Annals of Computer Science and Information Systems
Volume 41

Preprints of Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 8–11, 2024. Belgrade, Serbia

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

Annals of Computer Science and Information Systems, Volume 41

Series editors:
Maria Ganzha (Editor-in-Chief),
Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland
Leszek Maciaszek,
Macquarie University, Australia and Wrocław Universty of Economy, Poland
Marcin Paprzycki,
Systems Research Institute Polish Academy of Sciences and Management Academy, Poland
Dominik Ślęzak,
University of Warsaw, Poland and QED Software, Poland, and DeepSeas, USA

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
Hristo Djidjev,
Los Alamos National Laboratory, USA and Bulgarian Academy of Sciences, Bulgaria
Włodzisław Duch,
Nicolaus Copernicus University, Toruń, Poland
Hans-George Fill,
University of Fribourg, Switzerland
Ana Fred,
University of 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, China
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 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

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

POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

2024, Warszawa, Polskie
Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 41

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

USB: ISBN 978-83-973291-1-9
WEB: ISBN 978-83-973291-0-2

ISSN 2300-5963
DOI 10.15439/978-83-973291-0-2

© 2024, Polskie Towarzystwo Informatyczne
Ul. Solec 38/103
00-394 Warsaw
Poland

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

Cover art:
Paulina Wiewióra,
Elbląg, Poland

Also in this series:

Volume 40: Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), ISBN WEB: 978-83-969601-9-1, ISBN USB: 978-83-969601-0-8
Volume 39: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), ISBN WEB: 978-83-969601-6-0, ISBN USB: 978-83-969601-7-7, ISBN ART 978-83-969601-8-4
Volume 38: Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering, ISBN WEB: 978-83-969601-5-3
Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6
Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2
Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8
Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation ISBN 978-83-965897-8-1
Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4
Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0
Volume 31: Position Papers of the 17th Conference on Computer Science and Intelligence Systems, ISBN WEB: 978-83-965897-2-9, ISBN USB: 978-83-965897-3-6
Volume 30: Proceedings of the 17th Conference on Computer Science and Intelligence Systems, ISBN WEB: 978-83-962423-9-6, ISBN USB: 978-83-965897-0-5
Volume 29: Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business AnalyticsISBN WEB: 978-83-962423-7-2, ISBN USB: 978-83-962423-6-5

DEAR Reader, it is our pleasure to present to you Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS 2024), which took place on September 8-11, 2024, in Belgrade, Serbia.

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 2024 was chaired by Ivan Lukovic, while Dragana Makajić-Nikolić was the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute of Polish Academy of Sciences, The Faculty of Mathematics and Information Science Warsaw University of Technology, The Faculty of Electrical and Computer Engineering of the Rzeszów University of Technology, and The Faculty of Organizational Science of the University of Belgrade.

FedCSIS 2024 was technically co-sponsored by IEEE Poland Section, IEEE Serbia and Montenegro Section, Poland Section of IEEE Computer Society Chapter, Czechoslovakia Section of IEEE Computer Society Chapter, Serbia and Montenegro Section of IEEE Computer Society Chapter, Poland Section of IEEE Systems, Man, and Cybernetics Society Chapter, Poland Section of IEEE Computational Intelligence Society Chapter, Serbia and Montenegro Section of IEEE Computational Intelligence Society Chapter, Serbia and Montenegro Section of IEEE Education Society Chapter, Serbia and Montenegro Section of IEEE Young Professionals Affinity Group, Committee of Computer Science of Polish Academy of Sciences, Informatics Association of Serbia, and Mazovia Cluster ICT.

FedCSIS 2024 was organized in collaboration with the Strategic Partner: QED Software, and sponsored by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, Banca Intesa, Nelt Group, Netconomy, Elsevier, Journal of Computer Languages, OnlyOffice, Ascensio Systems d.o.o., Beograd, MDPI and Yettel Bank.

During FedCSIS 2024, the following Keynote and Invited lectures were delivered:

- Frank, Ulrich, University of Duisburg-Essen, Germany, *Multi-Level Language Architectures: Fostering Reuse, Integration and User Empowerment by Allowing for Additional Abstraction*
- Jovanović, Jelena, University of Belgrade, Serbia, *Learning analytics: Challenges and opportunities opened by AI*
- Kutyniok, Gitta, Ludwig-Maximilians-Universität München, Germany, *Reliable AI: Successes, Challenges, and Limitations*
- Tolvanen, Juha-Pekka, Metacase, Finland, *Languages for non-developers: what, how, where?*
- Dujmović, Jozo, San Francisco State University, USA, *Graded Logic and Professional Decision Making*

FedCSIS 2024 consisted of Main Track, with five Topical Areas and Thematic Sessions. Some of Thematic Sessions have been associated with the FedCSIS conference series for many years, while some of them are relatively new. The role of the Thematic Sessions is to focus and enrich discussions on selected areas, pertinent to the general scope of the conference, i.e. intelligence systems.

Each contribution, found in this volume, was refereed by at least two referees. They are presented in alphabetic order, according to 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.

Making FedCSIS 2024 happen required a dedicated effort of many people. We would like to express our warmest gratitude to the members of Senior Program Committee, Topical Area Curators, Thematic Session Organizers and to the members of FedCSIS 2024 Program Committee. In particular, we would like to thank those colleagues who have refereed the 184 submissions.

We thank the authors of the papers for their great contributions to the theory and practice of computer science and intelligence systems. We are grateful to the keynote and invited speakers, for sharing their knowledge and wisdom with the participants.

Last, but not least, we thank Ivan Lukovic and Dragana Makajić-Nikolić and the FON Team. We are very grateful for all your efforts!

We hope that you had an inspiring conference. We also hope to meet you again for the 20th Conference on Computer Science and Intelligence Systems (FedCSIS 2025) which will take place in Kraków, Poland, on September 14-17, 2025. We also hope that you will approve the evolution of the FedCSIS Conference concept, in the direction that properly addresses the current needs of research and applications. We want to continue looking at Computer Science from different angles but, at the same time, acknowledging the topic Intelligence Systems as the central point of everything that has to be considered.

Co-Chairs of the FedCSIS Conference Series:

Marek Bolanowski, *Rzeszów University of Technology, Poland*

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

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

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, and Warsaw University of Management, Poland*

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

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 8–11, 2024. Belgrade, Serbia

TABLE OF CONTENTS

COMMUNICATION PAPERS

A VARK learning style-based Recommendation system for Adaptive E-learning **1**
Fares Abomelha, Paul Newbury

Autism Spectrum Disorder Learning Application Framework (ASDLAF) Evaluation: ASD Users' Perspectives **9**
Yahya Almazni, Natalia Beloff, Martin White

DTA-SAU: A Conceptual Framework for Digital Transformation Adoption in Saudi Arabian Universities **19**
Rasha Alsulami, Natalia Beloff, Martin White

Enhancing Text Recognition of Damaged Documents through Synergistic OCR and Large Language Models **29**
Thomas Asselborn, Jens Dörpinghaus, Faraz Kausar, Ralf Möller, Sylvia Melzer

Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation **37**
Carina Geldhauser, Konstantin Malyshev

AI in the workplace: who is using it and why? A look at the driving forces behind artificial intelligence in German companies **45**
Christian Gerhards, Myriam Baum

Gradient Boosting Trees and Large Language Models for Tabular Data Few-Shot Learning **53**
Carlos Huertas

Towards Evolvable APIs through Ontological Analysis **61**
Nikolas Jíša, Robert Pergl

Job Shop Scheduling with Integer Programming, Shifting Bottleneck, and Decision Diagrams: A Computational Study **69**
Brannon King, Robert Hildebrand

Combining Local and Global Weather Data to Improve Forecast Accuracy for Agriculture **77**
František Koutenský, Jiří Pihrt, Miroslav Čepek, Vojtěch Rybář, Petr Šimánek, Michal Kepka, Karel Jedlička, Karel Charvat

Exploring Linguistic and Cultural Differences in Online Job Advertisement Analysis for NLP Applications **83**
Kai Krüger, Lea Grüner

Predictive System of the Water Deficit Analysis for the Black Sea Lowland (an example of the Kherson Region) **95**
Lyudmyla Kuzmych, Yuriy Kyrylov, Mykola Voloshyn

Toxic Molecule Classification Using Graph Neural Networks and Few Shot Learning. **105**
Bhavya Mehta, Kush Kothari, Reshmika Nambiar, Seema Shrawne

Integrating Artificial Intelligence Techniques in Cell Mechanics **111**
Muddasar Naeem, Mario Fiorino, Pia Addabbo, Antonio Coronato

Analyzing the Privacy of a Healthcare RFID Authentication Protocol **117**
Anca-Maria Nica, Ștefana Gheorghiță

Sensitivity Analysis in Air Pollution Modeling Supported by High Performance Supercomputers **125**
Tzvetan Ostromsky, Silvi-Maria Gurova, Meglena Lazarova

Handling Lot Sizing/Job Scheduling Synchronization through Path Search Algorithms **131**
Alain Quilliot, Aurélien Mombelli

Converting German Historical Legal Documents to TEI XML including challenges with Table Extraction **139**
Thomas Reiser, Petra Steiner

Smart Assistants for Enhancing System Security and Resilience **151**
Andrey Sadovykh, Dragos Truscan, Martin Schneider, Wissam Mallouli, Ana Cavalli, Cristina Seceleanu, Tanwir Ahmad

Psychological Needs as Credible Song Signals: Testing Large Language Models to Annotate Lyrics **159**
Eunsun Smith, Yinxuan Wang, Eric Matson

Integrating Artificial Intelligence-based programs into Autism Therapy: Innovations for Personalized Rehabilitation **169**
Fabrizio Stasolla, Enza Curcio, Antonio Zullo, Anna Passaro, Maricarla Di Gioia

Prediction of Temperature and Precipitation Changes for Serbia Using Time Series Models with Machine Learning **177**
Miroslav Tomić, Marija Đukić, Slavica Kordić, Vladimir Dimitrieski, Ivan Luković

FedCSIS 2024 Data Science Challenge: Predicting Stock Trends by a Multi-Dimensional Approach **185**
Quang Hieu Vu, Dymitr Ruta, Ling Cen, Ming Liu

Unused information in token probability distribution of generative LLM: improving LLM reading comprehension through calculation of expected values. **191**
Krystian Zawistowski

Author Index **199**

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 1–8
DOI: 10.15439/2024F5253
ISSN 2300-5963 ACSIS, Vol. 41

A VARK learning style-based Recommendation System for Adaptive E-learning

Fares Abomelha
0009-0000-4342-5859
Department of Computer science
King Khalid University
Department of Informatics
University of Sussex
Brighton, United Kingdom
fa470@sussex.ac.uk

Paul Newbury
0000-0001-6363-0286
Department of Informatics
University of Sussex
Brighton, United Kingdom
p.newbury@sussex.ac.uk

***Abstract*—Adaptive e-learning provides the best recommendations of learning resources according to the needs of the student, including learning style, knowledge level, personality, and the time they can spend on learning materials. Despite technological advancements, current e-learning platforms often fail to consider individual learning styles and knowledge gaps, leading to less effective learning experiences. This research evaluates the effectiveness of creating an adaptive e-learning system that uses the VARK learning model and a recommendation system to identify learning styles and provide personalized learning experiences to students based on their knowledge gap and learning preference in particular topics. The system first administers a VARK e-questionnaire to determine the student's learning style, followed by a pre-test to assess their knowledge level. Based on these assessments, the system assigns a personalized e-learning path aligned with the student's dominant learning style and addresses knowledge gaps in specific topics. The proposed system is expected to enhance learning experiences by providing personalized educational content that aligns with individual learning style and addresses specific knowledge deficiencies. This approach has the potential to substantially enhance educational outcomes and effectiveness of learning by delivering customized educational experiences that cater to the unique requirements of every student.**

***Index Terms*—adaptive e-learning, VARK learning style, recommendation system.**

I. INTRODUCTION

E-LEARNING systems have been gaining a lot of intentions over the past two decades and have experienced a significant increase in use over the last few years due to the pandemic. These mainly web-based systems provide anytime, anywhere, at any pace advantages to learning. This is because e-learning enables students to engage in active learning at any time and from any location, resulting in positive learning outcomes. T. A. el. Galil emphasizes this growth in e-learning stating that: “the e-learning market size has exceeded $315 billion in 2021 and is expected to climb to $400 billion by 2026. In 2021, 27 percent of E.U. citizens aged 16 to 74 years reported participating in online courses and since 2020, 98 percent of universities have moved their classes online.” [1]. These statistics highlight the critical role of e-learning in enhancing contemporary education.

Numerous e-learning platforms, such as Coursera, Udemy, and EdX, offer diverse online learning opportunities [2]. These platforms provide a variety of learning resources, allowing learners to choose the methods that best suit their educational goals. However, it is often mandatory for learners to engage with video content, as most courses require watching videos along with transcripts or additional learning materials [3]. Such platforms assume that all learners are identical in terms of their learning styles, which is not the case for most learners; some might prefer learning through audio or visual aids rather than video.

In the realm of online and electronic education, personalization and adaptive learning are two key features that can improve the educational process. On one hand, personalization has become a critical element in enhancing the learning process. This involves tailoring and customizing the educational content or services to meet the unique needs and learning styles of individuals or groups. This approach supports more effective learning strategies and increases learner engagement. On the other hand, adaptive learning systems learn from each user’s experiences, needs, and preferences. Such adaptive systems are crucial for providing a responsive and customized educational experience that can lead to better learning outcomes.

The increasing recognition of diverse learning preferences among students highlights the need for adaptive e-learning systems customized to individual styles. The VARK model, which categorizes learners into four distant groups as follows: visual, auditory, reading/writing, and kinesthetic modalities. It provides a foundational framework for understanding these preferences. However, existing e-learning platforms often fail to adequately incorporate these distinctions, leading to unsatisfactory learning experiences. By developing a new adaptive e-learning system that leverages the VARK model, we aim to create a more personalized and effective educational environment. This system will dynamically adjust content and instructional methods to match individual learning styles, potentially enhancing learning outcomes and student engagement in diverse educational situations.

The rationale behind adaptive e-learning is supported by educational psychologists who argue and shown evidence that differentiated instruction is a superior option to the traditional “one shoe fits all" method [4], [5]. Differentiated instruction theory states that to teach the learners effectively and efficiently, the tutor must teach and respond according to

©2024, PTI

the differences in the individual, which might be one or all the learner characteristics which include learning styles, cognitive styles, knowledge level, emotional state and other characteristics that effect the process of learning. By embracing these principles, adaptive e-learning systems can provide a more tailored educational experience, accommodating the unique needs of each learner and fostering better educational outcomes.

The proposed system in this study focuses on providing a novel adaptive e-learning system in terms of assisting students by identifying their learning style through VARK assessment and recommending content based on assessments within the system, engaging them in a modified learning pathway. The first step in creating an adaptive environment is often identifying students' learning styles. It has been observed that different students learn to process information using different learning styles. VARK is a learning model proposed by Fleming and Mills that gained popularity in 1992 [6]. This model is used to describe the several ways in which students learn. The VARK learning style model is specifically significant and is simplistic in nature as it focuses only on the content for adaptation.

With the intention of providing a new adaptive e-learning platform to improve the e-learning experience, we have developed a learning system based on the VARK model. This paper is structured as follows: Section II reviews related literature, Section III offers an overview of the system, and Section IV details the design of the system. Section V explains into the technical specifications of our proposed system. The paper ends with a conclusion in Section VI.

II. RELATED WORK AND BACKGROUND

This section reviews papers that investigate adaptive e-learning systems to accommodate various learning styles and recommendation systems, which align with the scope of this study. This is particularly relevant as other learning style models and recommendation systems are often difficult and complex to represent, requiring significant computational resources. To provide a comprehensive overview, Table I has been generated, which presents the common learning styles identified in several studies.

Amaniyan et al. conducted an experiment to examine the impact of personalized learning, tailored to the individual learning styles of nursing students. Approximately 160

TABLE I.

COMMON LEARNING STYLES IN EDUCATION

Learning Style Theory	Main Dimensions	Main Categories	Key Ideas
Dunn and Dunn Model categories [12].	Environmental preferences, Emotional preferences, Sociological preferences, psychological preferences.	– Environmental preferences: Lighting, temperature, noise, seating arrangements, etc. – Emotional preferences: Relaxation, motivation, activity level, etc. – Sociological preferences: Working alone or in groups, teacher, and peer interaction, etc. – Psychological preferences: Visual, auditory, kinesthetic, etc.	Students learn most effectively when their preferred learning styles are accommodated in the learning environment.
Felder-Silverman Model [13], [14].	Sensing/Intuitive, Visual/Verbal, Active/Reflective, Sequential/Global.	– Active/Reflective: Learning by doing and reflecting – Sensing/Intuitive: Learning through observation and understanding patterns – Visual/Verbal: Learning through visual aids or written text. – Sequential/Global: Learning through linear, step-by-step approaches or whole-picture approaches.	Different learners have different strengths and preferences in processing information and interacting with learning materials.
Kolb's Experiential Learning Model [15].	Reflective observation, active experimentation, abstract conceptualization, and concrete experience.	– Assimilators: active experimentation and direct experience. - Divergent: direct experience and critical reflection. - Assimilators: reflective observation and conceptual understanding. - Converges: active experimentation and conceptual understanding.	learning is a process of a four-stage cycle of reflective observation, active experimentation, abstract conceptualization, and concrete experience.
Honey and Mumford's Learning Style Model [16].	Activist, Reflector, Theorist, Pragmatist.	– Activists: Learning by doing and exploring. - Reflectors: Learning by observing and reflecting. – Theorists: Learning by understanding concepts and theories. – Pragmatists: Learning by applying ideas to real-life situations.	Learning styles are influenced by an individual's personality and preferred method of learning.

students had participated in the experiment. It was concluded that learners had a significant improvement when they studied with personalized learning material [7]. Other models are computationally expensive to implement. For instance, the work done by Rasheed and Wahid, used 25 attributes to predict the learning style following the Felder Silverman Learning Style Model (FSLM) of learning styles. Collecting, processing 25 attributes and using a machine learning model, when the number of students increase is computationally expensive [8].

People who are visual learners tend to remember information more effectively, when it is presented to them in the graphical form or pictorial representations. Auditory learners are those who prefer to listen to information that is presented to them verbally. Learners who take notes to learn and perform well on written tests fall into the category of reading and writing. People who are kinesthetic learners tend to learn by doing direct experiments or taking quizzes [9]. Usually, a questionnaire is administered to determine which category of learner the student belongs to. This questionnaire is a learning preference assessment tool that consists of 16 questions with four options to help determine the student's learning style. Zulfiani et al. in their research proved that science students prefer and learn more effectively using the kinesthetic and aural learning styles [10]. Diaz et al. proposed and developed a VARK learning style based adaptive learning system; they assessed on 100 students and concluded that the learning efficiency was improved due to it [11].

Recommendation systems are a key functionality of adaptive e-learning system. There are several approaches to implementing a recommendation system such as fuzzy logic, concept map, hybrid, multi-stage, and rule-based etc. As an example, Segal et al. disuse that content-based filtering is performed based on the similarity of the learning materials to the student's learning preference, topic interest and knowledge in that concept. Also, collaborative filtering uses similarities between users and items, both can be used to recommend learning resources. In e-learning, it can focus on a group of students [18]. Hybrid recommender system use two categories of recommending techniques, by first gathering resources according to users' choices and then filtering using the group preferences [19].

The multi-stage recommendations are a comparatively new concept in recommender systems where the recommendations go through a series of filtering, ranking, and ordering stages [20]. The mentor will keep track of the topic of interest and the materials in a database. Some of the concepts mentioned above are used in the proposed research in Section III to implement an adaptive e-learning environment. Karthika et al. [21] discuss the importance of adaptive e-learning systems that can accommodate the learner's preferences and knowledge levels of learning. Their work presents an intelligent and adaptive e-learning system for a software package that is fuzzy-based using fuzzy concept maps (FCM) and that gives e-learners the relevant domain content based on their knowledge level.

The learner model is used to identify the characteristics of e-learners such as both personal behaviour and the individual's level of knowledge to provide adaptive learning content. These learning resources are maintained in a shared database, and e-learners can retrieve the material they need for learning purposes. Following the learning process, the system provides assessment to assess the e-learner's understanding of the domain concept learned. A fuzzy cognitive map represents the learners' level of knowledge and make proper suggestions for related concepts using fuzzy sets and fuzzy rules. The 26 fuzzy rules help to classify the e-learner's knowledge level accurately.

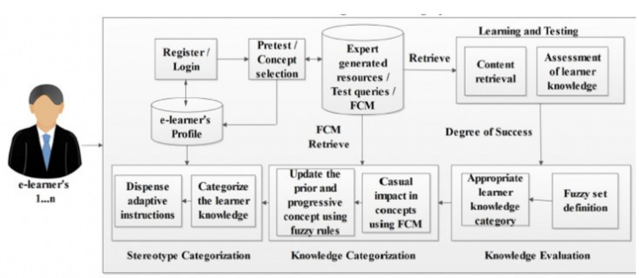

Learning and Testing
Register / Login
Pretest / Concept selection
Expert generated resources / Test queries / FCM
Retrieve
Content retrieval
Assessment of learner knowledge
e-learner's Profile
FCM Retrieve
Degree of Success
e-learner's 1...n
Dispense adaptive instructions
Categorize the learner knowledge
Update the prior and progressive concept using fuzzy rules
Casual impact in concepts using FCM
Appropriate learner knowledge category
Fuzzy set definition
Stereotype Categorization
Knowledge Categorization
Knowledge Evaluation

Fig. 1 Intelligent e-learning system [22]

The system proposed in this work as it appears in figure 1 aims to exceed the constraints that traditional e-learning systems by providing a fuzzy-based intelligent and adaptive e-learning system that can cater to the varying degrees of progress of e-learners. The evaluation of the proposed system shows encouraging outcomes in the accurate categorisation of online learners and their actual knowledge. The proposed system was evaluated only on programming learning material. Nevertheless, they have shown effective results in this domain.

The work of Benhamdi et al. [23] focus on developing a personalized approach to learning by recommending suitable learning materials according to students' prior knowledge, interests, retention abilities and preferences. The approach, known as NPR-EL (New multi-Personalized Recommender for E-learning), mixes content-based filtering with collaborative. The system is incorporated within a learning environment to provide personalised learning resources. The process begins with creating a course by specifying the learning scenario using an xml file called a manifest. The manifest includes elements such as roles, activities, methods, resources, title, and learning objectives. The novel recommender system, NPR-EL, recommends learning materials, which are added to the manifest. These recommended materials, along with the personalized manifest, form a Unit of Learning (UoL).

