion and Adoption of Information Technology: Proceedings of the first IFIP WG 8.6 working conference on the diffusion and adoption of information technology, Oslo, Norway, October 1995*, K. Kautz and J. Pries-Heje, Eds. Springer US, 1996, pp. 132–146. ISBN 9780387349824. [Online]. Available: [https://doi.org/10.1007/978-0-387-34982-4\\_10](https://doi.org/10.1007/978-0-387-34982-4_10)
- [22] V. Venkatesh, "Adoption and use of AI tools: a research agenda grounded in UTAUT," *Annals of Operations Research*, vol. 308, no. 1, pp. 641–652, 2022. doi: 10.1007/s10479-020-03918-9. [Online]. Available: <https://doi.org/10.1007/s10479-020-03918-9>
- [23] M. S. Farooq, M. Salam, N. Jaafar, A. Fayolle, K. Ayupp, M. Radovic-Markovic, and A. Sajid, "Acceptance and use of lecture capture system (LCS) in executive business studies: Extending UTAUT2," *Interactive Technology and Smart Education*, vol. 14, no. 4, pp. 329–348, 2017. doi: 10.1108/ITSE-06-2016-0015. [Online]. Available: <https://doi.org/10.1108/ITSE-06-2016-0015>
- [24] R. W. Rogers, "A Protection Motivation Theory of Fear Appeals and Attitude Change," *The Journal of Psychology*, vol. 91, no. 1, pp. 93–114, 1975. doi: 10.1080/00223980.1975.9915803. [Online]. Available: <https://doi.org/10.1080/00223980.1975.9915803>
- [25] —, "Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation," *Social psychology: A source book*, pp. 153–176, 1983.
- [26] B. G. Acosta-Enriquez, E. V. Ramos Farroñan, L. I. Villena Zapata, F. S. Mogollon García, H. C. Rabanal-León, J. E. M. Angaspilco, and J. C. S. Bocanegra, "Acceptance of artificial intelligence in university contexts: A conceptual analysis based on utaut2 theory," *Heliyon*, vol. 10, no. 19, p. e38315, 2024. doi: <https://doi.org/10.1016/j.heliyon.2024.e38315>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844024143460>
- [27] B. Kitchenham, S. Charters et al., *Guidelines for performing Systematic Literature Reviews in software engineering*. Keele, UK, 2007. [Online]. Available: [https://www.researchgate.net/publication/258968007\\_Kitchenham\\_B\\_Guidelines\\_for\\_performing\\_Systematic\\_Literature\\_Reviews\\_in\\_software\\_engineering\\_EBSE\\_Technical\\_Report\\_EBSE-2007-01](https://www.researchgate.net/publication/258968007_Kitchenham_B_Guidelines_for_performing_Systematic_Literature_Reviews_in_software_engineering_EBSE_Technical_Report_EBSE-2007-01)
- [28] W.-L. Hsu and A. D. K. Silalahi, "Exploring the paradoxical use of chatgpt in education: Analyzing benefits, risks, and coping strategies through integrated utaut and pmt theories using a hybrid approach of sem and fsqca," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100329, 2024. doi: <https://doi.org/10.1016/j.caeai.2024.100329>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X24001322>
- [29] M. M. Rana, M. S. Siddiquee, M. N. Sakib, and M. R. Ahamed, "Assessing AI adoption in developing country academia: A trust and privacy-augmented UTAUT framework," *Heliyon*, vol. 10, no. 18, p. e37569, 2024. doi: 10.1016/j.heliyon.2024.e37569. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844024136006>
- [30] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981. doi: 10.1177/002224378101800104. [Online]. Available: <https://doi.org/10.1177/002224378101800104>
- [31] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015. doi: 10.1007/s11747-014-0403-8. [Online]. Available: <https://doi.org/10.1007/s11747-014-0403-8>
- [32] R. F. Falk and N. B. Miller, *A primer for soft modeling*, ser. A primer for soft modeling. University of Akron Press, 1992. ISBN 9780962262845