The approach involves profiling the learners, clustering similar profiles, and predicting ratings for personalized recommendations. A questionnaire is used to gather information on learners' preferences, domain of interest, educational content types, and memory capacity. The questionnaire includes taxonomy and tests of varying difficulty levels. The proposed approach aims to enhance the learning experience by providing personalized recommendations based on individual learners' needs. However, utilizing xml files for the purpose of specifying learning scenarios may have technical limitations and might restrict the adaptability in connecting with other learning management systems.

The work of Esteban et al. [24] offers a hybrid recommendation system (RS) that uses content-based filtering (CBF) with collaborative filtering (CF) to suggest most appropriate courses to students based on student and course details. A Genetic Algorithm (GA) automatically finds the ideal RS configuration that includes the remaining parameters and the primary criterion. Actual data from the University of Cordoba's (Spain) Computer Science Degree, comprising 2500 inputs from 95 students and 63 courses spread over three academic years, was conducted in the experimental study. The findings of the experiment demonstrate a study of the most pertinent course recommendation criteria, the significance of utilizing a hybrid model to boost recommendation reliability by combining student and course data, and superior performance in comparison to earlier models.

Instructors, competencies, knowledge areas, and topics are considered to select the most comparable courses. These metrics may give relevance to each criterion and create a neighbourhood with the most comparable students. The content-based filtering approach calculates a similarity coefficient based on the course content. Additionally, the instructional manuals' contents are indexed to provide another similarity coefficient. The suggested Genetic Algorithm is an adaptation of Eshelman's CHC method, which employs adaptive search techniques to optimize the configuration of the recommendation system, ensuring the best fit for the given data.

The proposed approach is multi-step. Data description and processing come first. Then, the hybrid multi-criteria RS is described. This algorithm provides recommendations for university students regarding their course selection using many student and course factors. Finally, the optimized approach weights each criterion and automatically refines the remaining RS parameters. Using weights, this technique determines criteria significance. Thus, the most crucial factors are weighted higher. The approach also optimizes RS factors like similarity measurements and neighbourhood size. One shortcoming of CF systems is that new students or courses with limited previous data may not receive accurate recommendations, which is a common issue with CF systems.

Various such systems have been developed using machine learning techniques for adaptation in e-learning systems. Many systems use different parameters such as learning style, user preferences, motivation level, and others. The analysis of previous studies has resulted in the development of a proposed system designed to enhance learning outcomes. This system integrates the VARK model with a fuzzy logic and rule-based algorithm to assess and reveal to learners their dominant learning style and areas where they lack knowledge. It aims to deliver a prioritized selection of educational materials in various formats that are tailored to the learners' preferred learning styles. This customization is intended to rectify any misunderstandings in the learners' comprehension and to demonstrate their preferred mode of learning.

III. SYSTEM OVERVIEW

The proposed system focuses on providing an adaptive e-learning environment by understanding the student's learning style and engaging them in a bespoke learning path. This system entitled Flex Learning, is designed to enhance learners' performance, motivation, and engagement by recommending materials aligned with their dominant learning style. The system works as follow; students will be asked to complete a VARK questionnaire shortly after registering with this platform to identify their learning style.

The responses of the students are stored in the database and student profile. Later, the student is asked to check the list of topics they will learn, and a pre-test is performed to determine their knowledge level in those topics. The responses are recorded and classified at this stage too. Based on the previous responses, the recommendation system then ranks the learning content to the students based on their knowledge gap and their preferred learning style. Throughout the learning process, the student's learning is tracked through various tests and level of understanding on the topic.

The architecture of the adaptive e-learning system contains multiple essential components that collaborate effectively to facilitate an individualized educational experience. As demonstrated in figure 2, the system comprises separate components, each specifically designed to effectively engage and provide support to learners. The following sub-sections explore the structure and functionality of these components, showing how they interact in enhancing the learning experience.

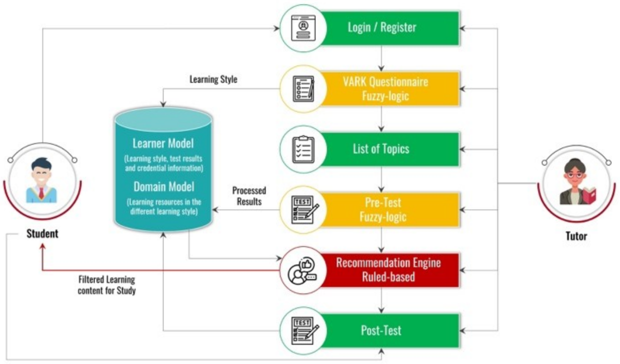


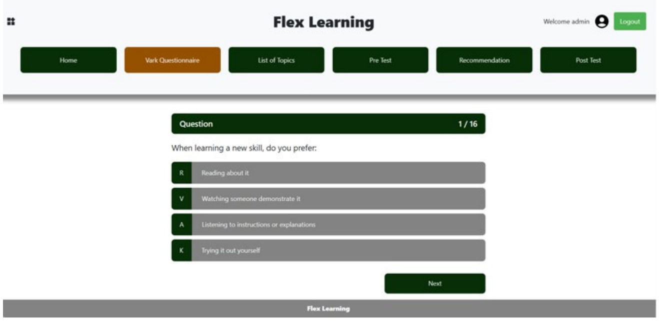
Fig. 2 The architecture of VARK Recommendation System.

A. *Login/Register*

The learners experience begins with the Login/Register module, which acts as the entry point to their individualized learning journey. In this system, learners can access their own profiles after a successful login, which enables customized educational experience that is aligned with their preferences and learning experiences. Also, the tutor component has administrative ability to generate reports and access to manage all system components and users.

B. *VARK Questionnaire*

A key component that makes use of the visual, audio, read/write, and kinaesthetic (VARK) learning style framework is the VARK questionnaire. This module provides learners 16 questions to determine their preferred learning method, revealing whether they learn best by visuals, auditory, reading/writing, or hands-on engagement as shown in figure 3. With the help of this essential data, the material distribution strategy is subsequently tailored according to the learner's preferences. The result of the VARK questionnaire identifies the learner’s dominant learning style.



When learning a new skill, do you prefer:

R Reading about it

V Watching someone demonstrate it

A Listening to instructions or explanations

K Trying it out yourself

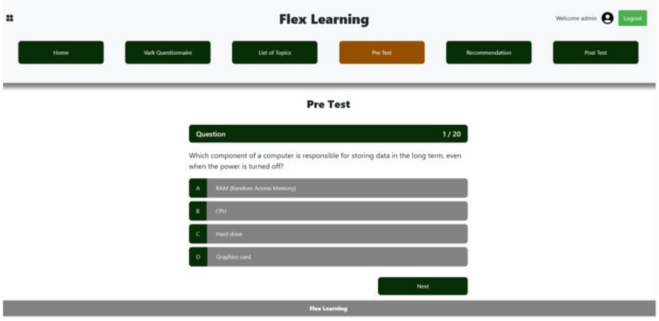
Fig. 3 VARK Assessment Page.

C. List of Topics

The topic list module gives learners the overview of the subjects that are covered on pre-test, and available on the learning system, giving learners responsibility to know what kind of questions they are seen that are covered in the system.

D. Pre-Test Questionnaire

The pre-test evaluates the learner’s current understanding of the topics. Its main objective is to identify knowledge gaps and time spent on each question to help prepare the recommendation system for accurate content recommendations in a ranked way. This process works in combination with the learner’s dominant learning style identified from the VARK questionnaire, but the pre-test itself does not influence the dominant learning style.



Which component of a computer is responsible for storing data in the long term, even when the power is turned off?

A RAM (Random Access Memory)

B CPU

C Hard drive

D Graphics card

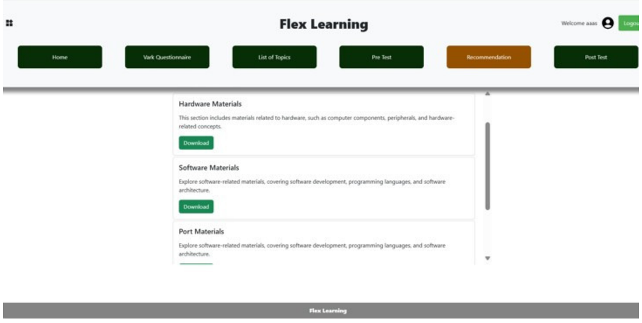
Fig. 4 Pre-Test Page.

E. Recommendation System

The centre of the adaptive e-learning architecture is the recommendation system. It evaluates the learner's VARK dominant learning style, and pre-test performance to rank the materials list for the learner, see figure 5.

F. Content Repository

The content repository contains a collection of educational materials, including texts, videos, chart/diagrams, and quizzes. Moreover, the learner’s profile in the repository contains static data that includes important details such as the username, password, email, major, age, gender, and registration number. Conversely, dynamic data includes the results that are associated with the learner's preferred learning style and learning materials. The relationships shown in figure 6 are indicated by the arrows, which show the direction of the relationship. For example, a user can be associated with one pre-test, one post-test, and one VARK assessment.



Hardware Materials

This section includes materials related to hardware, such as computer components, peripherals, and hardware-related concepts.

Software Materials

Explore software-related materials, covering software development, programming languages, and software architecture.

Port Materials

Explore software-related materials, covering software development, programming languages, and software architecture.

Fig. 5 Recommended Materials Page.

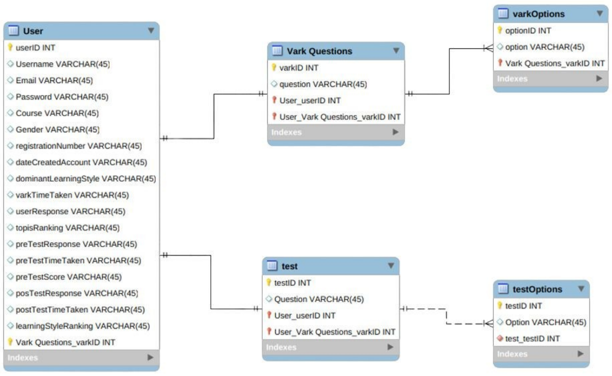


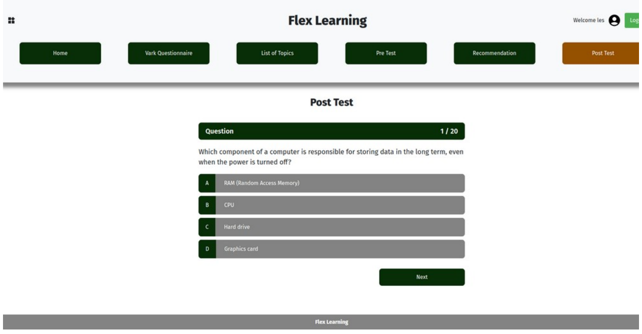
Fig. 6 Flex-Learning Entities Relationship (ER).

G. Post-Test Questionnaire

Upon completing the course materials, the post-test assesses the learner's knowledge retention and comprehension. This assessment, when reviewed with the pre-test results, informs the system's efficacy in content delivery. Learners receive instant results on their performance. See figure 7.

IV. SYSTEM DESIGN

React is an open-source JavaScript library that is open source which is utilised for developing web applications with many user interfaces. It enables programmers to produce UI



Which component of a computer is responsible for storing data in the long term, even when the power is turned off?

A RAM (Random Access Memory)

B CPU

C Hard drive

D Graphics card

Fig. 7 Post-Test Page.

components that are reusable, which facilitates rapid development processes and improves the user experience through efficient web page presentation [25]. React is known for its high performance, as it updates the user interface without needing to reload the entire web page, leading to faster response times and a smoother client-side experience [26].

A. Frontend (React)

A user interface can be designed through the composition of elements known as components [27]. The front end of a website comprises elements that are visible and interactive to the user, including menus. To create such a frontend web interface, specific tools, and technologies, which are typically a browser-controlled arrangement of HTML, CSS, and JavaScript, are required [28]. The frontend interface was built using React which provides a robust and user-friendly experience. The frontend is designed to help users throughout their learning journey, it guides them through the process, starting from filling out the VARK questionnaire to getting access to tailored resources and tests.

Important elements of the front-end architecture consist of:

1. *VARK Questionnaire:* The questionnaire, which is implemented using React components, allow users finding their dominant learning style between visual, audio, read/write, and kinaesthetic learning modes.
2. *Topic Access*: After completing the questionnaire, users can review the list of topics available in the system. This feature helps learners understand the subjects they will encounter in the pre-test. The interface interaction for accessing topics is built using React component, ensuring a smooth and user-friendly experience.
3. *Pre-test*: The pre-test evaluates the learner's current understanding of the topics. It is designed to identify knowledge gaps, which will inform the subsequent recommendation of learning materials. The pre-test is also implemented with React component, providing an interactive and responsive environment for users.
4. *Recommendations:* Resource recommendations are then assigned to the user based on the user's previous dominant learning style and pre-test performance; it is also aligned based on the user's most failed topic in the pre-test as ranked list.

B. Backend (Node.js)

Node.js is a framework that runs JavaScript on the server side and is designed to manage events. Operating as a single-threaded handle that responds to callbacks and never interferes with the primary thread, Node.js is an exceptionally efficient framework for developing web applications [29]. The backend framework, Node.js, manages data analysis, processing, and front-end communication. The backend architecture is made to handle user profiles and evaluations, analyse user responses from the VARK questionnaire effectively, and provide individualized recommendations through the usage of the fuzzy logic and rule-based algorithm.

Important elements of the backend architecture consist of:

1. *Application programming Interface (API) Endpoints:* The communication of requests and data between the frontend and backend is made via Node.js API endpoints. Receiving user responses from the VARK questionnaire, starting the algorithm to generate recommendations, and sending customised recommendations to the frontend are all handled by APIs.
2. *Algorithm Integration:* The algorithm is integrated into the backend to assess user input and produce tailored recommendations for learning materials. The algorithm is efficiently implemented using Node.js, which considers variables such as learning style preferences and based on learner answers on the pre-test outcomes to generate customised recommendations.
3. *User Profile Management:* Node.js manages user profiles, storing information such as learning preferences, assessment results, and resource interactions. User profile enables the application to reveal VARK result, pre-test, post-test and tracking user time spent on their progress. A personalized learning experience, tracking user progress and adapting recommendations over time.

C. Database (Json file)

JSON is a lightweight data structure that is constructed using the data types supported by the JavaScript programming language. Fundamentally, JSON documents are dictionaries composed of key-value pairings, with the value potentially being another JSON document; this configuration enables an unrestricted number of levels of nesting [30]. The project uses a JSON file to store the data. The data are the extracted by the node server and send through the server endpoint to the Frontend by the API.

V. TECHNICAL DETAILS

The technical implementation of the project involves utilising React for the frontend and Node.js for the backend, along with various libraries to enhance functionality and performance.

A. Frontend (React)

The frontend of our system is meticulously crafted using React, a powerful and flexible JavaScript library for building user interfaces. React's component-based architecture allows for reusable and maintainable code, ensuring a seamless and responsive user experience.

- **Component-Based Architecture**: Reacts component-based architecture allows for modular and reusable UI components, facilitating development and maintenance.
- **State Management**: React's state management capabilities, including hooks such as user state and user context, are utilised to manage application state and facilitate dynamic updates.

- **Routing**: Client-side routing using React router allows users navigate between application components and interfaces.
- **API Interaction**: Fetch APIs are utilised to communicate with the backend, sending requests and receiving responses for data retrieval and manipulation.
- **Styling**: CSS pre-processors such as Sass are used for styling UI components, providing flexibility and maintainability in design implementation.

B. Backend (Node.js)

Our backend, built with Node.js, serves as the robust backbone of the system, handling data processing, business logic, and communication with the database.

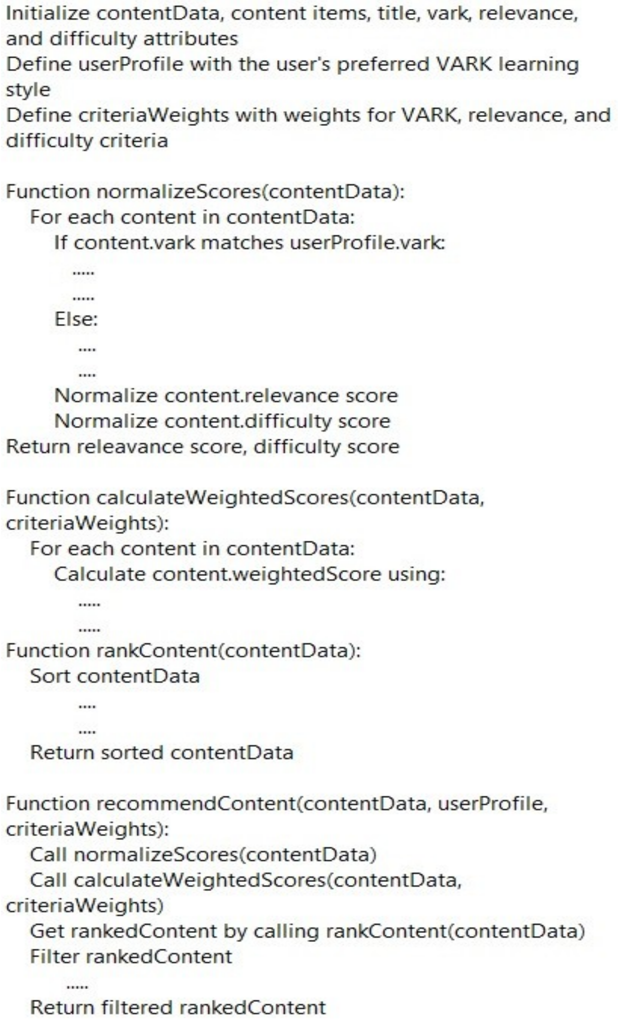
- **Express.js Framework**: Node.js uses Express.js as the web application framework, simplifying the creation of robust APIs and handling of HTTP requests.
- **Database Integration**: Json database solution is integrated with Node.js using libraries such as MongoDB Node.js driver for data storage and retrieval.
- **Authentication and Authorisation**: JSON Web Tokens (JWT) is employed for user authentication and authorisation, ensuring secure access to protected resources.
- **Algorithm Implementation**: The algorithm is implemented in Node.js using custom logic and external libraries, enabling analysis of user responses and generation of personalised recommendations.
- **RESTful API Design:** RESTful API design principles are followed to ensure consistency, scalability, and interoperability in communication between the frontend and backend.

C. Recommendation Algorithm

The heart of our system lies in its advanced recommendation algorithm, designed to deliver personalized content tailored to the user's unique preferences and learning style. In this section we provide a general overview of some of the utilized algorithms.

- **Normalization of Scores**: To ensure a fair comparison across different criteria, the algorithm begins by normalizing the VARK, relevance, and difficulty scores of each content item. This step standardizes the scores, bringing them to a common scale for accurate weighted scoring.
- **Ranking of Content**: With weighted scores from the pre-test in hand, the content items are sorted in descending order to prioritize the most relevant items.
- **Filtering Content**: The final recommendation list is filtered to include only those materials that match the user's preferred VARK learning style, ensuring a highly personalized experience.

The recommendation algorithm also includes the initialization of content and data structures, the use of user profile to implement content-based filtering, and personalized ranking as alternative approaches for users. Figure 9 illustrates a general overview of our utilized algorithm through a snapshot of pseudocode.



```
Initialize contentData, content items, title, vark, relevance,
and difficulty attributes
Define userProfile with the user's preferred VARK learning
style
Define criteriaWeights with weights for VARK, relevance, and
difficulty criteria

Function normalizeScores(contentData):
  For each content in contentData:
    If content.vark matches userProfile.vark:
      .....
      .....
    Else:
      ....
      ....
    Normalize content.relevance score
    Normalize content.difficulty score
Return releavance score, difficulty score

Function calculateWeightedScores(contentData,
criteriaWeights):
  For each content in contentData:
    Calculate content.weightedScore using:
      .....
      .....
Function rankContent(contentData):
  Sort contentData
      ....
      ....
  Return sorted contentData

Function recommendContent(contentData, userProfile,
criteriaWeights):
  Call normalizeScores(contentData)
  Call calculateWeightedScores(contentData,
criteriaWeights)
  Get rankedContent by calling rankContent(contentData)
  Filter rankedContent
    .....
  Return filtered rankedContent
```

Fig. 8 A Snapshot of the Pseudocode Utilized in the VARK learning style-based Recommendation system.

D. Deployment and Hosting

- **Platforms for Deployment:** For hosting and scalability, the application in implemented on platforms called AWS.
- **Monitoring and Logging:** Performance tracking, error tracking, and application metrics logging are done via monitoring tools such as New Relic.

The technical details outlined above provide an overview of the frontend and backend implementation, along with deployment considerations to ensure the reliability and security of the system.

VI. CONCLUSION AND FUTURE WORK

This study introduces A VARK learning style-based recommendation system for e-learning. This system has specifically developed to evaluate learners, indicate their chosen preferred method of learning and level of knowledge, place them on a learning path, and generate a ranked list of materials for learning to help find gaps in the learners' knowledge. A recommendation system has developed within the system to assess the learner's knowledge level on diverse topics and enhance their performance and entire level of learning. This system includes a pre/post-test component. The VARK learning style questionnaire is utilized to identify the learners' preferred learning style. It then suggests appropriate learning materials based on a ranked concept list that aligns with their preferred learning style.

To evaluate the proposed technology, we intend to conduct a pilot study in an online classroom setting for future research. While previous studies have employed various logical approaches and conducted experiments using different programming languages. The system that is proposed comprises multiple novel features and will go through testing within the subjects of a Computer Science area, involving groups of undergraduate students on first, and second year studying in this field. Also, the study will assess findings using a control group that undergoes the same procedures as previously described in the system overview, but without a customized learning pathway. In this group, students will simply be provided with all pertinent learning materials.

REFERENCES

[1] T. A. el. Galil, "E-Learning Statistics 2022: What the Data Show," *Al-Fanar Media*, Oct. 14, 2022. [Online]. Available: https://www.al-fanarmedia.org/2022/10/e-learning-statistics-2022-what-the-data-show/

[2] M. Conache, R. Dima, and A. Mutu, "A Comparative Analysis of MOOC (Massive Open Online Course) Platforms," *Informatica Ecnomica*, vol. 20, no. 2, pp. 4–14, Jun. 2016. [Online]. Available: https://doi.org/10.12948/issn14531305/20.2.2016.01

[3] C. Chrysoulas and M. Fasli, "Building an Adaptive E-Learning System," *Proceedings of the 9th International Conference on Computer Supported Education*, pp. 375–382, 2017. [Online]. Available: https://doi.org/10.5220/0006326103750382

[4] F. Rasheed and Abdul Wahid, "The Theory of Differentiated Instruction and Its Applicability: An E-Learning Perspective," *International Journal of Technical & Non-Technical Research*, vol. 9, no. 4, pp. 193–202, 2018.

[5] A. Miriam and A. Zayas, "Teaching Differentiated Instruction through E-Learning," *Association for the Advancement of Computing in Education (AACE)*, Nov. 4, 2019. [Online]. Available: https://www.learntechlib.org/p/211192/

[6] N. D. Fleming, "Teaching and Learning Styles: VARK Strategies," *IGI Global*, 2001.

[7] S. Amaniyan, Y. Bashiri, and M. Vasismoradi, "Comparison of the Conceptual Map and Traditional Lecture Methods on Students' Learning Based on the VARK Learning Style Model: A Randomized Controlled Trial," *SAGE Open Nursing*, vol. 6, Jan. 2020, Art. no. 2377960820940550. [Online]. Available: https://doi.org/10.1177/2377960820940550

[8] F. Rasheed and Abdul Wahid, "Learning Style Detection in E-Learning Systems Using Machine Learning Techniques," *Expert Systems with Applications*, vol. 174, Jul. 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.114774

[9] A. Ziadat, "The Impact of Using VAKT Strategy on Oral Reading and Reading Comprehension Skills of Elementary Students with Dyslexia," *International Journal of Learning, Teaching and Educational Research* , vol. 20, no. 2, pp. 121–136, Feb. 28, 2021. [Online]. Available: https://doi.org/10.26803/ijlter.20.2.7

[10] Z. Zulfiani, I. Suwarna, and M. Sujiyo, "Science Education Adaptive Learning System as a Computer-Based Science Learning with Learning Style Variations," *Journal of Baltic Science Education*, vol. 17, no. 4, pp. 711–727, Aug. 20, 2018. [Online]. Available: https://doi.org/10.33225/jbse/18.17.711

[11] F. Diaz, T. Rubilar, C. Figueroa, and R. Silva, "An Adaptive E-Learning Platform with VARK Learning Styles to Support the Learning of Object Orientation," *2018 IEEE World Engineering Education Conference (EDUNINE)*, Mar. 2018. [Online]. Available: https://doi.org/10.1109/EDUNINE.2018.8450990

[12] R. Dunn, "Understanding the Dunn and Dunn Learning Styles Model and the Need for Individual Diagnosis and Prescription," *Reading, Writing, and Learning Disabilities*, vol. 6, no. 3, pp. 223-247, 1990.

[13] R. M. Felder and L. K. Silverman, "Learning and Teaching Styles in Engineering Education," *Engineering Education*, vol. 78, no. 7, pp. 674-681, 1988.

[14] S. Graf, Kinshuk, and T. Leo, "In-depth Analysis of the Felder-Silverman Learning Style Dimensions," *Journal of Research on Technology in Education*, vol. 40, no. 1, pp. 79-93, 2007. [Online]. Available: https://doi.org/10.1080/15391523.2007.10782498

[15] S. A. McLeod, "Kolb-Learning Styles and Experiential Learning Cycle," *Simple Psychology*, 2017. [Online]. Available: https://www.simplypsychology.org/learning-kolb.html

[16] P. Honey and A. Mumford, "Learning Styles Questionnaire," *Organization Design and Development, Incorporated*, 1989.

[17] N. D. Fleming, "Teaching and Learning Styles: VARK Strategies," *IGI Global*, 2001.

[18] A. Segal, Z. Katzir, Gal, akov, G. Shani, and B. Shapira, "EduRank: A Collaborative Filtering Approach to Personalization in E-learning," in *Proc. 7th Int. Conf. Educ. Data Min. (EDM 2014)*, 2014, pp. 68–76.

[19] W. Chen, Z. Niu, X. Zhao, and Y. Li, "A Hybrid Recommendation Algorithm Adapted in E-Learning Environments," *World Wide Web*, vol. 17, no. 2, pp. 271–284, 2014, doi: 10.1007/S11280-012-0187-Z.

[20] R. Ak, B. Schifferer, S. Rabhi, and G. De Souza Pereira Moreira, "Training and Deploying Multi-Stage Recommender Systems," in *RecSys 2022 - Proc. 16th ACM Conf. Recomm. Syst.*, 2022, vol. 22, pp. 706–707, doi: 10.1145/3523227.3547372.

[21] R. Karthika, L. Jegatha Deborah, and P. Vijayakumar, "Intelligent E-Learning System Based on Fuzzy Logic," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7661–7670, 2020, doi: 10.1007/S00521-019-04087-Y.

[22] R. Karthika, L. Jegatha Deborah, and P. Vijayakumar, "Intelligent E-Learning System Based on Fuzzy Logic," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7664, 2020, doi: 10.1007/S00521-019-04087-Y.

[23] S. Benhamdi, A. Babouri, and R. Chiky, "Personalized Recommender System for E-Learning Environment," *Education and Information Technologies*, vol. 22, no. 4, pp. 1455–1477, 2017, doi: 10.1007/s10639-016-9504-y.

[24] A. Esteban, A. Zafra, and C. Romero, "Helping University Students to Choose Elective Courses by Using a Hybrid Multi-Criteria Recommendation System with Genetic Optimization," *Knowledge-Based Systems*, vol. 194, 105385, 2020, doi: 10.1016/J.KNOSYS.2019.105385.

[25] P. Rawat and A. N. Mahajan, "ReactJS: A Modern Web Development Framework," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 11, pp. 698-702, 2020.

[26] A. Boduch, R. Derks, and M. Sakhniuk, "React and React Native: Build Cross-Platform JavaScript Applications with Native Power for the Web, Desktop, and Mobile," *Packt Publishing Ltd.*, 2022.

[27] Chiarelli, *Beginning React: Simplify Your Frontend Development Workflow and Enhance the User Experience of Your Applications with React*. Packt Publishing Ltd., 2018.

[28] H. M. Abdullah and A. M. Zeki, "Frontend and Backend Web Technologies in Social Networking Sites: Facebook as an Example," in *2014 3rd International Conference on Advanced Computer Science Applications and Technologies*, Dec. 2014, pp. 85-89.

[29] D. Herron, "Node.js Web Development: Server-Side Web Development Made Easy with Node 14 Using Practical Examples," *Packt Publishing Ltd.*, 2020.

[30] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of JSON Schema," in *Proceedings of the 25th International Conference on World Wide Web*, Apr. 2016, pp. 263-273.

DOI: 10.15439/2024F3189
ISSN 2300-5963 ACSIS, Vol. 41

Autism Spectrum Disorder Learning Application Framework (ASDLAF) Evaluation: ASD Users' Perspectives

Yahya Almazni
dept. of Information Systems
Najran University
dept. of Informatics
University of Sussex
Brighton, United Kingdom
Ya275@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
dept. of Informatics
University of Sussex
Brighton, United Kingdom
N.Beloff@sussex.ac.uk

Martin White
0000-0001-8686-2274
dept. of Informatics
University of Sussex
Brighton, United Kingdom
M.White@sussex.ac.uk

Abstract*—Autism Spectrum Disorder (ASD) is a developmental condition that causes lifelong impairments in social interaction and communication skills. Besides early interventions and ASD services, technology has become essential in assisting ASD individuals in building and improving their communication and social skills. However, most technological innovations have been developed for Western countries. When considering Saudi Arabia as a developing country, people with ASD may have to deal with other difficulties, such as culture, privacy, and societal integration. To address this, we developed and evaluated a framework to investigate the factors that most influence the adoption of ASD Learning Applications (LA). Additionally, 16 hypotheses were defined to test specific relationships between the factors. A quantitative method was utilised to design our experimental study. This paper presents an evaluation of the Autism Spectrum Disorder Learning Application Framework (ASDLAF) from the perspective of individuals with ASD in Saudi Arabia. Utilising a quantitative approach, data was gathered through a survey with a total of 395 participants. Results of hypotheses testing revealed significant support for most factors except social rules (SR), religion (RE), and age (AG). These findings underscore the importance of cultural, technological, and pedagogical considerations in technology adoption for ASD individuals. The results also suggest that while the ASDLAF may be effective in addressing certain factors of ASD learning applications, there are still areas that need to be further developed to meet the specific cultural and religious needs of individuals in Saudi Arabia. Furthermore, it will be essential for developers and stakeholders to consider these factors in order to create more beneficial and effective learning tools for individuals with ASD in this region.

***Index Terms*—Autism Spectrum Disorder (ASD), Learning Applications (LA), Adoption Intention, Technology Acceptance Model, Saudi Arabia.**

I. INTRODUCTION

AUTISM Spectrum Disorder (ASD) is a complex neurodevelopmental disorder that impacts an individual's social, communication, and behavioural skills. Impairment in these skills reduces the ability of ASD individuals to understand social interaction [1], for instance, difficulty recognising facial expressions [2], which impacts their ability to engage in social activities. Globally, ASD prevalence has increased in recent decades. The estimated prevalence of ASD in Saudi Arabia is 42,000 cases [3]. According to Rapp et al. in their study [4], 4,900 adults in Canada have been diagnosed with ASD, while in the US, there are 48,500 American adults with ASD. To put these numbers in perspective, consider that the prevalence of ASD in the US is about 1 in 54 children, according to the Centers for Disease Control and Prevention (CDC), indicating a significant public health concern [5]. Given the prevalence of autism spectrum disorders (ASD) worldwide and their effects on diverse societies, it becomes evident that early interventions are essential in addressing the needs of individuals with ASD and improving their development.

Early interventions play a crucial role in helping individuals with ASD shape their behaviour and improve their social and communication skills, which can have a significant impact on overall development. Early interventions include a combination of therapies such as applied behaviour analysis (ABA), speech therapy, and developmental interventions. Research has emphasised the importance of early interventions in improving long-term outcomes for individuals with ASD, which can lead to significant improvements in communication, social skills, and adaptive behaviour [6]. These early interventions generally involve the family and caregivers in significant ways, empowering them to effectively support their child's needs and search for appropriate services. Many Arab countries experienced a shortage of ASD diagnoses, which led to burdens and complex impairments to ASD individuals and their families [7]. The lack of early interventions and diagnoses in these countries may be due to a lack of ASD awareness among societies. It is also possible that some individuals with autism remain undiagnosed due to the stigma surrounding mental health issues.

Previous studies have shown that technology can be an effective tool to assist individuals with ASD in improving their skills [8], enabling them to work at their own pace and level [9], [10]. The technology aims to mitigate ASD individuals' deficits and provide assistance to overcome difficulties [11]. With Learning Applications (LA), individuals with ASD have access to personalised, engaging, and interactive learning experiences. Mobile LA for individuals with ASD can focus on a variety of areas, including communication, social skills, and

©2024, PTI

Topical area: Information Technology for Business and Society

academic performance. Many applications use Video Self Modelling (VSM) based approaches to improve academic skills [12], [13]. VSM is a therapeutic technique that allows users to complete daily tasks effectively while recording themselves as visual evidence of their capabilities, which can result in increasing the confidence, motivation, and satisfaction of ASD users. The visual supports can be particularly beneficial for individuals with ASD who may struggle with verbal communication and language processing. Additionally, the interactive nature of mobile LA can keep individuals engaged and focused, ultimately leading to greater success in their learning and development. Overall, the use of ASD LAs tailored for individuals with ASD aims to enhance their learning experience and help them progress in various aspects of their lives.

However, most of the innovations regarding ASD have taken root in developed countries [14] or were primarily designed with Western cultural norms and values [15], which may not always be applicable or effective in other parts of the world. People with ASD in non-Western cultures may face difficulties as a result and may not benefit from interventions that are not appropriate for their culture. For example, in many Arab countries, including Saudi Arabia, there is a significant stigma surrounding mental health issues [16] [17], which can discourage families from seeking support for their loved ones with ASD. Additionally, language barriers can also cause significant challenges as the majority of the technological aids are designed with language and cultural considerations that differ significantly from those of native Arabic-speaking users [18], limiting their accessibility for people in need. Moreover, privacy is highly valued in Saudi Arabian culture [19], which can cause challenges for ASD individuals and their families when they seek support and services. This may decrease the possibility of ASD individuals being integrated into mainstream society. Thus, as mentioned earlier, there is a clear need for tailored interventions for individuals with ASD in non-Western cultures like Saudi Arabia, taking into consideration the unique cultural and societal context.

This paper aims to present our quantitative data analysis results of the survey conducted with ASD individuals in Saudi Arabia—the survey aimed at gathering data about ASD individuals' experiences with ASD LA. The analysis results are presented to provide a comprehensive understanding of the participants' views. First, we present a brief summary of our ASDLAF framework to set the context [20]. After that, we delve into our methodology validation process. Subsequently, we present the outcomes obtained from implementing our hypotheses testing, providing insightful analysis and interpretation of the results.

II. THE PROPOSED FRAMEWORK

The proposed framework called "Autism Spectrum Disorder Learning Application Framework for Saudi Adults (ASDLAF)" has been developed by the authors based on a critical evaluation of such models, including the technology acceptance model **(**TAM), and based on the existing literature, which aims to explore and examine the significant factors that assist adults with ASD in Saudi Arabia [20]. As shown in Fig. 1, the framework has a dependent variable that assesses the adoption of ASD LA, intermediate variables that analyse the usability, trust, acceptance, and effectiveness of

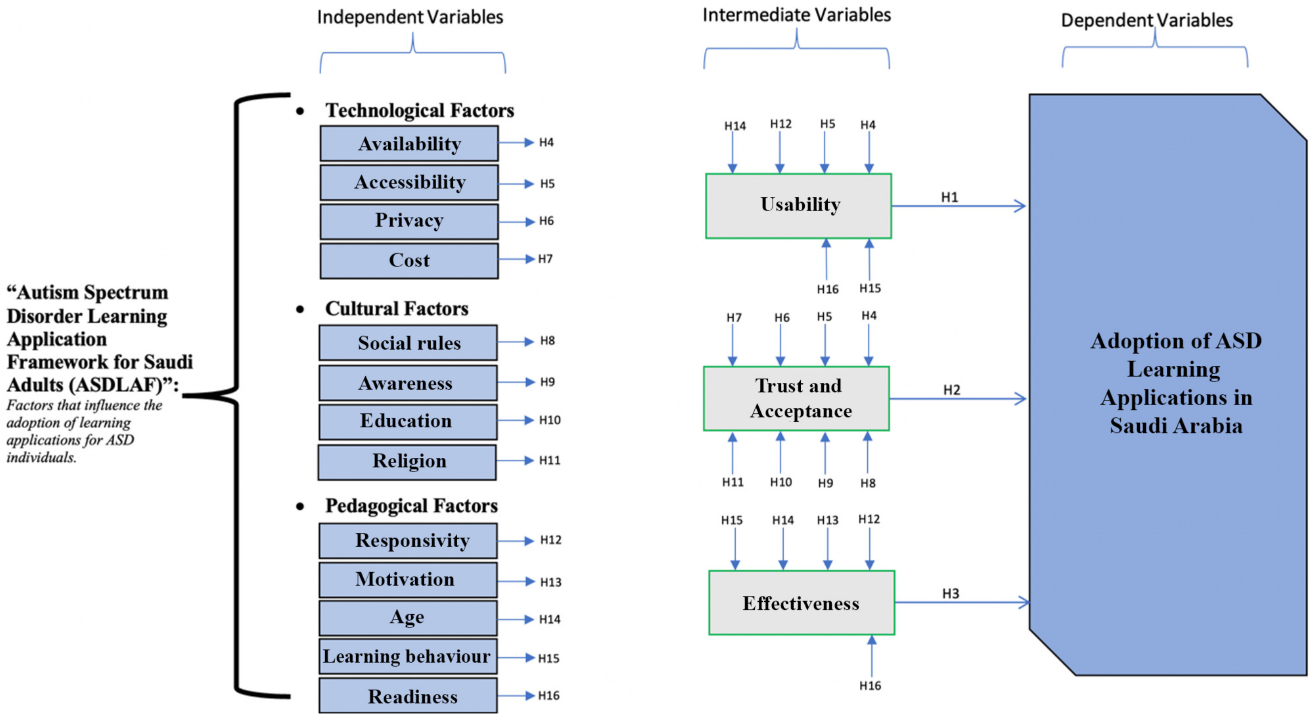


Fig. 1: Autism Spectrum Disorder Learning Application Framework for Saudi adults (ASDLAF) [16]

the technologies, and independent variables focused on technology, culture, and pedagogy that influence the adoption of ASD LA.

The intermediate variables of the proposed framework include Usability, Trust and Acceptance, and Effectiveness, while the technological factors are discussed in terms of how they affect the adoption of new technology. Culture is considered a vital factor that influences new technology adoption, and some cultural factors, such as education, social rules, awareness, and religion, shape the lifestyle of a society in a particular region. Pedagogical factors are also considered, as some individuals with ASD have learning difficulties, which can create significant barriers that influence them to accept and use LA. This study can contribute to a deeper understanding of how defined factors shape ASD individuals' perceptions of the adoption of ASD LAs. It also offers practical insights for policymakers, caregivers, and technology developers, which will help provide better solutions for individuals with ASD in Saudi Arabia.

III. METHODOLOGY VALIDATION

This study aims to investigate the factors that most influence the intention to adopt ASD LAs. The data was collected from a 68-item online survey which was distributed among ASD individuals in autism centres which provide care for people with autism aged above 18 years old. The respondents are informed that they can ask their family or caregivers for help completing the survey if they need it at any time. This occurred in the period between November and December 2023, with a total of 395 responses. The survey included 5-point Likert scale questions, aiming to validate our proposed ASDLAF framework and improve our understanding of the ASD users' experience of accepting ASD LA adoption. SPSS was used to analyse the survey data.

The study sample size was calculated based on the population of ASD in Saudi Arabia, which was estimated by [3]; therefore, the sample size required at least 381 responses to be gathered, with a confidence level of 95% and a confidence interval of 5%. Overall, a total of 412 responses were collected. In order to ensure data quality and validity, incomplete surveys were excluded, resulting in 395 responses for the statistical analysis phase.

A. Reliability and validity of the study

To ensure the survey's reliability, Cronbach’s Alpha α was used to assess the extent to which items in a scale are correlated with each other using the internal consistency method before proceeding to the hypotheses testing phase. Table I shows the Cronbach’s Alpha results and indicates that the reliability of all constructs is above 0.7, which is considered within the acceptable range according to [21], suggesting that α needs to be higher than or equal to 0.7.

TABLE I.
INTERNAL CONSISTENCY OF THE STUDY SURVEY INSTRUMENT

Construct	No of items	Cronbach’s Alpha
Usability (US)	3	0.869
Trust and Acceptance (TA)	3	0.844
Effectiveness (EF)	3	0.831
Availability (AV)	3	0.771
Accessibility (AC)	3	0.796
Privacy (PR)	3	0.881
Cost (CO)	3	0.874
Social rules (SR)	3	0.742
Awareness (AW)	3	0.876
Education (ED)	3	0.897
Religion (RE)	3	0.732
Responsivity (RS)	3	0.927
Motivation (MO)	3	0.887
Age (AG)	3	0.768
Learning behaviour (LB)	3	0.774
Readiness (RD)	3	0.794
Adoption of ASD Learning Applications in Saudi Arabia (AL)	3	0.909

This study uses face and content validity to ensure that the items or questions in the survey are relevant and appropriate for measuring the intended construct and, consequentially, enhance the validity of the findings. Validity is not directly calculable but rather inferred from accumulated evidence [22]; thus, the survey was reviewed and tested by 15 academic experts in both M-learning and ASD fields.

IV. RESULTS

This section provides an overview of the study results of the respondents including demographic characteristics and descriptive analysis, summarised in Table II. This section also delves into the framework analysis to explain the correlations between constructs.

A. Respondents’ demographic data

This study reflects the age statistics in Saudi Arabia as, according to the Saudi General Authority for Statistics, 51.1% of the overall population is less than 30 years old, and the median age of the population is 29 years old [23]. In recent decades, Saudi Arabia has experienced prompt demographic changes, with a growing youth population reshaping the social and economic landscape of the country. As more young Saudis rely on smartphones and mobile applications as essential parts of their daily lives [24], stakeholders need to adapt their strategies to target this demographic range effectively, especially for people needing healthcare services. Moreover, individuals often rely heavily on digital technologies during the adulthood transition period. Interestingly, the study also revealed that 29% of the survey respondents were in their 30s and 40s. It's possible that older respondents obtained their diagnosis later in life. Overall, there is a general trend of high agreement rates across all age groups, which indicates a positive attitude towards using ASD LAs regardless of age.

TABLE II.
SAMPLE CHARACTERISTICS (N = 395)

		Frequency	**Percentage**			**Frequency**	**Percentage**
Age	18–19	108	27.3	**Gender**	Male	211	53.4
	20–29	165	41.8		Female	184	46.6
	30–39	98	24.8	**Education level**	Never studied	9	2.3
	40–49	17	4.3		Elementary School	6	1.5
	50–59	7	1.8		Middle School	15	3.8
Time of Using Mobile Devices	1–3 hours	136	34.4		High school	136	34.4
	4–10 hours	159	40.3		Bachelor's degree	129	32.7
	11–20 hours	100	25.3		Master's degree	95	24.1
					PhD degree	5	1.3

In terms of gender, 53.4% of respondents were male, and 46.6 were female, as expected, given that the male population outnumbers the female population in Saudi Arabia [25]. Furthermore, the prevalence of autism in Saudi males is higher than in females [26]. This study reflects current gender demographics and autism prevalence rates in the Saudi context. It has been found that both genders generally express positive attitudes towards using ASD LAs.

Our study indicates a diverse range of educational backgrounds among individuals diagnosed with ASD. This variety of educational levels emphasises the necessity of understanding their unique challenges and the support they require, as the majority of respondents (34.4%) have a high school diploma, while 32.7% have a bachelor's degree. The relatively small percentage of respondents who never studied, attended elementary school, or went to middle school, with a total of 7.6%, indicates that the study sample mostly consists of individuals who have completed these early educational stages.

The survey analysis results indicated that participants with elementary school education levels have less intention of accepting ASD LAs adoption than others. This could be due to digital literacy skills, as people with limited education may face barriers to accessing and using technology [27]. Individuals with higher levels of education may have a better understanding of the potential benefits of ASD LAs in education, which may correspond with greater access to resources.

The widespread use of digital technologies in society is reflected in the high frequency of mobile device use among people with ASD. The examination of mobile device usage showed that the majority of respondents spend 1–10 hours per day on their devices, with 34.4% spending 1–3 hours. 40.3% of the examined population spends 4–10 hours, whereas 25.3% spend 11–20 hours. It is worth mentioning that excessive use of mobile devices can hinder ASD individuals' daily routines and their social interaction skills growth, leading to more severe issues such as isolation [28]. Overall, our data analysis results indicate that people engage with mobile devices in various ways. It also reveals the high agreement rates across different time categories (1–3 hours, 4–10 hours, and 11–20 hours), which indicates a positive intention to use autism learning applications.