- [33] A. Strzelecki, "To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology," *Interactive Learning Environments*, vol. 32, no. 9, pp. 5142–5155, 2024. doi: 10.1080/10494820.2023.2209881. [Online]. Available: <https://doi.org/10.1080/10494820.2023.2209881>
- [34] Z. Abidin, K. Phusavat, R. Arifudin, and E. Sugiharti, "Unveiling Factors Affecting Intentions to Use ChatGPT: Perspectives from Indonesian Higher Education Students," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, 2024. doi: 10.1109/ICIC64337.2024.10956977 pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/10956977>
- [35] M. A. Ayanwale, "Using Diffusion Theory of Innovation to Investigate Perceptions of STEM and Non-STEM Students' Adoption of Chatbot Systems in Higher Education: A Multiple Group Analysis," in *2024 IEEE Global Engineering Education Conference (EDUCON)*, 2024. doi: 10.1109/EDUCON60312.2024.10578835. ISSN 2165-9567 pp. 1–8, iSSN: 2165-9567. [Online]. Available: <https://ieeexplore.ieee.org/document/10578835>
- [36] T. Budhathoki, A. Zirar, E. T. Njoya, and A. Timsina, "ChatGPT adoption and anxiety: a cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT)," *Studies in Higher Education*, vol. 49, no. 5, pp. 831–846, 2024. doi: 10.1080/03075079.2024.2333937. [Online]. Available: <https://doi.org/10.1080/03075079.2024.2333937>
- [37] S. Grassini, M. L. Aasen, and A. Møgelvang, "Understanding University Students' Acceptance of ChatGPT: Insights from the UTAUT2 Model," *Applied Artificial Intelligence*, vol. 38, no. 1, p. 2371168, 2024. doi: 10.1080/08839514.2024.2371168. [Online]. Available: <https://doi.org/10.1080/08839514.2024.2371168>
- [38] A. A. Hernandez, M. B. Abisado, R. L. Rodriguez, and J. M. R. Imperial, "Predicting the Use Behavior of Higher Education Students on ChatGPT: Evidence from the Philippines," in *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 2023. doi: 10.1109/TALE56641.2023.10398324 pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/document/10398324>
- [39] E. M. Panggabean and A. D. K. Silalahi, "How do chatgpt's benefit-risk-coping paradoxes impact higher education in taiwan and indonesia?" *Computers and Education: Artificial Intelligence*, vol. 8, p. 100412, 2025. doi: <https://doi.org/10.1016/j.caeai.2025.100412>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X25000529>
- [40] A. Strzelecki, "Students' Acceptance of ChatGPT in Higher Education: An Extended Unified Theory of Acceptance and Use of Technology," *Innovative Higher Education*, vol. 49, no. 2, pp. 223–245, 2024. doi: 10.1007/s10755-023-09686-1. [Online]. Available: <https://doi.org/10.1007/s10755-023-09686-1>
- [41] J. Wang, "Students' Continuance Intention to Use Generative AI in Higher Education: Integrating the Technology Acceptance Model and Information Success Model," in *Proceedings of the 2025 3rd International Conference on Educational Knowledge and*

- Informatization*, ser. EKI '25. Association for Computing Machinery, 2025. doi: 10.1145/3765325.3765346. ISBN 9798400715846 pp. 115–124. [Online]. Available: <https://dl.acm.org/doi/10.1145/3765325.3765346>
- [42] M. A. Ayanwale and M. Ndlovu, “Investigating factors of students’ behavioral intentions to adopt chatbot technologies in higher education: Perspective from expanded diffusion theory of innovation,” *Computers in Human Behavior Reports*, vol. 14, p. 100396, 2024. doi: 10.1016/j.chbr.2024.100396. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2451958824000290>
- [43] H. Guo and T. Erdenebold, “Factors influencing intention to adopt an AI chatbot for learning in higher education: An integrated PLS-SEM, IPMA, and ANN approach,” *Computers and Education: Artificial Intelligence*, vol. 9, p. 100477, 2025. doi: 10.1016/j.caeai.2025.100477. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X25001171>