B. Hypotheses fit testing

In this part, we evaluate the fit of our hypothesised framework to the actual data obtained from our survey. The framework was developed using theoretical considerations and past research to highlight the relationships between key factors influencing the adoption of ASD LAs among individuals with ASD in Saudi Arabia.

First, a series of statistical studies were performed to determine the fit of each hypothesised relationship between structural factors (e.g., technical, cultural, and pedagogical) and ASD users' perceptions. These analyses included correlation coefficients and significance tests to examine the strength of the relationships proposed in our framework. The outcomes of our hypothesis fit testing offer valuable insights into the sufficiency of our theoretical framework. All of the hypotheses, H1 to H16, tested were formulated in our previous paper along with the ASDLAF framework development [20].

Usability (US)

The study indicates that there is a statistically significant moderately positive relationship between usability (US) and the intention to use ASD LAs (AL) with a correlation coefficient of 0.313 and a significance level (p-value) of less than 0.001. In other words, as the US of ASD LAs improves, individuals with ASD are more likely to use them. This underscores the importance of designing ASD learning applications with more US consideration by actively involving individuals with ASD in the design process and considering their unique needs. The findings support the hypothesised relationship between US and AL: *"H1. If ASD LAs' usability is increased, then this will lead to ASD users' increased intention to the adoption of ASD LAs."*

Trust and acceptance (TA)

It appears to be a strong positive relationship between the intention to utilise ASD LAs (AL) and trust and acceptance (TA) of the applications, as indicated by a correlation value of 0.448 at a significance level of less than 0.001. This suggests that there is an enormous increase in the intention of

ASD users to adopt these applications as trust and acceptance of ASD learning applications grow. It is crucial for developers and stakeholders to focus on building trust when designing ASD LAs, which can enhance the acceptance of these applications. Developers can ensure ASD LAs align with best practices by collaborating with ASD professionals to understand important considerations such as ethical guidelines and ASD users' privacy. The findings support the hypothesised relationship between TA and AL: *"H2. If ASD LAs' trust and acceptance is increased, then this will lead to ASD users' increased intention to the adoption of ASD LAs."*

Effectiveness (EF)

The findings reveal a strong positive correlation (*rs* = 0.414, $p < 0.001$) between the effectiveness (EF) of ASD LAs and the intention to use these applications, shedding light on a critical relationship in educational technology adoption within the ASD community. This result implies that as the effectiveness (ED) of ASD learning applications increases, there is a corresponding increase in the intention of individuals with ASD to adopt LAs. Computer-assisted learning (CAL) approaches can be effective in teaching social skills [29]. The findings support the hypothesised relationship between EF and AL: *"H3. If ASD LAs' effectiveness is increased, then this will lead to ASD users' increased intention to the adoption of ASD LAs."*

Availability (AV)

There is a statistically significant positive link ($p < 0.001$) between the intention to use ASD learning applications (AL) and availability (AV), as indicated by the observed correlation value of 0.454. This implies that there is an evident rise in the intention of ASD (AL) users to adopt these applications when the availability of ASD LAs increases. In addition to focusing on the needs of individuals with ASD, it is essential to support networks surrounding them by providing training and guidance for educators and caregivers on how to effectively integrate ASD LAs into educational and therapeutic settings. This will increase the availability of these applications in practical terms. The findings support the hypothesised relationship between AV and AL: *"H4. If ASD LAs are available to function, then this will lead to ASD users' increased intention to the adoption of ASD LAs because ASD users will trust that this service will be always available to use."*

Accessibility (AC)

The survey analysis results revealed a moderately positive correlation coefficient of 0.157 between accessibility (AC) and the intention to use ASD learning applications (AL) with a significance level of 0.002, indicating the vital relationship between AC and user intention to use ASD LAs. This alignment suggests that enhancing the AC will increase the intention of individuals with ASD to utilise these applications. Considering accessibility features such as visual support, customisation, and sensor-friendly design can ultimately contribute to creating a promising learning tool. The findings support the hypothesised relationship between AV and AL: *"H5. If ASD LAs are accessible properly for autistic users, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will use these applications more easily."*

Privacy (PR)

There is a weak positive relationship between privacy (PR) and the intention to use ASD learning applications (AL), as suggested by the obtained correlation coefficient ($r = 0.123$). This relationship has a significance score of 0.015, indicating statistical significance. The effect size is modest, suggesting that privacy concerns alone may not be the primary driver behind users' intention to use ASD LAs. Moreover, the modest correlation implies that users might subconsciously prioritise other factors over privacy. Users with limited technical expertise might find it difficult to assess the benefits and drawbacks of particular privacy measures, which would lead to a weaker correlation with adoption intention. The findings support the hypothesised relationship between PR and AL: *"H6. If ASD LAs' privacy is increased, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will trust that their personal data will be secured."*

Cost (CO)

Understanding the relationship between the cost (CO) of ASD LAs and the intention to use them is crucial in creating effective technological solutions for individuals with ASD. The findings of the survey revealed that a notable correlation coefficient of 0.174 was observed, with a statistically significant p-value of less than 0.001. These findings also indicate a positive relationship between the CO and AL, underscoring the essential role of affordability in technology adoption within the context of supporting individuals with ASD. Simply put, as the cost (CO) of ASD learning applications decreases or remains reasonable, there is an increase in the intention of ASD (AL) users to adopt these applications. This result is consistent with economic theories of demand, which suggest that lower costs encourage higher adoption rates [30]. The findings support the hypothesised relationship between CO and AL: *"H7. If ASD LAs are at a free or reasonable cost, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will accept these applications available with low cost."*

Social rules (SR)

The findings show a correlation coefficient of 0.087, with a significance level of 0.083, suggesting an insignificant relationship between social rules (SR) and the intention to use ASD LAs (AL). The result fails to reach conventional levels of significance ($p < 0.05$), implying that social rules may not be considered a substantial influence on the intention of individuals with ASD to engage with LAs designed specifically for their needs. Social rules might be a major barrier for certain people to use ASD LAs, but they might not have much of an impact on other individuals' decisions to accept technology; individuals are not the same [31]. This diversity can result in a wider range of responses, leading to a weaker correlation value. The findings reject the hypothesised relationship between SR and AL: *"H8. If social rules prevent ASD users* *to use learning applications, then this will lead to ASD users' decreased intention to the adoption of ASD LAs because they will not socially accept these applications."*

Awareness (AW)

The correlation coefficient of 0.160 and a significance level of 0.001 suggest a weak positive correlation between awareness (AW) and the intention to engage with ASD learning applications (AL). This implies that individuals with ASD have the intention to be more interested in using LAs as autism awareness grows. This finding highlights the value of raising awareness of ASD and the advantages of using particular LAs for those who are on the spectrum. Awareness is not just about understanding ASD individuals—it's about enhancing their quality of life [32], by increasing their empowerment, inclusion, and access to resources that support ASD individuals' unique way of learning. The findings support the hypothesised relationship between AW and AL: *H9. "If awareness of autism is increased, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will have a better understanding to accept these LAs."*

Education (ED)

The results indicate a weakly positive correlation between education (ED) and the intention to use ASD learning applications (AL), with a correlation coefficient value of 0.211 and a statistically significant p-value of less than 0.001. In other words, as the level of education of ASD users increases, there is also a significant increase in their interest in using these applications. Higher education levels might be associated with a better appreciation of the potential benefits of ASD LAs in supporting their learning skills. Furthermore, individuals with higher education levels may be more proficient in using technology or are more likely to use ASD learning applications. The findings support the hypothesised relationship between CO and AL: *H10. "If ASD users are educated, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will have proper skills and accept these LAs."*

Religion (RE)

The results revealed that the correlation coefficient is 0.019, suggesting a very weak relationship between religion (RE) and the intention to use ASD learning applications (AL). Furthermore, the high p-value of 0.706 indicates that this correlation is not statistically significant at the conventional level of 0.05. Religion is a complex phenomenon, as there are different religious traditions and beliefs that have been practiced for centuries. The findings reject the hypothesised relationship between SR and AL: *H11. "If religious rules limit the use of ASD LA, then this will lead to ASD users' decreased intention to the adoption of ASD LAs because they will not accept these LAs."*

Responsivity (RS)

The result of the correlation test showed a positive relationship between RS and AL, = 0.206, p = <0.001. This indicates that users with ASD tend to adopt ASD learning applications that have more responsive interfaces. Developers can use this finding to prioritise features that enhance the responsivity within ASD learning applications, such as customisable settings, clear instructions, and sensory-friendly design elements. Individuals with ASD often have unique preferences [33] and various learning styles; thus, it is important to consider these individual differences and design proper applications that meet the specific needs of each user. The findings support the hypothesised relationship between RS and AL: *"H12. If ASD LAs provide proper responsive interfaces, then this will lead to ASD users' increased intention to the adoption of ASD LAs because ASD LAs will be effective and easy to use."*

Motivation (MO)

The survey analysis results revealed a strong positive correlation coefficient of 0.531 between motivation (MO) and the intention to use ASD learning applications (AL) with a significance level of <.001. It suggests that the motivation factor acts as a driving force behind the intention to adopt and use ASD learning applications. People with ASD can certainly learn continuously in a comfortable environment, thanks to the revolution of current technological advancements such as artificial intelligence and augmented reality. These trending technologies can enhance the learning experience of ASD individuals and improve their social skills and communication abilities by creating virtual scenarios for practicing real-life interactions in a safe setting. The findings support the hypothesised relationship between RS and AL: *H13. "If ASD users have motivations to use learning applications, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will use these LAs effectively."*

Age (AG)

Despite the weak positive correlation, rs = 0.092, between age (AG) and the intention to use ASD learning applications (AL), the correlation is not statistically significant at the conventional level of 0.05. While the correlation coefficient indicates a weak relationship, there might still be some influence of age on the intention to use ASD learning applications. However, in today's digital era, age has become less restricted by traditional age principles. Nevertheless, the digital age has also allowed older individuals to adapt to new technologies. It can be concluded that there is no significant relationship between age and the intention to use ASD learning applications based on the data collected; thus, the findings reject the hypothesised relationship between RS and AL: *"H14. If ASD LAs are suitable for adultescent age, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will find these LAs effective and easy to use."*

Learning behaviour (LB)

The correlation coefficient of 0.209, with a significance level of $p < 0.001$, emphasises a positive statistically significant relationship between learning behaviour (LB) and (AL). This correlation suggests that certain learning behaviours may encourage individuals with ASD towards the adoption of

ASD LAs. More importantly, technology can be a promising tool to provide a proper learning approach, as visual processing is a powerful choice for people with ASD [34]. Therefore, individuals with ASD can improve their communication skills and engage more effectively in social interactions in a safe environment. The findings support the hypothesised relationship between RS and AL: *"H15. If ASD LAs are aligned with ASD individuals learning behaviours, then this will lead to ASD users' increased intention to the adoption of ASD LAs because they will use these LAs effectively and easily."*

Readiness (RD)

The intention to use ASD learning applications (LA) and readiness (RD) appear to have a positive relationship, as indicated by the correlation coefficient of 0.214, with a significance level of less than 0.001, indicating a statistically significant relationship. Simply put, when individuals feel confident enough to use LAs, they are more likely to have the intention to adopt them. Therefore, this intention to adopt can lead to increased usage, which can ultimately result in the individual experiencing the full benefits of LAs. Using ASD LAs in their daily routine can contribute to improving their overall personal skills. The findings support the hypothesised relationship between RS and AL: *"H16. If ASD users are ready to use LA independently, then this will lead to ASD users' increased intention to the adoption of ASD LA because they will use these LA effectively and easily."*

This study discusses the critical factors that influence the adoption of ASD LAs by examining the validity of our proposed framework, ASDLAF. Statistical analyses were conducted to evaluate the relationships between various factors and users' intentions to adopt ASD LAs. As shown in Table III, the findings indicate significant positive correlations between usability, trust and acceptance, effectiveness, availability, accessibility, motivation, learning behaviour, and readiness and the intention to use ASD LAs. Thus, developers and stakeholders are encouraged to prioritise these factors to increase the intention of ASD individuals to adopt the LAs.

Factors such as readiness and motivation are crucial in driving the intention to adopt ASD LAs. These two factors can also be important incentives that increase the independence and empowerment of the ASD individuals. However, social rules, religion, and age factors showed weaker or insignificant correlations, which made their impact less pronounced. These factors may not directly influence the decisions to use LAs since ASD individuals and their families may prioritise immediate factors such as usability and effectiveness. despite that, these factors should also be considered when designing LAs as they may have significant indirect implications. The updated framework that is presented in Fig. 2, has been revised after the evaluation of the factors in ASDLAF [20] and the results of the hypotheses testing.

Developers can create ASD learning applications that are not only effective in educational settings but also a tool that will be considered supportive for those ASD individuals requiring unique needs. These applications should be tailored to each individual's specific learning style and needs to increase the level of LAs effectiveness. For example, individuals with ASD have different sensory preferences. When it comes to the use of smart phones, it is crucial to consider the need to customise these stimuli to avoid overwhelming users who might be sensitive to bright lights, sudden noises, or persistent vibrations. Thus, some adjustments should be applied such as having a customisable alerts for the sudden noise issue, and using gentle vibrations in order to decrease the defects of persistent vibrations [35]. The sensory considerations are crucial

TABLE III.
SUMMARY OF THE HYPOTHESES TESTS RESULTS

Factors	Hypotheses test	Relationship direction	Relationship Strength
US → AL	Supported	Positive	Moderate (rs = 0.313 **) sig
TA → AL	Supported	Positive	Strong (rs = 0.448 **) sig
EF → AL	Supported	Positive	Strong (rs = 0.414 **) sig
AV → AL	Supported	Positive	Strong (rs = 0.454 **) sig
AC → AL	Supported	Positive	Weak (rs = 0.157 **) sig
PR → AL	Supported	Positive	Weak (rs = 0.123 **) sig
CO → AL	Supported	Positive	Weak (rs = 0.174 **) sig
SR → AL	Rejected	Positive	No Significant Relationship
AW → AL	Supported	Positive	Weak (rs = 0.160 **) sig
ED → AL	Supported	Positive	Weak (rs = 0.211 **) sig
RE → AL	Rejected	Positive	No Significant Relationship
RS → AL	Supported	Positive	Weak (rs = 0.206 **) sig
MO → AL	Supported	Positive	Strong (rs = 0.531 **) sig
AG → AL	Rejected	Positive	No Significant Relationship
LB → AL	Supported	Positive	Weak (rs = 0.209 **) sig
RD → AL	Supported	Positive	Weak (rs = 0.214 **) sig

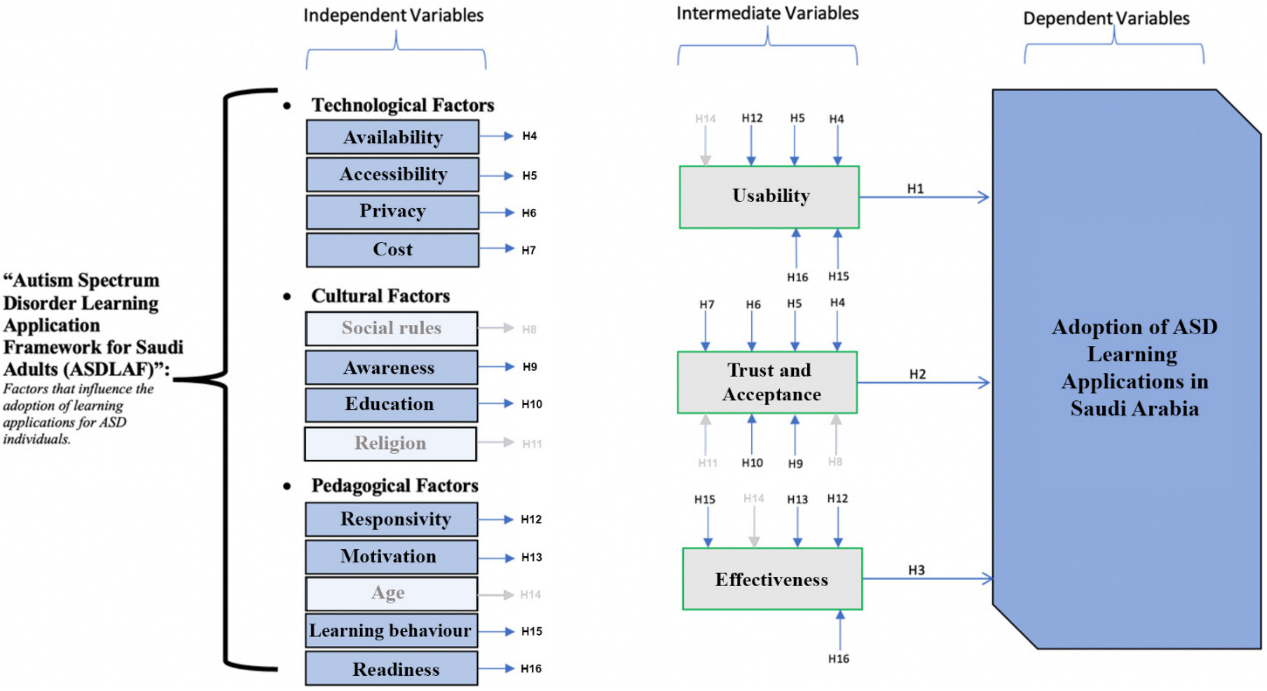


Fig. 2: The revised framework

to increase the motivation of ASD individuals to use technology as a supportive tool.

Another important consideration is the applications usability as our survey analysis found that usability is a vital factor in adopting the use of ASD LAs. A user-friendly interface is essential for facilitating ease of use and reducing frustration, which can impact the overall user experience of the users. Designing a user-friendly interfaces can help ASD users interact ideally [36], by creating a simple and predictable interfaces with clear instructions. The usability of LAs plays a significant role in how well the application is utilized by its target users. Therefore, how ready ASD users are to use LAs and the level of their learning behaviour are also considered key factors in increasing the likelihood of ASD users accepting and engaging with the application.

Additionally, this study emphasises the prominence education and awareness in enhancing the learning experiences of individuals with ASD. Several studies [13][33] have revealed that, in Arab countries, individuals with ASD are not provided with proper support. This could be due to the lack of awareness in this region. Participants of our survey have emphasised that better awareness leads to increase their intention to adopt ASD LAs.

The financial aspect is another essential consideration for ASD individuals and their families as some of them face significant financial burdens due to therapeutic and medical expenses. Thus, LAs developers should consider their price models so that users can afford using them. Plus, providing trial demos can allow the users to evaluate the effectiveness of the applications features before purchasing. Moreover, developers may collaborate with non-profit organizations to provide a financial assistance for the families who cannot afford the applications. During the transition to adulting, the cost of providing care for ASD individuals is high [38], as well as the medical expenses [39]; thus, keeping the price of applications at a reasonable range will greatly increase their acceptance among ASD society.

As stated earlier, privacy is a valuable factor in the Saudi society [19], let alone the unique sensitivities that ASD individuals usually have. Maintaining sensitive information such as sexual, physical, psychological information is considered very vital and exposing them could impact their dignity[40]. Therefore, developers must priorities this ethical consideration by involving ASD individuals, their families, and caregivers in the design process, which will increase the trust and acceptance of the ASD LAs. This involvement will help in creating a more personalised and effective learning experience and also minimise the risk of information exposure and unauthorised access.

V. CONCLUSION

In conclusion, this study sheds light on the critical factors influencing the adoption of Autism Spectrum Disorder Learning Applications (ASD LAs) among individuals in Saudi Arabia. By conducting a quantitative survey analysis, we validated the proposed Autism Spectrum Disorder Learning Application Framework (ASDLAF) and tested 16 hypotheses. This research provides valuable insights into the experiences of individuals with ASD regarding ASD LAs.

These findings highlight the importance of factors such as usability, trust and acceptance, effectiveness, availability, accessibility, motivation, learning behaviour, and readiness.

Whereas the impact of the social rules, religion, and age factors is less significant in determining the success of this type of technology. Overall, these results emphasise the need for further research and development in the field of technology adoption to gain a deeper comprehension of the factors that influence individuals' decisions to accept new innovations.

The future work involves a qualitative interview analysis, which can offer deeper insights into the experiences, perspectives, and challenges faced by individuals with Autism Spectrum Disorder (ASD) regarding the adoption of Learning Applications (LAs) in Saudi Arabia. The interview will be conducted with ASD caregivers, providing rich data that can support the quantitative findings.

REFERENCE:

[1] I. M. Hopkins *et al.*, ‘Avatar Assistant: Improving Social Skills in Students with an ASD Through a Computer-Based Intervention’, *J Autism Dev Disord*, vol. 41, no. 11, pp. 1543–1555, Nov. 2011, doi: 10.1007/s10803-011-1179-z.

[2] J. Bishop, ‘The Internet for educating individuals with social impairments: Educating individuals with social impairments’, *Journal of Computer Assisted Learning*, vol. 19, no. 4, pp. 546–556, Dec. 2003, doi: 10.1046/j.0266-4909.2003.00057.x.

[3] S. M. Al-Salehi, E. H. Al-Hifthy, and M. Ghaziuddin, ‘Autism in Saudi Arabia: Presentation, Clinical Correlates and Comorbidity’, *Transcult Psychiatry*, vol. 46, no. 2, pp. 340–347, Jun. 2009, doi: 10.1177/1363461509105823.

[4] P. T. Shattuck, A. M. Roux, L. E. Hudson, J. L. Taylor, M. J. Maenner, and J.-F. Trani, ‘Services for Adults with an Autism Spectrum Disorder’, *Can J Psychiatry*, vol. 57, no. 5, pp. 284–291, May 2012, doi: 10.1177/070674371205700503.

[5] M. J. Maenner *et al.*, ‘Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016’, *MMWR Surveill. Summ.*, vol. 69, no. 4, pp. 1–12, Mar. 2020, doi: 10.15585/mmwr.ss6904a1.

[6] S. Camarata, ‘Early identification and early intervention in autism spectrum disorders: Accurate and effective?’, *International Journal of Speech-Language Pathology*, vol. 16, no. 1, pp. 1–10, Feb. 2014, doi: 10.3109/17549507.2013.858773.

[7] A. Seif Eldin *et al.*, ‘Use of M-CHAT for a multinational screening of young children with autism in the Arab countries’, *International Review of Psychiatry*, vol. 20, no. 3, pp. 281–289, 2008, doi: 10.1080/09540260801990324.

[8] S. Bölte, O. Golan, M. S. Goodwin, and L. Zwaigenbaum, ‘What can innovative technologies do for Autism Spectrum Disorders?’, *Autism*, vol. 14, no. 3, pp. 155–159, May 2010, doi: 10.1177/1362361310365028.

[9] O. Golan and S. Baron-Cohen, ‘Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia’, *Develop. Psychopathol.*, vol. 18, no. 02, Jun. 2006, doi: 10.1017/S0954579406060305.

[10] A. L. Wainer and B. R. Ingersoll, ‘The use of innovative computer technology for teaching social communication to individuals with autism spectrum disorders’, *Research in Autism Spectrum Disorders*, vol. 5, no. 1, pp. 96–107, Jan. 2011, doi: 10.1016/j.rasd.2010.08.002.

[11] A. Rapp, F. Cena, R. Castaldo, R. Keller, and M. Tirassa, ‘Designing technology for spatial needs: Routines, control and social competences of people with autism’, *International Journal of Human-Computer Studies*, vol. 120, pp. 49–65, Dec. 2018, doi: 10.1016/j.ijhcs.2018.07.005.

[12] C. E. Burton, D. H. Anderson, M. A. Prater, and T. T. Dyches, ‘Video Self-Modeling on an iPad to Teach Functional Math Skills to Adolescents With Autism and Intellectual Disability’, *Focus Autism Other Dev Disabl*, vol. 28, no. 2, pp. 67–77, Jun. 2013, doi: 10.1177/1088357613478829.

[13] M. A. Prater, N. Carter, C. Hitchcock, and P. Dowrick, ‘Video self-modeling to improve academic performance: A literature review’, *Psychol. Schs.*, vol. 49, no. 1, pp. 71–81, Jan. 2012, doi: 10.1002/pits.20617.

[14] S. A. Samadi and R. McConkey, ‘Autism in Developing Countries: Lessons from Iran’, *Autism Research and Treatment*, vol. 2011, pp. 1–11, 2011, doi: 10.1155/2011/145359.

[15] C. E. Hill, K. D. Loch, D. Straub, and K. El-Sheshai, ‘A Qualitative Assessment of Arab Culture and Information Technology Transfer’:, *Journal of Global Information Management*, vol. 6, no. 3, pp. 29–38, Jul. 1998, doi: 10.4018/jgim.1998070103.

[16] A. Al-Krenawi, ‘Mental health practice in Arab countries’, *Current Opinion in Psychiatry*, vol. 18, no. 5, pp. 560–564, Sep. 2005, doi: 10.1097/01.yco.0000179498.46182.8b.

[17] R. E. Gearing, M. J. MacKenzie, R. W. Ibrahim, K. B. Brewer, J. S. Batayneh, and C. S. J. Schwalbe, ‘Stigma and Mental Health Treatment of Adolescents with Depression in Jordan’, *Community Ment Health J*, vol. 51, no. 1, pp. 111–117, Jan. 2015, doi: 10.1007/s10597-014-9756-1.

[18] A. Al-Wabil, H. Al-Shabanat, R. Al-Sarrani, and M. Al-Khonin, ‘Developing a Multimedia Environment to Aid in Vocalization for People on the Autism Spectrum: A User-Centered Design Approach’, in *Computers Helping People with Special Needs*, vol. 6180, K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, Eds., in Lecture Notes in Computer Science, vol. 6180. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 33–36. doi: 10.1007/978-3-642-14100-3_6.

[19] T. abu-Gazzeh, ‘Privacy as the Basis of Architectural Planning in the.pdf’, *Architecture & Behaviou*, vol. 11 (3–4), pp. 269–288, 1996.

[20] Y. Almazni, N. Beloff, and M. White, ‘ASDLAF: A Novel Autism Spectrum Disorder Learning Application Framework for Saudi Adults’, in *Intelligent Systems and Applications*, vol. 542, K. Arai, Ed., in Lecture Notes in Networks and Systems, vol. 542. , Cham: Springer International Publishing, 2023, pp. 442–458. doi: 10.1007/978-3-031-16072-1_33.

[21] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis: A global perspective*, 7th ed. London: Pearson, 2010.

[22] K. White, W. R. Borg, and B. R. Worthen, *measurement and evaluation in the schools.pdf*. New York: Longman, 1993.

[23] General Authority for Statistics, ‘Population Summary Report’, Riyadh, Saudi Arabia, statistic 2, 2022.

[24] A. Alkhateeb, R. Alboali, W. Alharbi, and O. Saleh, ‘Smartphone addiction and its complications related to health and daily activities among university students in Saudi Arabia: A multicenter study’, *J Family Med Prim Care*, vol. 9, no. 7, p. 3220, 2020, doi: 10.4103/jfmpc.jfmpc_1224_19.

[25] The General Authority for Statistics (GAStat), ‘Population in Kingdom by Gender, Age Group,and Nationality (Saudi/Non-Saudi)’, 2019. [Online]. Available: https://www.stats.gov.sa/en/1007-0

[26] A. Khan, H. AlGhadeer, A. Al-Qassimi, T. Al-Jubran, H. Al-Momen, and M. Al-Nazzal, ‘Autism in Saudi Arabia, a challenge to Saudi families: a cross-sectional study’, *IJMDC*, pp. 1453–1458, 2020, doi: 10.24911/IJMDC.51-1595277794.

[27] N. Khanlou, A. Khan, L. M. Vazquez, and M. Zangeneh, ‘Digital Literacy, Access to Technology and Inclusion for Young Adults with Developmental Disabilities’, *J Dev Phys Disabil*, vol. 33, no. 1, pp. 1–25, Feb. 2021, doi: 10.1007/s10882-020-09738-w.

[28] P. Ghanouni, T. Jarus, J. G. Zwicker, and J. Lucyshyn, ‘The Use of Technologies Among Individuals With Autism Spectrum Disorders: Barriers and Challenges’, *J Spec Educ Technol*, vol. 35, no. 4, pp. 286–294, Dec. 2020, doi: 10.1177/0162643419888765.

[29] S. Ramdoss, W. Machalicek, M. Rispoli, A. Mulloy, R. Lang, and M. O’Reilly, ‘Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: A systematic review’, *Developmental Neurorehabilitation*, vol. 15, no. 2, pp. 119–135, Apr. 2012, doi: 10.3109/17518423.2011.651655.

[30] A. Marshall, *Principles of Economics*. Prometheus, 1997.

[31] D. W. Dew and G. M. Alan, ‘Rehabilitation of Individuals with Autism Spectrum Disorders.pdf’, Institute on Rehabilitation Issues Monograph No. 32, Washington, DC: The George Washington

University, Center for Rehabilitation Counseling Research and Education, 2007.

[32] M. A. Alsehemi, M. M. Abousaadah, R. A. Sairafi, and M. M. Jan, ‘Public awareness of autism spectrum disorder’, *NSJ*, vol. 22, no. 3, pp. 213–215, Jul. 2017, doi: 10.17712/nsj.2017.3.20160525.

[33] S. R. Colorosa and C. J. Makela, ‘Integrative Literature Review-Styles of Learning for Autism Spectrum Disorders and Human Resource Development- Informing Performance Management’, 2014.

[34] J. M. Cafiero, ‘Technology Supports for Individuals with Autism Spectrum Disorders’, *J Spec Educ Technol*, vol. 27, no. 1, pp. 64–76, Mar. 2012, doi: 10.1177/016264341202700106.

[35] A. M. O’Brien, R. W. Schlosser, C. Yu, A. A. Allen, and H. C. Shane, ‘Repurposing a Smartwatch to Support Individuals With Autism Spectrum Disorder: Sensory and Operational Considerations’, *J Spec Educ Technol*, vol. 36, no. 4, pp. 215–226, Dec. 2021, doi: 10.1177/0162643420904001.

[36] S. Khan, M. N. Tahir, and A. Raza, ‘Usability issues for smartphone users with special needs &#x2014; Autism’, in *2013 International Conference on Open Source Systems and Technologies*, Lahore, Pakistan: IEEE, Dec. 2013, pp. 107–113. doi: 10.1109/ICOSST.2013.6720615.

[37] Y. M. Al-Farsi, M. M. Al-Sharbati, O. A. Al-Farsi, M. S. Al-Shafaee, D. R. Brooks, and M. I. Waly, ‘Brief Report: Prevalence of Autistic Spectrum Disorders in the Sultanate of Oman’, *J Autism Dev Disord*, vol. 41, no. 6, pp. 821–825, Jun. 2011, doi: 10.1007/s10803-010-1094-8.

[38] M. L. Ganz, ‘The Lifetime Distribution of the Incremental Societal Costs of Autism’, *Archives of Pediatrics & Adolescent Medicine*, vol. 161, no. 4, pp. 343–349, Apr. 2007, doi: 10.1001/archpedi.161.4.343.

[39] A. V. S. Buescher, Z. Cidav, M. Knapp, and D. S. Mandell, ‘Costs of Autism Spectrum Disorders in the United Kingdom and the United States’, *JAMA Pediatrics*, vol. 168, no. 8, pp. 721–728, Aug. 2014, doi: 10.1001/jamapediatrics.2014.210.

[40] F. Mohammadi, M. Rakhshan, Z. Molazem, N. Zareh, and M. Gillespie, ‘Caregivers’ perception of dignity in teenagers with autism spectrum disorder’, *Nurs Ethics*, vol. 26, no. 7–8, pp. 2035–2046, Nov. 2019, doi: 10.1177/0969733018796679.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 19–28
DOI: 10.15439/2024F8469
ISSN 2300-5963 ACSIS, Vol. 41

DTA-SAU: A Conceptual Framework for Digital Transformation Adoption in Saudi Arabian Universities

Rasha Alsulami
0009-0002-0898-3999
Department of Information Systems Management, Jeddah University, Jeddah, Saudi Arabia
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: ra452@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: n.beloff@sussex.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: m.white@sussex.ac.uk

Abstract—**Digital transformation in universities is reshaping academic landscapes globally, and Saudi Arabia. Universities are keen to provide high quality learning using different types of teaching methods and up-to-date technology. Thus, the shift to digitisation in Saudi universities is unavoidable, and must embrace digital transformation to satisfy students requirements. Digital transformation can promote learning and teaching through using digital tools. A literature review revealed that although there are few studies conducted on digital transformation adoption in Saudi universities, to date, there is a lack of coherent and comprehensive theoretical approach to this topic, particularly regarding adoption barriers. Our study aims to address this gap by presenting a conceptual framework for Digital Transformation Adoption in Saudi Arabian Universities (DTA-SAU), which combines the affecting factors. A mixed method (qualitative and quantitative) approach will be used to validate the DTA-SAU framework. The results will assist universities in developing countries in digital transformation adoption process**.

Index Terms—**digital transformation, technology adoption models, Saudi Arabian universities.**

I. INTRODUCTION

THE GLOBAL economy has been impacted by changes that have been unprecedentedly swift [1]. In order to stay competitive and survive in a fast-moving business environment, governments as well as companies and corporations have had to adapt and adopt a mindset that is open to implementing change [2]. With these changes in mind, both public and private sector organisations have implemented digital initiatives [3]. It is generally accepted that digital technology can be usefully deployed to enhance workplace processes and make workflow more efficient, faster and easier.

In Saudi Arabia, some of these digital initiatives have been implemented to transform a number of governmental sectors such as telecommunications, education and healthcare. This has been carried out with the aim of enhancing operations and boosting business value, as well as achieving a development that is sustainable, globally effective and which raises the amount that the digital economy contributes to Saudi Arabia GDP, and ultimately improves Saudi citizens' quality of life.

Saudi Vision 2030 clearly lays out objectives for the digital transformation of the education system, and to this end, the government have been striving to ensure those goals are met. In 2018, the Saudi Ministry of Education inaugurated the 'Future Gate' project, in collaboration with the educational technologies company TETCO. The project aims to establish digital technologies throughout the Saudi educational system, including the digitisation of curricula and establishing smart classrooms to replace paper documents and traditional classrooms [4]. This initiative aims to enhance student-teacher interaction, facilitate a combination of in-person and online learning, and improve the overall educational experience and achievements of young learners.

It worth documenting that digital transformation plays a crucial role in driving the future of higher education institutions because it focuses on developing an innovative process regarding learning, teaching, library facilities, blackboard, and other learning environment [5]. This, in fact, enables universities to drive digital transformation to improve the learning process and use online teaching facilities. For businesses, organizations, government organizations and academic institutions, digital transformation is currently seen as an unavoidable trend. Organizations are being forced by this trend to change how they handle operations, the provision of goods and services, marketing initiatives, and every other aspect of the businesses.

Numerous advantages come with digital transformation, including lower operational expenses, a larger consumer base, and quicker and more accurate decision-making. Moreover, adopting online learning reduces costs and develops students and faculty in the use of computing technology facilities. Universities are organizations that do research, offer educational services, and advance human knowledge. Therefore, such institutions may not be able to draw in, train, and lead scientists, students, and enterprises without successfully implanting digital transformation. However, digital transformation faced all types of challenges that include technical aspects related to maintain reliable and valid digitalization system, organizational aspects represented in managing and implementing transformation plans, legal and

©2024, PTI

Topical area: Information Technology for Business and Society

security aspects as well as users such as, students, lecturers, human resources related obstacles [6]. Another major challenge identified is related to financing. A considerable amount of money is spent on digital equipment systems, conversion, and educational transformation[6]. Therefore, institutions must navigate these challenges carefully to realize the full potential of digital transformation.

Higher education institutions must overcome significant obstacles related to the digital transformation to incorporate digital tools and technology into teaching and learning processes as well as change and adjust current systems, processes, communication channels, and all other academic and administrative activities carried out by the institution [7]. Hence, decision making regarding adoption of digital transformation in Saudi universities entail identifying the main factors affecting the adoption process.

A) Research Questions

The study questions derived from the primary goal of the study as follows:

1. What are the major challenges and benefits of adopting Digital Transformation in Saudi Arabian Universities?
2. According to–Top Management members, Academic staff, IT employees and students, which factors influence Saudi Arabian Universities to adopt Digital Transformation, and to what extent?
3. What would be a conceptual framework for Digital Transformation Adoption in Saudi Arabian Universities?

By answering these questions, this study aims to create a framework to assist Saudi Arabian universities in understanding factors that impact the adoption of digital transformation and will provide a basis for researchers who are investigating this phenomenon not only at the Saudi level but at regional and international levels.

There are four sections to this paper. The first part reviews the literature in order to identify the most significant factors that have been found to influence the adoption of digital transformation. The second part presents the various technology adoption theories and framework, and this is followed by an outline of the framework used in the current study. The final section describes the methods utilised to collect the data and how the framework will be validated.

II. LITERATURE REVIEW

Several studies were carried out and reported in the literature to identify the main factors that influence the adoption of technology in the higher education and business sectors.

Successful digital transformation in universities depends on stakeholders' awareness of digital tools and technologies. Raising awareness can reduce resistance to change and develop a more positive attitude towards digital transition by helping to eliminate uncertainty and fear of the unknown [8]. Stakeholders understand the benefits and advantages of digital tools and how they are used to increase productivity, improve learning outcomes, or open new possibilities for research and cooperation [9]. It is worth mentioning that stakeholders are more likely to support the adoption and push for its successful implementation when they are aware of the potential advantages and positive effects of digital technologies.

One of the most influential factors found to impact the adoption of digital technologies has been top management support [10], [11]. Moreover, in the context of SMEs, top management support has been shown to be of particular significance in encouraging a positive attitude towards digital transformation [12]. In developing nations like Saudi Arabia, the ICT decision-maker is most frequently a member of a SME's senior management team, and the adoption of ICT is directly impacted by his or her support and decision [13]. It can be understood that top management support is the major factor in adopting digital transformation since decisions regarding funding and personnel cannot be made without support from top management.

Another serious barrier noted in previous studies is security and privacy, which are established worldwide issues for technology adoption and implementation [14]. A qualitative study conducted on the risks associated with participating in e-commerce found that clients are worried about the security and privacy of their data, such as emails and names, that could be used for marketing and other purposes [15]. It can be suggested that developing a high level of security and privacy is likely to improve the intention of institutions to implement digital transformation. A safe and secure environment fosters trust, safeguards sensitive data, and encourages stakeholders to embrace the changes wrought by digital transformation programmes.

The significance of providing benefits to users in digital transformation is embedded in users of digital technologies and tools. The benefits clients perceive positively affected digital transformation adoption in the United States business sector [16]. The usefulness of digital tools such as mobile technologies positively affected lecturers' perceptions of digital tools [17]. Giving benefits to users (usefulness) is likely to boost the desire of institutions to implement digital transformation. A user-centred strategy focusing on actual advantages for stakeholders fosters higher acceptance, utilisation, and support for digital technology, resulting in a successful and effective digital transformation throughout the institution.

Resistance to change is regarded as one of the most critical factors which prevent the adoption of digital transformation in higher education institutions. Changing the academic culture is complex and challenging in many cases, it is gradual [18]. For example, in a study conducted on higher education institutions in India, the authors claimed that many academic staff were not positive towards changing teaching methods or using new digital technologies, while some were reluctant to change [19]. Resistance to change is not only an issue in higher education institutions but also a global issue in all sectors. This could be attributed to the fact that resistance to change is a fear of uncertainty and lack of interest to change. Refuse to change may be attributed to particular habits, that refer to an individual's reaction to their traditional environment [20]. Instructors' resistance to change is the most critical element influencing digital transformation adoption. The capacity of teachers to adopt new technologies and services will contribute to adopting digital transformation successfully [21]. Finally, resolving resistance to change in digital technology is critical for increasing the intention to accept digital transformation in universities. Universities may build a more positive and supportive atmosphere by understanding and minimizing opposition and encouraging stakeholders to embrace and successfully use digital technology for the institution's progress.

Study by [22] was conducted at a British university and investigated the attitudes of 48 staff members to the adoption of e-learning. The study found that the most significant factors were staff skills and attitudes, student expectations and institutional infrastructure. Even in the context of this developed country, study participants complained of a lack of direction and specifically tailored support, which indicates that a strategy for providing such direction, resources and support is required in educational institutions wishing to implement e-learning.

Universities would like the benefits of adopting a new invention to be proportionate with the expenses of adopting the innovation. Digital transformation has reduced the costs of services provided to students and lecturers. A study conducted in Saudi Arabia [23] investigated the effect of adopting and integrating digital transformation on organisational and spending on digitalisation in Saudi universities. This study found that using digital transformation and services provided reduced students' withdrawal from the courses, which in turn positively affected the universities' academic performance. Although the study [23] investigated the students' and lecturers' views of digital transformation, it overlooked the technical issues related to required skills.

In brief, this literature review section helps emphasize the most important factors that influence digital transformation adoption. Additionally, identified significant gaps, such as a lack of studies on Saudi higher education. Furthermore, all previous research explored the elements that influence the adoption of digital transformation from a single or two points of view; in our study, we investigated these influencing factors from the perspectives of-students, academic staff, IT departments, and senior management. To close these gaps, building a conceptual framework for digital transformation in Saudi Arabian institutions has become an absolute imperative.

III. ADOPTION THEORIES AND FRAMEWORKS

Several frameworks and theories have been developed and refined to investigate and assess the factors that influence technology adoption in a particular sector. This study conducted a thorough review of the literature relevant to this research to determine the theories used to explore technology adoption by organisations. The findings of this investigation indicated the most widely employed theories:

- Technology Organization and Environment (TOE) framework [24], [25]
- Diffusion of Innovation (DOI) [26],
- TAM Integrated with TOE framework [27], [28]
- DOI integrated with TOE [29], [30], [31]
- Technology Acceptance Model (TAM) [32], [33].

There have been many studies on what influences the decision to use IT technologies [34]. This study will use the Technology Acceptance Model (TAM), a model developed by Davis [35], as well as the TOE framework developed by Tornatzky et al. [36]. Integrating the two frameworks improves the examination of the factors. The TOE and TAM work harmoniously as they explore universities' intention to adopt digital transformation from different aspects addressed by the technological, organisational the research will exclude the environmental context, but will add the social and financial context.

IV. PROPOSED CONCEPTUAL FRAMEWORK

This research has developed a new conceptual framework by combining the TAM model and TOE framework with some modifications to achieve our research objectives.

There are a number of technological, organisational and social concerns regarding the adoption of digital transformation at Saudi universities. The TAM (Technology Acceptance Model) and TOE (Technology – Organisation - Environment) frameworks were utilised in order to analyse the factors that influence the adoption of innovative digital technologies, as these were deemed to offer a comprehensive cover of the technological, organisational, social and financial contexts related to the research objectives. In tandem, the frameworks offer a means of obtaining insight into both individual factors and the socio-technical governance of Saudi universities that potentially influence acceptance and implementation of digital transformation. This insight provides Saudi university stakeholders the means with which to develop more effective ways to deploy digital transformation adoption initiatives that may lead to improving efficiency in administration and ultimately to better educational outcomes.

A. The Theory of Technology, Organisation, and Environment

The TOE Framework examines how an organisation adopts and uses new technology and how the environment might affect this [37]. It assesses adoption of current technologies in an organisation by separating the elements that influence the adoption process into three groups: technological context, organisational context, and environmental context. [37]. TOE is seen as an appropriate framework for investigating decisions to adopt innovation in organisations at the organisational, rather than user level. Many previous studies have employed TOE to examine the adoption of technological innovation [38], [39].

B. The Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM) is used extensively to describe individuals' intent to use new technology. It is based on the Theory of Reasoned Action (TRA) and its extension, the Theory of Planned Behaviour (TPB) [35]. The TAM focuses on perceived usefulness and ease of use, which can significantly affect digital transformation in Saudi universities. Many studies confirm that perceived ease of use positively impacts users' intent to use technology [40], [41]. Therefore, the more users believe that digital technology is easy to use, the more likely this technology is to be adopted.