- [44] A. Habibi, M. Muhaimin, B. K. Danibao, Y. G. Wibowo, S. Wahyuni, and A. Octavia, “Chatgpt in higher education learning: Acceptance and use,” *Computers and Education: Artificial Intelligence*, vol. 5, p. 100190, 2023. doi: <https://doi.org/10.1016/j.caeai.2023.100190>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X23000693>
- [45] S. Hoi, Y. Yu, and L. Ye, “How Perceived Organizational Support Influences University Students’ Intention to Use AI Language Models in Course Learning: An Exploratory Study Based on the Technology Acceptance Model,” in *Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence*, ser. ICIEAI '23. Association for Computing Machinery, 2024. doi: 10.1145/3660043.3660182. ISBN 9798400716157 pp. 781–786. [Online]. Available: <https://dl.acm.org/doi/10.1145/3660043.3660182>
- [46] F. Jazim, Y. H. Al-Mamary, and A. A. Abubakar, “Developing an integrated model to explore key factors influencing university students’ behavioral intentions to use ChatGPT in enhancing higher education in the hail region,” *Acta Psychologica*, vol. 259, p. 105445, 2025. doi: 10.1016/j.actpsy.2025.105445. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0001691825007589>
- [47] A. E. E. Sobaih, I. A. Elshaer, and A. M. Hasanein, “Examining Students’ Acceptance and Use of ChatGPT in Saudi Arabian Higher Education,” *European Journal of Investigation in Health, Psychology and Education*, vol. 14, no. 3, pp. 709–721, 2024. doi: 10.3390/ejihpe14030047. [Online]. Available: <https://www.mdpi.com/2254-9625/14/3/47>
- [48] L. I. D. Faruk, R. Rohan, U. Nindrutirukun, and D. Pal, “University Students’ Acceptance and Usage of Generative AI (ChatGPT) from a Psycho-Technical Perspective,” in *Proceedings of the 13th International Conference on Advances in Information Technology*, ser. IAIT '23. Association for Computing Machinery, 2023. doi: 10.1145/3628454.3629552. ISBN 9798400708497 pp. 1–8. [Online]. Available: <https://dl.acm.org/doi/10.1145/3628454.3629552>
- [49] I. Arin, A. Farkhan, F. Muhamad, and F. T. Anjani, “Analysis of the Impact of ChatGPT Utilization on the Levels of Laziness and Productivity,” in *2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA)*, 2024. doi: 10.1109/ICTIIA61827.2024.10761147 pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/document/10761147>
- [50] A. Habibi, A. Mukminin, A. Octavia, S. Wahyuni, B. K. Danibao, and Y. G. Wibowo, “ChatGPT acceptance and use through UTAUT and TPB: A big survey in five Indonesian Universities,” *Social Sciences & Humanities Open*, vol. 10, p. 101136, 2024. doi: 10.1016/j.ssaho.2024.101136. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2590291124003334>
- [51] F.-C. Li, D. T. Thanh Phuong, T.-Y. Chou, A. D. K. Silalahi, and L. T. Hue, “Exploring the use of GPT Chatbots, in Education: A Comparative Study on User Experiences, in Taiwan and Vietnam,” in *2024 International Conference on Information Management and Technology (ICIMTech)*, 2024. doi: 10.1109/ICIMTech63123.2024.10780856. ISSN 2837-2778 pp. 1–6, iISSN: 2837-2778. [Online]. Available: <https://ieeexplore.ieee.org/document/10780856>
- [52] N. A. S. Asiri and A. A. Alharbi, “Measuring Graduate Students’ Use of ChatGPT in Academic Writing Based on the Technology Acceptance Model,” *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, vol. 20, no. 1, pp. 1–19, 2025. doi: 10.4018/IJWLTT.392931. [Online]. Available: <https://www.igi-global.com/gateway/article/www.igi-global.com/gateway/article/392931>
- [53] T. T. A. Ngo, T. T. A. Vo, and M. T. Phan, “The psychology of AI adoption in education: University students’ intentions to use Large Language Models for learning from a TAM and TPB perspective,” *Acta Psychologica*, vol. 261, p. 105789, 2025. doi: 10.1016/j.actpsy.2025.105789. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0001691825011023>
- [54] W. Vidarshika, N. Dayapathirana, and A. Ranasinghe, “Understanding AI Chatbot Adoption in Education: The Role of Perceived Usefulness, Ease of Use, and Anthropomorphic Tendencies,” in *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 2025. doi: 10.1109/SCSE65633.2025.11031020. ISSN 2997-7363 pp. 1–6, iISSN: 2997-7363. [Online]. Available: <https://ieeexplore.ieee.org/document/11031020>
- [55] A. Wicaksono, F. I. Barunastra, M. C. J. G. Suhari, and F. Jingga, “Students Level of Trust in the Use of AI Chatbots in Higher Education: A Quantitative Analysis Adopting Technology Acceptance Model,” in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, vol. 7, 2024. doi: 10.1109/IC3I61595.2024.10829196 pp. 305–310. [Online]. Available: <https://ieeexplore.ieee.org/document/10829196>
- [56] N. Annamalai, B. Bervell, D. O. Mireku, and R. P. K. Andoh, “Artificial intelligence in higher education: Modelling students’ motivation for continuous use of ChatGPT based on a modified self-determination theory,” *Computers and Education: Artificial Intelligence*, vol. 8, p. 100346, 2025. doi: 10.1016/j.caeai.2024.100346. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X24001498>
- [57] Y. Luo and W. Y. Leong, “Exploring the factors influencing of teachers’ acceptance of artificial intelligence in higher education english teaching,” in *2024 International Conference on Intelligent Education and Intelligent Research (IEIR)*, 2024. doi: 10.1109/IEIR62538.2024.10960147 pp. 1–5. [Online]. Available: <https://ieeexplore.ieee.org/document/10960147>
- [58] H. Yang, N. Xu, X. Lin, and W. Zhang, “Integrating AI literacy into the TAM-TPB model to explain students’ intention to use educational AI through MASEM approach,” *Computers in Human Behavior Reports*, vol. 20, p. 100833, 2025. doi: 10.1016/j.chbr.2025.100833. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2451958825002489>
- [59] Z. You, M. Guo, G. Wainaina, K. Qi, and H. Guo, “AI Literacy and Academic Performance of Chinese University Students: Mediating Role of Learning Attitude,” in *Proceedings of the 2nd International Conference on Educational Knowledge and Informatization*, ser. EKI '24. Association for Computing Machinery, 2024. doi: 10.1145/3691720.3691805. ISBN 9798400710230 pp. 495–503. [Online]. Available: <https://dl.acm.org/doi/10.1145/3691720.3691805>
- [60] F. M. C. Remegio and R. Asahid-Cheng, “Decoding Acceptance through Technology Acceptance Model: A Descriptive Study of ChatGPT Usage Across Academic Disciplines,” in *2024 4th International Conference on Educational Technology (ICET)*, 2024. doi: 10.1109/ICET62460.2024.10868886 pp. 398–402. [Online]. Available: <https://ieeexplore.ieee.org/document/10868886>
- [61] L. Alenazi and E. Alhalal, “Factors affecting Artificial Intelligence usage intention among nursing students: Unified theory of acceptance and use of technology,” *Nurse Education Today*, vol. 152, p. 106780, 2025. doi: 10.1016/j.nedt.2025.106780. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0260691725002163>
- [62] F. Casillo, V. Deufemia, and C. Gravino, “Automatic identification of privacy and security requirements: a systematic literature review,” *Requirements Engineering*, vol. 31, no. 1, p. 1, 2025. doi: 10.1007/s00766-025-00455-4. [Online]. Available: <https://doi.org/10.1007/s00766-025-00455-4>