C. The proposed conceptual research framework

We developed a framework called DTA-SAU: Framework of Digital Transformation Adoption in Saudi Arabian Universities. This is to investigate the factors that influence the adoption of digital transformation. Previous studies on this subject were built upon. This included models and theories that were frequently used to measure the various elements that impact new technology adoption. This section presents and describes the research framework and hypothesis; see Fig.1. In this research, four theoretical contexts with associated factors have been defined as elements that influence the adoption of digital transformation. The following contexts are considered:

1. Organisational Context
2. Social Context
3. Technological Context
4. Financial Context

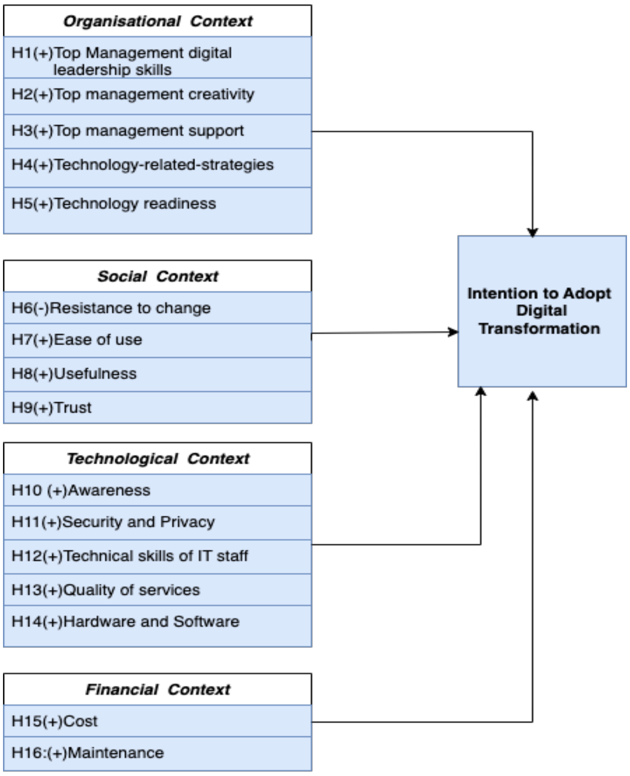

Organisational Context
H1(+)Top Management digital leadership skills
H2(+)Top management creativity
H3(+)Top management support
H4(+)Technology-related-strategies
H5(+)Technology readiness
Social Context
H6(-)Resistance to change
H7(+)Ease of use
H8(+)Usefulness
H9(+)Trust
Technological Context
H10 (+)Awareness
H11(+)Security and Privacy
H12(+)Technical skills of IT staff
H13(+)Quality of services
H14(+)Hardware and Software
Financial Context
H15(+)Cost
H16:(+)Maintenance
Intention to Adopt Digital Transformation

Fig 1.The proposed conceptual research framework of Digital Transformation Adoption in Saudi Arabian Universities (DTA-SAU)

1) The organizational context

a) Top Management digital leadership skills

Digital leadership skills is regarded as a critical factor in innovation process models in organisations. The process is deemed as something created and changed constantly with certain backgrounds. Digital leadership skills is, in fact, a mixture of leadership skills and digital skills which contribute to improving organisational performance [42]. Digital leadership positively affected organisational performance in Jordan [43]. Leaders’ skills are almost represented in data analysis, knowledge of web-development, programming language, financial management, operating systems, data security and knowledge of cloud computing. Many digital transformation projects failed because of the lack of leadership digital skills and knowledge about digital technologies and facilities [44], [45]. Many studies indicated the crucial role of top management digital leadership skills in the success of digital transformation [46], [47], [48]. Furthermore, some studies found a significant relationship between leaders’ digital skills and the adoption of digital transformation. The authors indicated that leadership technical competencies have successfully affected the process of digital transformation [49]. These results are consistent with other studies [50], [51], [52], [53]. Therefore, top management digital leadership skills are considered a potentially critical factor that influences the adoption of digital transformation, and this leads to the following hypothesis:

> *H1: A high level of* ***top management digital leadership skills*** *will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

b) Top management creativity

Literature documents that leaders know instinctively that creativity is the life blood of their organizations. This saying means that leaders’ creative data may lead to thinking about new programmers and action plans to adopt [54]. For example, they may adopt and integrate digital technologies into the organizational system. Therefore, leaders’ mission is finding out better ideas than the ideas their organizations are used to working with or committed to adopting. This, in fact, can be applied to the adoption of digital transformation in organizations and universities. Creativity is a manifold phenomenon, and humans can conceive new approaches in tangible products [55]. Creativity in organizations is the source of design, improves processes and can shape the work environment. Therefore, top management creativity is a potential important factor that influences adoption of digital transformation, this leads to the following hypothesis:

> *H2: A high level of* ***top management’ creativity*** *will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

c) Top management support

Top management support and the adoption of digital transformation focus on understanding the importance of digital tools, funding human resources and personnel initiatives, and participating in activities related to digital transformation [56]. It could be suggested that top management support is the primary factor in adopting digital transformation since decisions regarding funding and personnel can only be made with support from top management. Top management support consists of shaping and forming the organizational structure and context to make it more adaptable to adopt digital and technological tools to the characteristics of the organisation [11]. Therefore, universities must adapt technologies to students' and academic staff's needs, such as e-learning, blackboard, and interaction groups. Several studies ascertain that support provided by top management is one of the most important organizational factors that enhance the adoption of digital transformation [11], [12]. Top management support plays a crucial role in the success related to the adoption of digital transformation [57]. Therefore, the next hypothesis can be proposed:

H3: A high level of ***top management support*** *will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

d) Technology-related-strategies.

Information technology strategies are seen as one of the main aspects that affect the adoption of digital transformation [58]. Business strategies can lead to information technology strategies, but the two strategies should be aligned. Digital transformation strategies in organizations focus on activities that can respond to rapid changes in organizational performance and achievement. Strategies should be related to digital technologies such as cloud computing, big data and mobile technologies [45], [59]. Technology-related strategies are essential in achieving digital transformation in universities. In relation to universities, digital transformation strategies can be adopted in new online teaching methods, improving blackboard performance, learning environment, and creating cooperative learning between higher education institutions. Therefore, technology-related strategies are considered a potentially critical factor that influences the adoption of digital transformation, and this leads to the following hypothesis:

H4: Making ***Technology-related strategies****, will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

e) Technology readiness

Technology readiness refers when evaluating the suitability of network technologies and an organization's systems for adoption, it is essential to consider their level of maturity. Technology readiness substantially influences the leadership's decision-making process on the adoption of innovation [60]. The question that can be raised here is whether the universities are ready to face and adopt digital transformation in general. Universities' readiness for digital transformation focuses on students, academic staff, and information technology departments. Students and academic staff are the end users of technology tools and facilities, while IT staff are responsible for preparing to adopt digital technologies. Many studies showed that a robust ICT infrastructure is another critical factor for success in any organization [61], [62], [63]. In related research, it was discovered that adopting new technologies is aided by the readiness of ICT infrastructure [64]. These considerations have generated the following hypothesis:

H5: ***Technology readiness*** *will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

2) Social context

a) Resistance to change

It is crucial to address resistance to change when organizations intend to adopt digital transformation. This is very important when executing digital transformation efforts in organizations which include public institutions, and resistance to change is a typical obstacle. People may hesitate to adopt new digital technologies for various reasons, including apprehension about the unknown, doubts about their ability to use the devices and worry about potential disturbances to their routines. Resistance to change is regarded as one of the main factors which hinder the adoption of digital transformation in higher education institutions. Refusing to change obstructs ICT adoption and incorporation into education systems [65]. The mindset of teachers and their characteristic unwillingness to shift from the old way of doing things is an obstacle to ICT adoption [65]. Therefore, the hypothesis proposes that resistance to change in the use of digital technologies will have a negative effect on the intention to adopt digital transformation in Saudi Arabian.

H6: ***Resistance to change*** *in using digital technologies will decrease the intention to adopt* ***digital transformation*** *in Saudi Arabian universities.*

a) Ease of use

Usability in digital transformation is an important aspect of successfully adopting digital technology in any organization, especially Saudi Arabian. What degree are digital technologies are easy to use is one factor which affects the use of digital tools. For instance, technical support or IT departments need to make sure that the user interface of digital tools and apps is intuitive, visually appealing, and simple to use. User-friendly features and capabilities are also important, but users need little training and technical knowledge [66]. IT departments are required to make digital technologies accessible to all users, including those with varying abilities and requirements. This needs clear instructions and support, and thorough user support is provided to assist stakeholders in efficiently using digital tools. Users' feedback is critical, and IT departments must use different mechanisms to collect it and continually enhance the user experience [67]. Therefore, Saudi Arabian universities that prioritise the ease of use of digital technology are more likely to have a positive intention to adopt digital transformation. Thus, the next hypothesis can be proposed:

H7: A high level of ***ease of use*** *for digital technologies will increase the intention to adopt* ***digital transformation*** *in Saudi Arabian universities*

b) Usefulness

The importance of providing benefits to users in digital transformation is embedded in users of digital technologies and tools. A key element in digital transformation initiatives is user-centeredness. Universities are more likely to embrace and promote the implementation of digital technology when they focus on offering actual benefits and usefulness to their stakeholders, which include students, faculty, staff, and administrators. The benefits perceived by clients positively influenced the adoption of digital transformation in the US company sector [16]. In a survey on digital transformation in higher educational institutions in the United States [68], the authors found that respondents reported many benefits of digital technologies directly related to students' success in examinations [68]. The usefulness of digital tools such as mobile technologies, positively affected lecturer perception digital tools [17]. Therefore, according to the hypothesis, usefulness of digital tools to users is likely to have a positive intention to adopt digital transformation. Thus, the next hypothesis can be proposed:

*H8: **Usefulness** of digital tools will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

c) Trust

Trust is a critical component in digital transformation efforts in universities. Universities must have faith in digital tools and technology's dependability, security, and efficacy as they integrate them into their operations, teaching, research, and student services. Shifting from a social concept of trust to a digital one requires modelling and managing trust to build trustworthy digital systems [69]. A study was carried out on confidence in digital technologies among academic staff in universities. The authors found that lecturers still had no confidence in digital tools because there is still no confidence in pedagogical and educational thinking to integrate these technologies [70]. It is worth noting that ensuring service quality fosters stakeholder trust and satisfaction, promoting a supportive attitude towards digital transformation. Some confidence levels may affect the trust in digital tools and technologies. For example, the first level is reliability where stakeholders' users believe that digital technologies will perform as expected and provide consistent and accurate results [71]. The second level, trust in digital technologies will protect data integrity and prevent unauthorized modification or manipulation [72]. Therefore, according to the hypothesis, universities with a high level of trust in digital tools are more likely to have a positive intention to adopt digital transformation.

*H9**: Users' Trust** in digital technologies will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

3) Technological Context

a) Awareness

The successful implementation of digital transformation in universities depends on stakeholders' awareness of digital tools and technologies. Students, teachers, staff, administrators, and other university constituents impacted by digital transformation projects are stakeholders. Stakeholders are more likely to be open to change and supportive of transformation projects when they comprehensively understand the advantages, functions, and potential applications of digital tools [9].The rationale behind this is that by raising the awareness of students, teachers, staff, and administrators about the benefits and possibilities of digital tools and technology, universities can build enthusiasm for embracing these changes-Therefore, increasing stakeholders' awareness of digital tools and technology is a crucial first step towards fostering the intention to adopt digital transformation in Saudi Arabian universities. Thus, the next hypothesis can be proposed:

*H10: **Stakeholders' awareness** of digital tools will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

b) Security and privacy

Security and privacy are paramount when integrating digital technologies into university operations. Students, faculty members and administrative staff want assurance that their personal information will be kept confidential and secure. A study in Algeria confirmed that the adoption behaviour of e-commerce is mostly influenced by security challenges and system risk [73]. To address this concern, universities must implement strong cybersecurity measures, adhere to privacy regulations, and ensure secure data handling. Corporate privacy policies should align with customer requirements to foster trust, which can encourage online engagement [74]. Moreover, universities should acknowledge and mitigate the risks associated with transferring money online, and the security and privacy of personal data should be of paramount concern. Therefore, according to the hypothesis, universities that prioritise building a high level of security and privacy are more likely to have a positive intention to adopt digital transformation.

*H11: Developing a high level of **security and privacy** will increase the intention to adopt **digital transformation** in Saudi Arabian universities**.***

c) Technical skills in IT departments

Any organization, including Saudi Arabian universities must have technical skills to implement digital transformation efforts successfully. As universities integrate digital technology into their operations, teaching, research, and student services, having a trained IT staff is critical to ensuring these technologies are deployed, maintained, and supported effectively. Well-trained IT staff are critical for successfully implementing digital transformation in universities [75]. This skilled workforce implements and maintains the technologies and security protocols to safeguard digital assets. Building trust and confidence among stakeholders, including students, teachers, and IT staff, is integral. Therefore, according to the hypothesis, universities that have well-trained IT staff are more likely to have a positive intention to adopt digital transformation.

*H12: A high level **technical skills in an IT department** will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

d) Quality of services

The quality of services offered plays a significant role in driving the adoption of digital transformation in universities. High-quality services motivate students, academic staff, and other stakeholders to use digital tools actively. To ensure high-quality services, universities should prioritize users' needs and preferences, making digital tools accessible and reliable. If a low-quality service is provided, some stakeholders may be reluctant to change, but if users have a positive experience with high-quality digital services, they are more likely to overcome reluctance to change and be more open to adopting new technologies [76]. Assuring the quality of services is likely to boost the willingness of public institutions to embrace digital transformation [77]. Therefore, quality of services of digital tools are considered a potentially critical factor that influences the adoption of digital transformation, and this leads to the following hypothesis:

*H13**:** Ensuring the **quality of services** will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

e) Hardware and software compatibility

Hardware and software compatibility is vital for successful digital transformation in universities. Compatibility ensures that diverse technologies can coexist seamlessly, which enables the integration of hardware devices and software applications. Hardware and software compatibility extends to interoperability, allowing hardware and software to communicate and exchange data effectively. Compatibility with multiple operating systems and devices enables users to access digital tools irrespective of their platform preferences. Updating hardware and software to ensure compatibility with the latest technologies and security requirements is essential. High compatibility reduces technical challenges and conflicts between hardware and software, making it easier for stakeholders to adopt and use digital solutions. This, in turn, enables universities to respond swiftly to future technological advancements [78]. Therefore, the hypothesis proposes that universities with high levels of hardware and software compatibility are more likely to have a positive intention to adopt digital transformation.

*H14: A high level of **hardware and software compatibility** within institutions will increase the intention to adopt **digital transformation in** Saudi Arabian universities.*

4) Financial context

a) Cost-effectiveness

Universities would like the benefits of adopting a new invention to be proportionate with the expenses of adopting the innovation. Low-cost inventions are more likely to be used [79]. In the context of electronic data interchange, cost-effectiveness is a crucial variable [80]. Moreover, powerful PCs, reduced hardware and software prices, and ready-to-use, user-friendly software make it more likely for universities to adopt new information technologies, eliminating costs as barriers to universities adopting new information technologies [81]. Many studies have found that online education is cost-effective compared to traditional classroom instruction [82], [83].

Therefore, Universities assess the cost in relation to the advantages before adopting new technologies. Using technology in learning environments is cost-effective once the technology is in place and the organizational culture is amended to embrace these new learning patterns [84]. Therefore, cost-effectiveness is considered a potentially critical factor that influences the adoption of digital transformation; this leads to the following hypothesis:

*H15: **Cost-effectiveness** of digital tools will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

b) Maintenance

In universities, maintenance is a vital part of digital transformation. As universities integrate digital tools and facilities into their operations, teaching, research, and student service, ensuring their correct operation and dependability over time is critical to reap the digital transformation's benefits [85]. Regular maintenance ensures that digital tools and facilities run consistently and optimally, resulting in a great user experience [86]. Proactive maintenance reduces annoyance and resistance to adopting digital technology by minimizing interruptions caused by unanticipated system faults [87]. Proper maintenance helps to preserve the long-term benefits of digital transformation, demonstrating the worth of these technologies. Therefore, a well-maintained digital infrastructure instils trust in users, encouraging them to embrace and effectively employ digital tools. Keeping suitable levels of digital tools and facilities is likely to promote the desire of institutions to implement digital transformation. Proactive maintenance assures digital technology's dependability, performance, and ongoing advantages, creating a pleasant atmosphere that encourages stakeholders to embrace and support the institution's transformation activities. Therefore, the hypothesis proposes that, universities which prioritize appropriate maintenance of digital tools and infrastructures are more likely to have a positive intention to adopt digital transformation.

*H16: **An appropriate level of maintenance** will increase the intention to adopt **digital transformation** in Saudi Arabian universities.*

V. CONCLUSION AND FUTURE WORK

This study has developed a conceptual framework specifically adapted for the exploration of technology adoption in Saudi Arabian universities. The TAM and TOE technology acceptance models have been merged and modified to create the DTA-SAU framework which identifies all the crucial factors that potentially affect adoption of digital transformation in Saudi Arabian universities.

The study takes a mixed method approach to gather data and validate the DTA-SAU framework. The names of all twenty-eight universities were entered to SPSS software. Simple random sampling (SRS) technique was used to draw the sample of Saudi universities. According to SRS method, each university had an equal chance of being selected in the sample. Each university was given a unique ID number which was used in drawing the sample. Seven universities were selected randomly which would represent all Saudi universities. Phase one uses a quantitative methodology and involves distributing a survey to respondents in seven Saudi Arabian universities, targeting various groups of respondents (students, academic staff, IT support departments and top management members). There will be dedicated questionnaires designed specifically for each target group~~.~~ We aim for at least 400 responses to be collected from Saudi Arabian universities to validate the proposed framework and the hypotheses.

Phase two of the study will use a qualitative approach, and semi-structured interviews are conducted. The targeted group for the interviews will be only the top management members in Saudi Arabian universities, including Deans, Deputy Deans, Heads of Department, Deputy Heads of Department and Head of IT Support Department. These interviews are intended to produce qualitative information that would encapsulate the tangible context in which decision-making occurs. For data analysis we will use IBM-SPSS for quantitative method (questionnaires) and NVivo for qualitative method (interviews).

As a result, the DTA-SAU conceptual framework will give decision-makers in universities a better understanding of the challenges and benefits of digital transformation in Saudi Arabian universities and assist them in a more informed decision-making process.

REFERENCES

[1] S. Furusten, "Institutional Theory and Organizational Change". Edward Elgar Publishing, 2nd edition, pp.116-122 , 2023, https://doi.org/10.4337/9781035307210

[2] S.Benn,D. Dunphy & A.Griffiths, "Organizational change for corporate sustainability". (3rd ed.) Routledge, Taylor and Francis Group, (2014) , https://doi.org/10.4324/9781315819181

[3] S. Shariffuddin and J. Razali, "Transformation of university colleges to full-pledged universities: A proposed conceptual framework for Malaysian higher learning institutions," International Journal of Advanced and Applied Sciences, vol. 4, no. 12, pp. 168-173, 2017. DOI:10.21833/ijaas.2017.012.030

[4] Tetco , "Tetco For Educational Technologies",2022, http://tetco.sa/.

[5] B. Schmarzo, "What is digital transformation", (2017), https://www.cio.com/article/3199030/what-is-digital-transformation.html.

[6] V. Maltese,"Digital transformation challenges for universities: ensuring information consistency across digital services". Cataloging & Classification, 56 (7), 1-20, (2018), https://doi.org/10.1080/01639374.2018.1504847

[7] G. Rodríguez-Abitia and G. Bribiesca-Correa,"Assessing digital transformation in universities".*Future Internet*, (13), 2-16. (2021), DOI:10.3390/fi13020052

[8] G.Oliveira and R. Rabechini,"Stakeholder management influence on trust in a project: a quantitative study". *International Journal of Project Management*, 37(1):131-144,(2019), doi: https://doi.org/10.1016/j.ijproman.2018.11.001

[9] S. Misra and D. Wilson, "Thriving innovation ecosystems: synergy among stakeholders, tools, and people" ,2023, https://doi.org/10.48550/arXiv.2307.04263

[10] R. Sharma, and P. Yetton,"The contingent effects of training, technical complexity, and task interdependence on successful information systems implementation," MIS Quarterly, 31(2), 219-238, 2007, https://doi.org/10.2307/25148789

[11] S. Bayona, and S. Iriarte, "IT projects success factors: a literature review," International Journal of Information Systems and Project Management, vol. 8, no. 2, pp. 49-78, 2020, doi: 10.12821/ijispm080203.

[12] S.S.Alam & M.K.M.Noor, " ICT adoption in small and medium enterprises: an empirical evidence of service sectors in Malaysia". International Journal of Business and Management, 4(2), 112,2009, doi:10.5539/ijbm.v4n2p112

[13] S.Bruque & J. Moyano, "Organisational determinants of information technology adoption and implementation in SMEs: the case of family and cooperative firms". Technovation, 27(5), 241–253, 2007,https://doi.org/10.1016/j.technovation.2006.12.003

[14] M.Alshehri & S.Drew,"Challenges of e-government services adoption in Saudi Arabia from an e-ready citizen perspective", ,2010,https://figshare.utas.edu.au/articles/journal_contribution/Challenges_of_e-Government_Services_Adoption_in_Saudi_Arabia_from_an_e-Ready_Citizen_Perspective/22944389/1

[15] L. Dawson, S. Minocha &M. Petre," Exploring the total customer experience in e-commerce environments". In *Proceedings of the IADIS International Conference e-Society* (pp. 945-948),2003, https://www.academia.edu/3050271/Exploring_the_total_customer_experience_in_e_commerce_environments

[16] R.Kapur, "Significance of digital technology". *International Journal of Transformations in Business Management*, 8(II), 20-33,2018, [Online]: https://ijtbm.com/admin/uload/03%20Radhika%20Kapur.pdf

[17] N.Samuel, S.Onasanya,and C.Olumorin,"Perceived usefulness, ease of use and adequacy of use of mobile technologies by Nigerian university lecturers". International Journal of Education and Development using Information and Communication Technology (IJEDICT), 14 (3), 5-16,2018,https://www.learntechlib.org/p/188292/.

[18] T.Gkrimpizi, , V.Peristeras, &I. Magnisalis,"Classification of barriers to digital transformation in higher education institutions: Systematic literature review",Education Sciences, 13(7), 746,2023, https://doi.org/10.3390/educsci13070746

[19] B.R.Aditya , R.Ferdiana, & S.S .Kusumawardani,"Identifying and prioritizing barriers to digital transformation in higher education: a case study in Indonesia". *International Journal of Innovation Science*, *14*(3/4), 445–460,2022, https://doi.org/10.1108/IJIS-11-2020-0262

[20] S.Alateyah, "Identified factors affecting the Intention of Saudi Arabian citizens to adopt e-government services". International Journal of Innovation, Management and Technology 2014. (doi:10.7763/IJIMT.2014.V5.527).

[21] A.Alhubaishy, &A. Aljuhani,"The challenges of instructors' and students' attitudes in digital transformation: A case study of Saudi Universities",Education and Information Technologies, 26(4), 4647-4662,2021, https://doi.org/10.1007/ s10639-021-10491-6

[22] E.King & R. Boyatt. "Exploring factors that influence adoption of e-learning within higher education". British Journal of Educational Technology, 46(6), 1272-1280, (2015), https://doi.org/10.1111/bjet.12195

[23] H.Brdesee, "A divergent view of the impact of digital transformation on academic organizational and spending efficiency: a review and analytical study on a university E-service." Sustainability, 13(13), 7048. 2021, https://doi.org/10.3390/su13137048

[24] A.Tashkandi and I. Al-Jabri, "Cloud computing adoption by higher education institutions in Saudi Arabia: analysis based on TOE", International Conference on Cloud Computing (ICCC). pp. 1–8. IEEE (2015), doi: 10.1007/s10586-015-0490-4.

[25] C.Low, Y. Chen and M.Wu, "Understanding the determinantsof cloud computing adoption," Ind. Manag. data Syst. 2011, doi/10.1108/02635571111161262/full/html

[26] Q.Al Ajmi, , R.A.Arshah, A.Kamaludin, A.SSadiq, M.A.Al-Sharafi," A conceptual model of e-learning based on cloud computing adoption in higher education institutions",International Conference on Electrical and Computing Technologies and Applications (ICECTA). pp. 1–6. IEEE ,2017, doi*:* 10.1109/ICECTA.2017.8252013

[27] H.Gangwar,H. Date, R.Ramaswamy ."Understanding determinants of cloud computing adoption using an integrated TAM-TOE model," J. Enterp. Inf. Manag, 2015, /doi/10.1108/JEIM-08-2013-0065/full/html

[28] A.Chiniah, A.E.U.Mungur and K.N. Permal,"Evaluation of cloud computing adoption using a hybrid TAM/TOE model," In: Information Systems Design and Intelligent Applications. pp. 257–269. Springer ,2019, doi:10.1007/978-981-13-3338-5_24

[29] N. Al Mudawi, N. Beloff and M.White,"Cloud computing in government organizations-towards a new comprehensive model", 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/CBDCom/IOP/SCI). pp. 1473–1479. IEEE ,2019, doi*:* 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00266

[30] A.Alkhalil,R. Sahandi and D. John," An exploration of the determinants for decision to migrate existing resources to cloud computing using an integrated TOE-DOI model." J. Cloud Comput. 6, 1–20 ,2017, doi:10.1186/s13677-016-0072-x

[31] N.Alkhater, G.Wills, R.Walters, "Factors influencing an organisation's intention to adopt cloud computing in Saudi Arabia". ,IEEE 6th international conference on cloud computing technology and science. pp. 1040–1044. IEEE (2014), doi*:* 10.1109/CloudCom.2014.95

[32] I.Arpaci," Antecedents and consequences of cloud computing adoption in education to achieve knowledge management." Comput. Human Behav. 70, 382–390 ,2017, https://doi.org/10.1016/j.chb.2017.01.024

[33] Z.Ali., B.Gongbing and A.Mehreen, "Understanding and predicting academic performance through cloud computing adoption: a perspective of technology acceptance model.", J. Comput. Educ. 5, 297–327 ,2018, Online at: https://mpra.ub.uni-muenchen.de/104638/

[34] J.W.Lian, D.C.Yen and Y.T. Wang, "An exploratory study to understand the critical factors affecting the decision to adopt cloud computing in Taiwan hospital.", International Journal of Information Management, 34, 28–36, 2014, https://doi.org/10.1016/j.ijinfomgt.2013.09.004

[35] F.Davis , "Perceived usefulness, perceived ease of use, and user acceptance of information technology.", MIS Quarterly, 13(3), 319-340 ,1989, doi:10.2307/249008

[36] L. G. Tornatzky, M. Fleischer, and A. K. Chakrabarti, *Processes of technological innovation*. Lexington Books, 1990

[37] F. Alharbi, A. Atkins, and C. Stanier, "Understanding the determinants of cloud computing adoption in Saudi healthcare organisations," Complex & Intelligent Systems, vol. 2, no. 3, pp. 155–171, 2016, doi: 10.1007/s40747-016-0021-9.

[38] S. Lippert and C. Govindarajulu, "Technological, organizational, and environmental antecedents to web services adoption." Communications of the IIMA, 6(1): p.14 ,2006, doi: https://doi.org/10.58729/1941-6687.1303

[39] I.Al-Jabri, and M. Alabdulhadi," Factors affecting cloud computing adoption: perspectives of IT professionals.", International Journal of Business InformationSystems,23(4): p. 389-405, 2016, https://doi.org/10.1504/IJBIS.2016.080215

[40] V.Venkatesh, M.G. Morris, G.B. Davis , & F.D. Davis, "User acceptance of information technology: toward a unified view". MIS Quarterly, 425–478,2003, DOI:10.2307/30036540

[41] S. Y. Park,M. Nam& S. Cha,"University students' behavioral intention to use mobile learning: evaluating the technology acceptance model". British Journal of Educational Technology, 43(4), 592–605,2012, DOI: 10.1111/j.1467-8535.2011.01229.x

[42] L.W.Wasono and A. Furinto,"The effect of digital leadership and innovation management for incumbent telecommunication company in the digital disruptive era.", International Journal of Engineering and Technology, 7, 125–130,2018, doi:10.14419/ijet.v7i2.29.13142

[43] D. Husban,M. Almarshad, and M. Atahrawi, "Digital leadership and organization's performance, the mediating role of innovation capability", International Journal *of Entrepreneurship*, 25(5), 1-18,2021, https://www.abacademies.org/articles/digital-leadership-and-organizations-performance-the-mediating-role-of-innovation-capability.pdf

[44] P. Hoberg, H. Krcmar, G. Oswald,and B.Welz, "Skills for digital transformation" research report, Initiative for digital transformation (IDT) at the Technical University of Munich", Chair for Information Systems (2015). https://www.8pillars.com.au/wp-content/uploads/2018/09/Skills-For-Digital-Transformation-UniOfMunich.pdf

[45] G. Vial,"Understanding digital transformation: a review and a research agenda." *J. Strat. Inf. Syst.* 28 (2) 2021.doi:10.1016/j.jsis.2019.01.003

[46] K.Lenkenhoff,U. Wilkens, M. Zheng,T. Sube,B .Kuhlenkooter and X. Ming,"Key challenges of digital business ecosystem development and how to cope with them.", *Procedia CIRP*, 73, 167-172. , 2018,https://doi.org/10.1016/j.procir.2018.04.082

[47] B.Voigt,K.Manz and U.Wilkens,"What leadership pattern can be Observed in IPS2 work systems when compared with production and service?", *Procedia CIRP*, 16:277–282,2014, https://doi.org/10.1016/j .procir.2014.02.015

[48] L.A. Vasconcelos, A.L.F. Facin, M.S. Salerno and R.K.Ikenami," Unpacking the innovation ecosystem construct: evolution, gaps and trends.", *Technological Forecasting and Social Change*, 136, 30-48,2018, https://doi.org/10.1016/j.techfore.2016.11.009

[49] K. Gilli, M. Nippa and M. Knappstein,"Leadership competencies for digital transformation: an exploratory content analysis of job advertisements.", *German Journal of Human Resource Management*, 37(1), 50-75, 2023, https://doi.org/10.1177/23970022221087252

[50] C. Guzman-Oritz, N. Navarro-Acosta, N. andW. Florez,"Impact of digital transformation on the individual job performance of insurance companies in Peru.", International Journal of Data and Network Science 4, 337-346,2020, doi: 10.5267/j.ijdns.2020.9.005

[51] M. F. Klus and J. Müller,."The digital leader: what one needs to master today's organisational challenges.", *J Bus Econ* **91**, 1189–1223 ,(2021), https://doi.org/10.1007/s11573-021-01040-1

[52] M. Wade and N. Obwegeser,"How to choose the right digital leader for your company.", *MIT Sloan Management Review*, 60(4):1-4,2019, https://www.researchgate.net/publication/333186011_How_to_Choose_the_Right_Digital_Leader_for_Your_Company

[53] M. Sousa and A. Rocha,"Digital learning: developing skills for digital transformation of organizations",2019, https://doi.org/10.1016/j.future.2018.08.048

[54] W. Klemm,"Leadership: creativity and innovation", 1990, online:https://www.researchgate.net/profile/W-R-Klemm/publication/318034968_Leadership_and_Creativity/links/59b6b4acaca2722453a47e13/Leadership-and-Creativity.pdf

[55] J. Goertz,"Creativity: An essential component for effective leadership in today's schools." *Roeper Review*, 22(3), 158-162,2000, https://doi.org/10.1080/02783190009554024

[56] B. Ragu-Nathan,C. Apigian, S. Ragu-Nathan and Q. Tu,"A path analytic study of the effect of top management support for information systems performance," Omega - International Journal of Management Science, 32(6), 459-471, 2004, https://doi.org/10.1016/j.omega.2004.03.001

[57] R. Sabherwal, A.Jeyaraj and C.Chowa,"Information system success: Individual and organizational determinants," Management Science, 52(12), 1849-1864. 12. 2006, https://doi.org/10.1287/mnsc.1060.0583

[58] C. Matt, T.Hess, and A. Benlian,"Digital transformation strategies." *Business & Information Systems Engineering*, 57(5):339-343, 2015,https://doi.org/10.1007/s12599-015-0401-5

[59] P. Regina, and E.De Capitani,"Digital innovation and migrants' integration: notes on EU institutional and legal perspectives and criticalities." *Social Sciences* 11: 144,2022, https://doi.org/10.3390/socsci11040144

[60] Y. Chang,S.F. Wong ,U. Eze and H. Lee,"The effect of IT ambidexterity and cloud computing absorptive capacity on competitive advantage."Industrial Management and Data Systems, 119: 613–638, 2019, doi: https://doi.org/10.1108/IMDS-05-2018-0196.

[61] R. Martins, T.Oliveira and M.A. Thomas," An empirical analysis to assess the determinants of SaaS diffusion in firms." *Computers in human behavior ,2016* ,https://doi.org/10.1016/j.chb.2016.03.049

[62] E. Camargo and M.Y. Wang,"A pilot study on the internationalization of Taiwanese agri-biotech SMEs: A Technology-Organization-Environment (TOE) perspective." *Portland International Conference on Management of Engineering and Technology*, 2015-Septe: 1207–121,2015, doi**:** 10.1109/PICMET.2015.7273220

[63] T. R. Gebba and M.R. Zakaria,"E-government in Egypt: an analysis of practices and challenges." *International Journal of Business Research* (on-line), 4: 11–25, 2015, doi:10.24102/ijbrd.v4i2.576

[64] M.S. Azam, "Services adoption: processes by firms in developing nations." *Advances in Business Marketing and Purchasing* (on-line), 23: 7–290,2015, https://doi.org/10.1108/S1069-096420150000023005

[65] A.Alahmari and L.Kyei-Blankson,"Adopting and implementing an e-learning system for teaching and learning in Saudi public K-12 schools: the benefits, challenges, and concerns." World Journal of Educational [17] Research, 3(1), 11-32,2016, doi:10.22158/wjer.v3n1p11

[66] OECD/European Union,"Digital transformation and capabilities. in supporting entrepreneurship and innovation in higher education in Italy." OECD Publishing,2019, https://doi.org/10.1787/43e88f48-en

[67] S. Sasmoko, L. Mihardjo, F.Alamsjah,and E.Elidjen," Dynamic capability: the effect of digital leadership on fostering innovation

capability based on market orientation" ,Manag. Sci. Lett. 9, 1633–1644,2019, doi: 10.5267/j.msl.2019.5.024

[68] D. C. Brooks and M. McCormack,"Driving digital transformation in higher education "EDUCAUSE., 2020, Online: https://library.educause.edu/resources/2020/6/driving-digitaltransformation-in-higher-education

[69] Z. Yan,"Trust management in mobile environments: autonomic and usable models." 1st edition: IGI Global,2013,doi:10.4018/978-1-4666-4765-7

[70] S. Greener, and C. Wakefield,"Developing confidence in the use of digital tools in teaching.", Electronic Journal of E-Learning, 13(4),260-267.,2015,Online-at https://files.eric.ed.gov/fulltext/EJ1062118.pdf

[71] A. Salovaara,K. Lyytinen and E.Penttinen,"High reliability in digital organizing: mindlessness, the frame problem, and digitaloperations." MIS Q., 43 (2), 555-578,2019, https://doi.org/10.25300/MISQ/2019/14577

[72] M.Connor&M.Siegrist, "Factors influencing people's acceptance of gene technology: the role of knowledge, health expectations, naturalness, and social trust". Science Communication, 32(4), 514–538,2010, https://doi.org/10.1177/1075547009358919

[73] Z.Belkhamza, S.A. Wafa, "The effect of perceived risk on the intention to use ecommerce: the case of Algeria". Journal of Internet Banking and Commerce, 14, 1, 1- 10. (2009), https://www.icommercecentral.com/open-access/the-effect-of-perceived-risk-on-the-intention-to-use-ecommerce-the-case-of-algeria.php?aid=38301&view=mobile

[74] N. Subramanian and A. Jeyaraj,"Recent security challenges in cloud computing." *Computers & Electrical Engineering*, 71, 28-42.2018, https://doi.org/10.1016/j.compeleceng.2018.06.006

[75] A.R. Alharbi, "The current state of teachers' ICT use in classrooms in boys' secondary schools in Saudi Arabia". PhD thesis, University of Glasgow, UK. (2019a). https://theses.gla.ac.uk/73018/1/2018AlharbiPhD.pdf

[76] N. C. Fernandez and P. Lago," Characterizing the contribution of quality requirements to software sustainability." *Journal of System Software*,137,289-305,2018, https://doi.org/10.1016/j.jss.2017.12.005

[77] Y.Wu,H. Li, R. Luo, and Y. Yu,"How digital transformation helps enterprises achieve high-quality development? Empirical evidence from Chinese listed companies." *European Journal of Innovation Management,* ISSN: 1460-1060, 2023, https://doi.org/10.1108/EJIM-11-2022-0610

[78] G.Simmons , G.A. Armstrong , M.R.Durkin ,"A conceptualization of the determinants of small business website adoption."International Small Business Journal 26(3): 351–389. 2018, https://doi.org/10.1177/0266242608088743

[79] L. G. Tornatzky and K. J. Klein, "Innovation characteristics and innovation adoption-implementation: a meta-analysis of findings," in *IEEE Transactions on Engineering Management*, vol. EM-29, no. 1, pp. 28-45, Feb. 1982, doi: 10.1109/TEM.1982.6447463.

[80] G. Premkumar , K.Ramamurthy , S. Nilakanta ,"Implementation of electronic data interchange: an innovation diffusion perspective." Journal of Management Information Systems 1994;11(2):157±86, 1994,https://doi.org/10.1080/07421222.1994.11518044

[81] P.Palvia , D.B.Means , W.M.Jackson ,"Determinants of computing in very small businesses.",Information and Management 27:161±74,1994, https://doi.org/10.1016/0378-7206(94)90044-2

[82] S.Bartolic-Zlomislic,& C.Brett, "Assessing the costs and benefits of telelearning: a case study from the Ontario Institute for Studies in Education of the University of Toronto". (1999)., [Online]. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4a861ceaacaf8aea7c223d6c96f01168b83c3fcf

[83] J. Cukier, "Cost-benefit analysis of telelearning: developing a methodology framework". distance education, 18(1), 137−152. (1997). DOI:10.1080/0158791970180110

[84] C.Schooley,"The ROI of elearning,2009, https://sussex.idm.oclc.org/login?url=https://www.proquest.com/magazines/roi-e-learning/docview/197265223/se-2?accountid=14182.

[85] A,Saihi ,M. Awad and M.Ben-Daya, "Quality 4.0: leveraging industry 4.0 technologies to improve quality management practices – a systematic review", *International Journal of Quality & Reliability Management*, Vol. 40 No. 2, pp. 628-650, 2023, https://doi.org/10.1108/IJQRM-09-2021-0305

[86] J.Bokrantz, A. Skoogh, C.Berlin, and J. Stahre,"Maintenance in digitalised manufacturing: Delphi-based scenarios for 2030." *International Journal of Production Economics*, 91, 154-169,2017, https://doi.org/10.1016/j.ijpe.2017.06.010

[87] L.Silvestri, A. Forcina,V. Introna ,A.Santolamazza,and V.Cesarotti, "Maintenance transformation through industry 4.0 technologies: A systematic literature review." *Computers in Industry,* 19(3), 27-36, 2020, https://doi.org/10.1016/j.compind.2020.103335

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 29–36
DOI: 10.15439/2024F7400
ISSN 2300-5963 ACSIS, Vol. 41

Enhancing Text Recognition of Damaged Documents through Synergistic OCR and Large Language Models

Thomas Asselborn*, Jens Dörpinghaus‡§, Faraz Kausar†, Ralf Möller*, Sylvia Melzer†
* Universität Hamburg, Institute for Humanities-Centered AI, Warburgstraße 28, 20354 Hamburg, Germany,
Email: {thomas.asselborn, ralf.moeller}@uni-hamburg.de
, https://orcid.org/0009-0005-3011-7626, https://orcid.org/0000-0002-1174-3323
† Universität Hamburg, Centre for the Study of Manuscript Cultures,
Warburgstraße 26, 20354 Hamburg, Germany,
Email: faraz.kausar@studium.uni-hamburg.de, sylvia.melzer@uni-hamburg.de, https://orcid.org/0000-0002-0144-5429
‡ Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
§ University of Koblenz, Germany,
Email: jens.doerpinghaus@bibb.de, https://orcid.org/0000-0003-0245-7752

***Abstract*—Optical Character Recognition (OCR) remains a highly relevant area of research in pattern recognition. Its applications span various domains, including supporting reading for the visually impaired, interpreting Morse codes, capturing postal addresses, evaluating emails, scanning price tags and passports, and extracting text from digitised documents. As the volume of digitised data continues to grow, challenges arise in capturing the semantic structure of documents through logical structure analysis and providing data suitable for information retrieval to answer specific research questions. While classic OCR processes like Tesseract and OCRopus work well for contemporary digitised documents, there is room for improvement in text and word recognition of historical documents that are severely damaged. Large Language Models (LLMs) like GPT-4 can be effectively used for text recognition tasks, utilising their advanced natural language processing capabilities to interpret and reconstruct unclear or damaged text, offering potential for improving the overall text recognition process. However, challenges arise additionally when documents contain e. g. a mixture of single-column and double-column text, images and text, or words not known or blocked by the agents.**

This article aims to find a suitable combination of OCR models and LLMs to accurately add missing words to texts according to their original versions.

I. INTRODUCTION

EVEN after 90 years, OCR (**O**ptical **C**haracter **R**ecognition) is still a very topical area of research in the field of pattern recognition. The areas of application for this technology are very wide-ranging. These include supporting reading for the blind, interpreting Morse codes, automatically capturing postal addresses, evaluating e-mails, scanning price tags and passports and extracting text from digitised documents [4]. The number of digitised data is steadily increasing and various repositories [17], [12], [32] with different search strategies have already been created. With regard to the data analysis of digitised data, there is an increasing challenge to capture the semantic structure of a document through logical structure analysis and to provide data that is suitable for retrieving information to answer individual research questions and consists of more than just a comparison of character strings. In addition, to promote document understanding, relevant information is required that not only results from text extraction, but also uses other data from images, notes, drawings, fonts, font colours, locations, document structures, etc.

The classic OCR processes include the open source OCR Tesseract or OCRopus. These methods have been extended with regards to the possibilities of machine learning to meet the above-mentioned challenges among others. One extension, for example, are the OCR-D software modules [23]. While the text recognition of these processes works very well for digitised documents from the present time, there is still room for improvement in the text and word recognition of historical documents that are severely damaged, e.g. by water damage, glued-on notes, perforations, holes, mould, etc. which can lead to words that are no longer recognisable. Some documents may be inscribed or obviously crossed out in later years, particularly in passages reflecting past ideology.

LLM (**L**arge **L**anguage **M**odel)s such as GPT-3 can be used effectively for text recognition tasks. Their advanced natural language processing capabilities allow them to interpret and reconstruct unclear text in addition to filling in spaces with damaged or missing text. These capabilities offer potential to improve the text recognition process. A simple implementation of first recognising text and then using an agent such as

---

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 ‘Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures’, project no. 390893796.

©2024, PTI

ChatGPT[1], Perplexity.ai[2] or UHHGPT[3] to add words becomes a challenge when documents contain a mixture of single-column and double-column text or images and text, or the words used are not part of the LLM (either not known or blocked by the agent).

Some OCR approaches were tested and it became apparent that text recognition of tables poses a major challenge. Only by correctly identifying the image and text regions that belong together can the text be generated in the correct order or the use of agents make sense. Prompt engineering also plays an important role in ensuring that an existing text is produced as it might be written, rather than a new text. The agent must first assume the role of the writer from the relevant time so that the desired correct text is produced. The challenge we face in this article is to find a suitable combination of OCR models or tools and an LLM so that the missing words from texts are added according to the original version of the text.

We have tested our approach on the digitised journals of the Godeffroy Natural History Museum[4], which existed in Hamburg from 1861 to 1885 and on a German legal document corpus. The Federal Institute for Vocational Education and Training maintains a collection of occupation-related documents with legal bases, which reflect about 85 years of German VET (**V**ocational **E**ducation and **T**raining) history. In recent years, this collection has been systematically recorded for the first time, resulting in precise knowledge of its contents on the one hand and the state of preservation of the individual documents on the other.

The results obtained provide an overview of various combined approaches and show that even with poor OCR results, the use of LLMs still delivers good results overall.

II. RELATED WORK

Vocational Education and Training

Understanding the practicalities of reform implementation is crucial for effective VET policy transfer and adaptation.

A lot of different research has been done in the digitisation of documents in recent years: For example, historical Finnish newspapers, see [17], or historical publications of the Bundesanzeiger, see [12]. The historical development of vocational training regulations has only been studied to a very limited extent [15], while the general history and development of the labour market in relation to occupations receives much attention, see [37], [13], [29], [21]. Other works focus on the current development of regulations, see [18], and their analysis is also widely considered [10], [26], [2]. However, it remains unclear whether this is due to the fact that historical resources are currently not publicly available.

The Historical international standard classification of occupations (HISCO) is a publicly available dataset of comparable occupations that would be a prerequisite to make historical occupations and regulations interoperable. It was introduced in 2002 [19] and is available as a database at https://historyofwork.iisg.nl/index.php, where several datasets can be downloaded. However, the list of German occupations is incomplete. Another relevant dataset is prepared as Ontologie historischer, deutschsprachiger Berufs- und Amtsbezeichnungen (see https://www.geschichte.uni-halle.de/struktur/hist-data/ontologie/), but is currently not publicly available. Classifications for GDR (**G**erman **D**emocratic and **R**epublic) occupations are also not yet digitally available, while their mapping to standards like KldB is widely discussed [9], [1]. Another dataset is offered as “Genealogie der Berufe”, but is only available as a web service (see https://www.bibb.de/dienst/berufesuche/de/index_berufesuche.php/). Also worth mentioning is the seminal work by Wolf-Dieter Gewande, who in 1999 for the first time compiled unpublished recognition data and traced the development of more than 1300 occupations to the present, see [11].

While very little research has been done on the historical regulation of vocational education and training in Germany, we can identify a second research gap: Data integration should be accompanied by linked data sets for occupational classifications that are not currently available. Thus, the integration of older data such as the KldB 1975, 1988 and 1992 is crucial.