- [63] G. P. Tomaselli, N. S. Pinto, and C. Acuña, “Requirements management methods and practices for improving the quality of agile software development processes: a review of the literature,” *Requirements Engineering*, vol. 30, no. 4, pp. 371–398, 2025. doi: 10.1007/s00766-025-00448-3. [Online]. Available: <https://doi.org/10.1007/s00766-025-00448-3>
- [64] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa, “Automatic assessment of text-based responses in post-secondary education: A systematic review,” *Computers and Education: Artificial Intelligence*, vol. 6, p. 100206, 2024. doi: 10.1016/j.caeai.2024.100206. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X24000079>

# Author Index

|                                          |         |                                       |         |
|------------------------------------------|---------|---------------------------------------|---------|
| <b>A</b> bayomi-Alli, Olusola . . . . .  | 81      | <b>M</b> askeliūnas, Rytis . . . . .  | 81      |
| Alshehri, Saleh Z . . . . .              | 1       | Mikhov, Rossen . . . . .              | 67      |
| <b>B</b> eloff, Natalia . . . . .        | 1       | Myasnichenko, Vladimir . . . . .      | 67      |
| Bolanowski, Marek . . . . .              | 89      | <b>N</b> kinyili, Tiberius . . . . .  | 73      |
| Borowiecki, Paweł . . . . .              | 11      | <b>O</b> dusami, Modupe . . . . .     | 81      |
| Bylina, Beata . . . . .                  | 11      | Ogara, Solomon . . . . .              | 73      |
| <b>C</b> iaschi, Matteo . . . . .        | 17      | Omwenga, Vincent . . . . .            | 73      |
| <b>D</b> amaševičius, Robertas . . . . . | 81      | <b>P</b> ap, Zsigmond . . . . .       | 111     |
| Dan, Daniel . . . . .                    | 17      | Paszkiewicz, Andrzej . . . . .        | 89      |
| Dörpinghaus, Jens . . . . .              | 25      | Pataricza, András . . . . .           | 111     |
| <b>E</b> sser, Alexander . . . . .       | 93      | Polychronopoulos, Nikos . . . . .     | 59      |
| <b>F</b> idanova, Stefka . . . . .       | 67      | <b>R</b> eiser, Thomas . . . . .      | 93, 119 |
| Földvári, András . . . . .               | 111     | Riediger, Volker . . . . .            | 93      |
| <b>G</b> avranis, Andreas . . . . .      | 59      | Rockenfeller, Robert . . . . .        | 25      |
| Gera, Anchal . . . . .                   | 25      | <b>S</b> dobnyakov, Nikolay . . . . . | 67      |
| Gerhards, Christian . . . . .            | 37      | Sowa, Piotr . . . . .                 | 89      |
| Gołąbek, Andrzej . . . . .               | 89      | Speckesser, Stefan . . . . .          | 101     |
| <b>H</b> ajri-Gabouj, Sonia . . . . .    | 45      | Steiner, Petra . . . . .              | 93      |
| Harbaoui, Imen . . . . .                 | 45      | <b>T</b> oldi, Balázs Ádám . . . . .  | 111     |
| Hein, Kristine . . . . .                 | 53      | <b>V</b> asileva, Veronika . . . . .  | 93, 119 |
| <b>J</b> ürjens, Jan . . . . .           | 93, 119 | <b>W</b> hite, Martin . . . . .       | 1       |
| <b>K</b> han, Imran . . . . .            | 1       | <b>X</b> u, Lei . . . . .             | 101     |
| Kirilov, Leoneed . . . . .               | 67      | <b>Y</b> oussef, Rabaa . . . . .      | 45      |
| Kostadinovska, Katerina . . . . .        | 53      |                                       |         |
| Kozanidis, George . . . . .              | 59      |                                       |         |
| Kozik, Mateusz . . . . .                 | 89      |                                       |         |