Text recognition

To support as many application areas as possible with OCR, individual OCR modules were developed in various research projects, including the OCR-D project [23]. During the OCR-D project, the OCR-D software was developed to allow the easy combination of a variety of so-called processors – independent tools for specific tasks – to define workflows tailored to the peculiarities of different templates and thus automate the process of a large quantity of prints, particularly from the 16th to 18th centuries (cf. [33], [34], [35]. It is only through such automation that it is possible to make large collections available in their entirety as full-text to the scientific community.

OCR4all is a web application that offers a semi-automatic workflow tailored to digitise historical documents [30]. However, when compared to Tesseract, OCR4all’s performance is not as strong. While OCRopus is an open source software, it also falls short of Tesseract’s performance. ABBYY FineReader, a commercial tool, typically provides only slightly better results than Tesseract, see [14], [5].

LLMs for Text Recognition

LLMs such as GPT-3 and GPT-4 have shown remarkable capabilities in understanding and generating human-like text. Although they are primarily used for NLP (**N**atural **L**anguage **P**rocessing), recent research has investigated the use of LLMs to improve text recognition. [31], [36]

To the best of our knowledge none of the literature we found specifically discussing the synergies between LLMs and OCR techniques. This therefore appears to be a relatively new area of research with little direct literature available to date. The

---

[1] https://chat.openai.com

[2] https://www.perplexity.ai/

[3] https://uhhgpt.uni-hamburg.de/

[4] https://www.biodiversitylibrary.org/item/244246

sources indicate some related work on the use of LLMs for text recognition tasks, but a comprehensive overview of their synergies with OCR is not readily available based on these search results.

GPT4-o has demonstrated proficiency in multilingual applications. For German text input specifically, GPT4-o emerges as the preferred choice. Its superior performance in handling multilingual tasks, including those involving German, makes it particularly well-suited to process German text. [20]

III. USE CASES

Document collections constitute a vital component of vocational training research. VET and CVET (**C**ontinuing **V**ocational **E**ducation and **T**raining) in Germany have been subject to regulation since the 1920s. Over the course of many decades and through various political regimes, including the Third Reich, the German Democratic Republic, and the Federal Republic of Germany, these regulations have undergone significant evolution. The job archive at the Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung, BIBB) houses numerous historical VET and CVET regulations, which are largely inaccessible.

The key stakeholders in continuing vocational training in Germany include: (a) educational institutions, (b) companies and enterprises, (c) employees, and (d) sponsors. In light of the aforementioned transformation processes, it is imperative that these stakeholders adapt to the evolving conditions and requirements that they face. These challenges have shifted over time, necessitating a comprehensive overview and analysis of how regulatory documents reflect the aforementioned evolving requirements, changes, and challenges. For instance, it is crucial to identify which educational content is increasingly offered and demanded, in order to draw conclusions about the development needs of both the vocational and continuing education systems. The research-based development of the vocational education system aims not only to ensure the economy's competitiveness at a systemic level but also to combat unemployment and stabilise the social security system [8]. However, the historical regulations are not currently available in a digitised format. Given that these documents span a long period and multiple states (the German Empire, the GDR, and the FRG (**F**ederal **R**epublic of **G**ermany)), the challenges for OCR and data infrastructure are substantial.

The German Committee for Technical Education (DATSCH) was established in 1908, and from that point forward, a series of documents were created with the objective of standardising occupations. The age of the documents has resulted in the deterioration of the paper, with approximately two-thirds written in Fraktur script and the remainder in various Latin fonts. Preservation varies, with some documents exhibiting well-preserved characteristics and others displaying signs of water damage, glued notes, perforations, or mould. Some documents have later inscriptions or are crossed out, particularly passages reflecting Nazi ideology.

The majority of early order specifications are in DIN A5 format, with special formats ranging from pocket-sized job descriptions (DIN A6) to large inserts up to about DIN A1. A smaller part of the collection includes legal regulations from the Federal Republic of Germany, subject to BBiG, HwO, specific health profession laws, or federal school legislation. This collection encompasses training regulations, amendments, corrections, framework curricula, and advanced training regulations from the federal government, federal states, and competent bodies.

The GDR materials pertain to training and advanced education, including training documents for skilled worker training and socialist vocational training (Ausbildungsunterlagen für die Facharbeiterausbildung and Ausbildungsunterlagen für die sozialistische Berufsbildung), training plans (Ausbildungspläne), and equipment norms. The GDR documents are frequently bound as booklets or books, comprising up to 323 pages in A4 and A5 formats. They are typically printed in two columns with a typewriter font.

The diverse formats and conditions of the documents present significant OCR challenges. The use of Fraktur and multiple Latin scripts, along with physical damages complicates digitisation. Moreover, the prevalence of special formats, two-column layouts, and typewriter fonts in GDR materials serves to compound the difficulty in creating accurate digital copies. Consequently, novel approaches to support OCR on these very specific documents are of great importance for further research on (vocational) education.

At the Centre for the Study of Manuscript Cultures (CSMC) at the Universität Hamburg, there are a number of application areas in which OCR, like OCR-D, is used in a wide variety of workflows which can then support the evaluation of historical prints and documents. In the special research project "Sonderforschungsbereich" (SFB) 950: Manuscript Cultures in Asia, Africa and Europe, the empirical diversity of manuscript cultures was researched on the basis of the material. This resulted in numerous digital copies. Further digital copies have been and are being produced by the current DFG Cluster of Excellence 2176 - Understanding Written Artefacts (UWA). Automated character and word recognition of the digitised material using OCR can support research activities in the evaluation of historical prints and manuscripts. Historical manuscripts are often not in best condition and may include damages in a similar way described previously; potentially it may be even more dramatic. The application would not only save time, but it would also be conceivable to establish separate software modules to be included automatically after the standard OCR process to create a new, improved application so that damaged texts can also be restored.

IV. OPTICAL CHARACTER RECOGNITION (OCR)

OCR is the process of recognising text in a scanned or photographed document, image-only PDF containing text and similar types of documents. The goal of this process is to convert this text into a machine readable format, e.g. as a plain text file, so that further processing can be performed. [16] This process contains multiple discrete steps that are sometimes also called activities.

A. Activities

OCR modules are available and described in more detail in [24]. Some activities are presented in the following:

- *Binarization* is the process of converting an image into a binary representation, where each pixel is either black or white.
- *Dewarping* is used to correct distortions in scanned documents caused by the curvature of the page or the scanning process itself.
- *Despeckling* aims to remove small, isolated spots or noise from scanned images, improving their overall quality.
- *Deskewing* is the process of straightening a skewed or rotated document image, aligning it properly for further processing.
- *Font identification* analyses the shapes and characteristics of characters in a document to determine the font or fonts used.
- *Segmentation* is the process of dividing an image into meaningful regions or components for further analysis.
- *Region segmentation* identifies and separates different regions within a document, such as text blocks, images, and tables.
- *Region classification* categories the segmented regions of a document into different types, such as text, graphics, or tables.
- *Line segmentation* breaks down text regions into individual lines, enabling line-by-line processing.
- *Line recognition* analyses and interprets the content of each line, extracting relevant information.
- *OCR* is the process of converting scanned texts images into machine-readable text data.
- *Text recognition* analyses and interprets the segmented text, converting it into machine-readable format.

The existing OCR-D modules can be easily combined in an individual workflow, which ensures very good adaptability of the OCR-D modules, at least for the evaluation of digital prints.

B. OCR-D Workflow Application

As mentioned briefly above, the use of the OCR-D software modules and workflows for the full-text digitisation of the "Journal of the Godeffroy Museum". This collection was used because the results can be published without any problems in terms of copyrights, etc. and have the same representative requirements for text recognition of table contents.

Fig. 1 shows the result after running the Tesseract OCR module for a table. Green indicates which text characters were recognised correctly and red indicates which text characters were recognised incorrectly. We used the workflow:

1) Binarization
2) Region segmentation
3) Line segmentation
4) Line recognition

The workflow with just a few steps shows that an improvement in region recognition is necessary. Extending the workflow with additional steps usually leads to better results, but unfortunately this does not apply in the case of tables. A subsequent application of agents could lead to no meaningful texts, since row-by-row evaluation is the wrong reading direction for tables. Thus, for the recognition of texts, we have used other OCR tools.

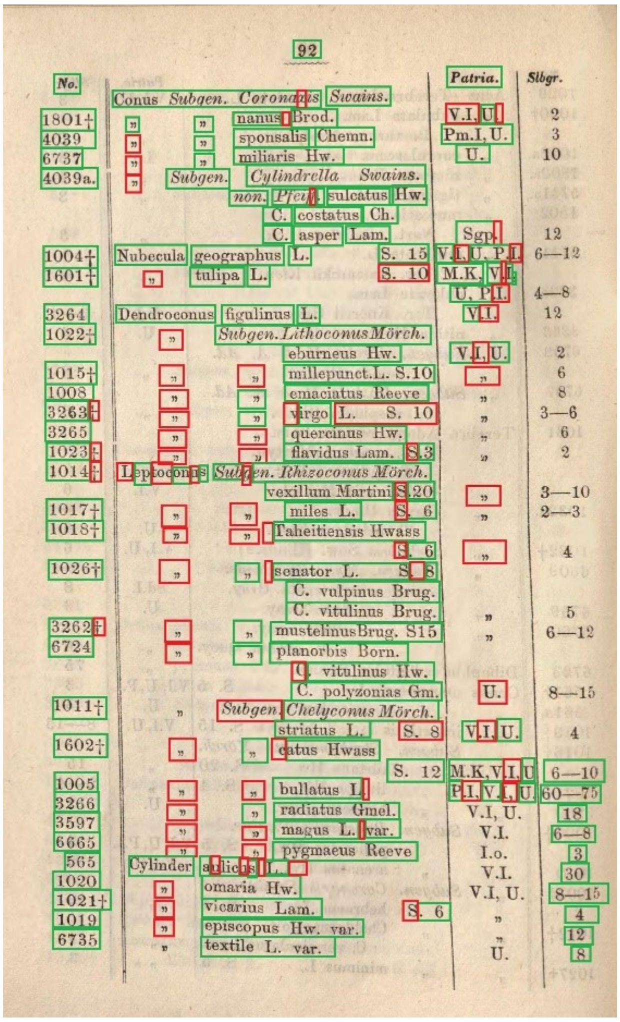
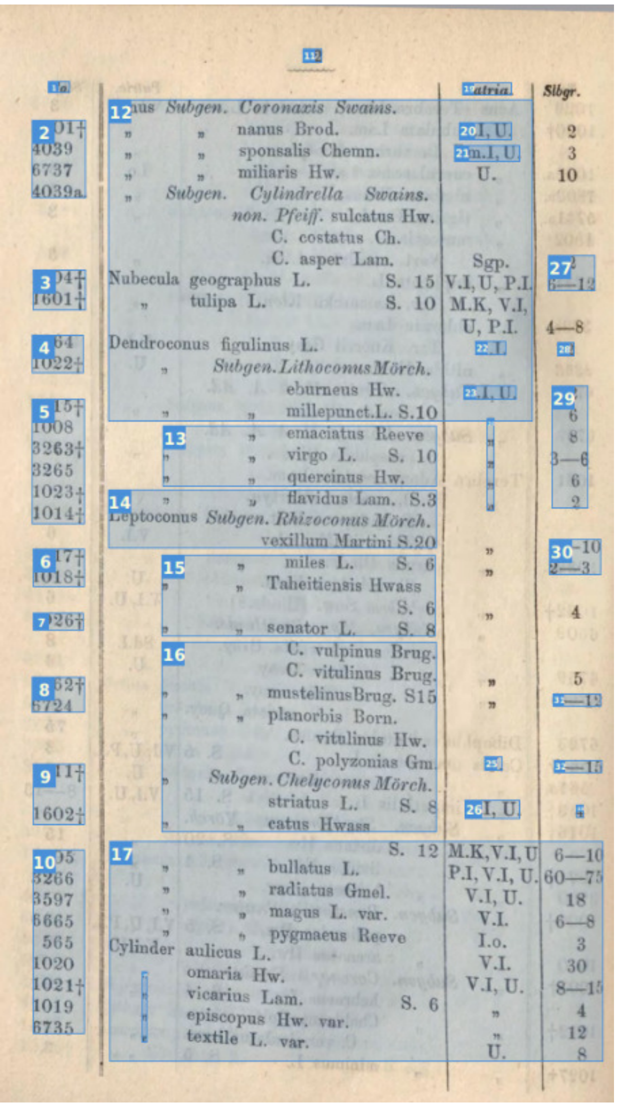


Fig. 1. Table from https://www.biodiversitylibrary.org/item/ 244246 page 92 with the OCR-D results

C. gImageReader

We also tested the OCR software gImageReader, which offers the possibility of correction by a user via a user interface. It is a GUI (**G**raphical **U**ser **I**nterface) front-end for Tesseract [22]. Thus, it provides the option for people that have no skills in using the command line to use Tesseract. Fig. 2 shows the areas marked by the software. The user has the option of adding further areas manually or deleting other areas. Here, the contiguous text areas of the table were not recognised convincingly well.

Using gImageReader on the example [3] from the VET corpus shown in Fig. 3, we got the following result:



112

atria Slbgr.
us *Subgen. Coronaxis Swains.*
" " nanus Brod. I, U. 2
" " sponsalis Chemn. m.I, U. 3
" " miliaris Hw. U. 10
" *Subgen. Cylindrella Swains.*
non. Pfeiff. sulcatus Hw.
C. costatus Ch.
C. asper Lam. Sgp.
Nubecula geographus L. S. 15 V.I, U, P.I.
" tulipa L. S. 10 M.K, V.I, U, P.I. 4—8
Dendroconus figulinus L. I.
" *Subgen. Lithoconus Mörch.*
eburneus Hw. I, U.
" " millepunct. L. S.10 6
" emaciatus Reeve 8
" " virgo L. S. 10 3—6
" " quercinus Hw. 6
" " flavidus Lam. S.3 2
Leptoconus *Subgen. Rhizoconus Mörch.*
vexillum Martini S.20 "
" miles L. S. 6 "
" " Taheitiensis Hwass S. 6 " 4
" " senator L. S. 8
C. vulpinus Brug.
C. vitulinus Brug. " 5
" " mustelinus Brug. S15 "
" " planorbis Born.
C. vitulinus Hw.
C. polyzonias Gm.
" *Subgen. Chelyconus Mörch.*
striatus L. S. 8 I, U.
" " catus Hwass S. 12 M.K, V.I, U 6—10
" " bullatus L. P.I, V.I, U. 60—75
" " radiatus Gmel. V.I, U. 18
" " magus L. var. V.I. 6—8
" " pygmaeus Reeve I.o. 3
Cylinder aulicus L. V.I. 30
omaria Hw. V.I, U. 8—15
vicarius Lam. S. 6 " 4
episcopus Hw. var. " 12
textile L. var. U. 8

Fig. 2. The text areas marked by the OCR software gImageReader after performing the segmentation

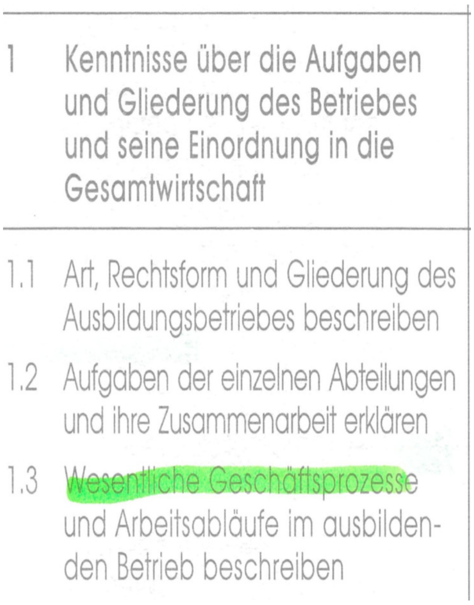
```
1 Kenntnisse Uber die Aufgaben
inhalte gem. und Gliederung des Betriebes
Berufsbild und seine Einordnung in die
Gesamtwirtschaft

1.1 Art, Rechtsform und Gliederung des
Kenntnisse und Ausbildungsbetriebes
    beschreiben

-|1.9 Aufgaben der einzelnen Abteilungen

und ihre Zusammenarbeit erkldren

und Arbeitsabldufe im ausbilden-
den Betrieb beschreiben
```



1 Kenntnisse über die Aufgaben
und Gliederung des Betriebes
und seine Einordnung in die
Gesamtwirtschaft

1.1 Art, Rechtsform und Gliederung des
Ausbildungsbetriebes beschreiben

1.2 Aufgaben der einzelnen Abteilungen
und ihre Zusammenarbeit erklären

1.3 Wesentliche Geschäftsprozesse
und Arbeitsabläufe im ausbilden-
den Betrieb beschreiben

Fig. 3. Excerpt from “AKTUALISIERTE AUSBILDUNGSPLANEMPFEHLUNGEN Datenverarbeitungskaufmann Datenverarbeitungskauffrau” including human annotation [3]

D. *OCR4all*

We used OCR4all with the standard workflow as described in [25] to recognise the same text example shown in Fig. 3. Again, the results provided were not sufficient:

```
Kenntnisse über die Aufgoben
und Gliederung des Betriebes
und seine Einordnung in die
Gesomtwirtschoft
Ausbildungs-
holbjohre
afele
P
-
```

E. *ABBYY FineReader*

Finally, we used ABBYY FineReader 15 OCR-Editor to recognise the texts as well. Here, the result was better than what we got from all other OCR engines used in our sample testing.

The ABBYY FineReader output is:

```
1 Kenntnisse über die Aufgaben und Gliederung
    des Betriebes und seine Einordnung in die
    Gesamtwirtschaft
1.1 Art Rechtsform und Gliederung des
    Ausbildungsbetriebes beschreiben
1.2 Aufgaben der einzelnen Abteilungen und
    ihre Zusammenarbeit erklären
1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37\Zz0Gy
    | l-G-M |ö l övG
und Arbeitsabläufe im ausbildenden Betrieb
    beschreiben
```

Text recognition of tables works quite well with this tool. The texts that were marked in the original were not recognised. A jumble of letters was output. No meaningful search can be activated for this bad case. This can be remedied to a certain extend by a LLM, as shown next.

V. COMBINATION OF OCR AND LLMS

ChatGPT is an advanced language model developed by OpenAI, designed to generate human-like text and engage in natural and coherent conversations with users. It is built upon the GPT (**G**enerative **P**re-trained **T**ransformer)-3.5 architecture. It can maintain context in conversations and respond in a contextually relevant manner, making it suitable for a wide range of applications. There are other agents, for example UHHGPT, that have been developed and use some of the same functions such as GPT-3.5 or GPT-4 omni. UHHGPT is a graphical user interface based on HAWK-Ki[5] that utilises the OpenAI API to send queries to and receive answers from one of the OpenAI GPT models. Its main goal is to provide anonymous access to the models since the user´s query cannot be traced back to the individual member of the university. Additionally, it offers access to GPT-4 omni for free for all members of the university. The prompt sent to OpenAI is always augmented by "You are a helpful assistant who works at the University of Hamburg in Germany." We have used UHHGPT with the function GPT-4 omni for our experiments. The following outputs show the same text excerpts from Section IV.

Prompt:

```
The following text was recognized after an
     OCR. Complete the words as they could
    have been written: [OCR output above is
     part of this prompt].
```

The output of UHHGPT with the function GPT-4 omni (OCR4all text) is:

```
Kenntnisse über die Aufgaben
und Gliederung des Betriebes
und seine Einordnung in die
Gesamtwirtschaft
Ausbildungsjahre
Anleitung
Praktikum
```

The output of UHHGPT with the function GPT-4 omni (gImageReader) is:

```
1. Kenntnisse über die Aufgaben
inhalte gemäß und Gliederung des Betriebes
    ,
Berufsbild und seine Einordnung in die
Gesamtwirtschaft.

1.1 Art, Rechtsform und Gliederung des
Kenntnisse- und Ausbildungsbetriebes
    beschreiben.

1.9 Aufgaben der einzelnen Abteilungen
und ihre Zusammenarbeit erklären
und Arbeitsabläufe im ausbildenden
Betrieb beschreiben.
```

The output of UHHGPT with the function GPT-4 omni (ABBYY FineReader text) is:

```
1 Kenntnisse über die Aufgaben und
    Gliederung des Betriebes und seine
    Einordnung in die Gesamtwirtschaft
1.1 Art, Rechtsform und Gliederung des
    Ausbildungsbetriebes beschreiben
1.2 Aufgaben der einzelnen Abteilungen und
     ihre Zusammenarbeit erklären
1.3 Die Betriebsvorgänge und Arbeitsablä
    ufe im ausbildenden Betrieb beschreiben
```

VI. RESULTS

We considered the best combination for our data set (ABBYY FineReader with GPT-4 omni) for further evaluation. Extensive evaluations would have to be carried out before a generally valid statement could be made about the best combinations. A few of our key findings are the following.

Bad input leads to bad output

One of the findings we had was that a erroneous detection with the standard OCR process will also lead to corrections by UHHGPT that are likely erroneous as well. This becomes clearer when looking at one example. After only using ABBYY FineReader, we got the following detection:

1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37 \Zz0Gy | l-G-M |ö l ö vG und Arbeitsabläufe im ausbildenden Betrieb beschreiben

(translated into English: 1 id V'V: vJtQ tz 1 rll r G 1 1 kZz \37 \Zz0Gy | l-G-M |ö l ö vG and describe work processes in the training company). The first part could mean anything but based on the semantics, UHHGPT provided the word "Aufbau" (structure) as an appropriate correction. While this word may be fitting, the correct words used where "Wesentliche Geschäftsprozesse" (key business processes) which is semantically not identical to "Aufbau".

A similar problem may also occur when looking at tables. When the regions of the table were not identified correctly, as seen in Section IV-B, and the results are provided as if the table columns were all part of the same sentence, UHHGPT will nevertheless try to find a sentence that may be fitting combining elements of originally different sentences.

---

[5] https://github.com/HAWK-Digital-Environments/HAWKI

The larger the relevant context provided, the better the results

When looking at the results we got after correction with UHHGPT, we noticed that results are generally better when the context provided to it is longer. One example for that was present in the text where a cut-out is shown in Fig. 3. Sometimes, OCR wrongly recognised the chapter number written before every entry, e.g. instead of 1.3 we got 1. When trying to correct that error using UHHGPT, it was able to correct that to 1.3 if the complete table of contents but it did not correct it if only this specific line was given. There are a few factors that may impacts how much context a GPT can take into consideration.

First, GPT has, depending on the specific model, a specific context window. This context window measures how many tokens before the currently generated token are taken into consideration during generation [6]. GPT 4 omni has a context window of 128,000 tokens while GPT 3.5 Turbo only has one of 16,385 tokens [27]. Using the estimates from [28], [38], GPT-3.5 Turbo has a context window of roughly 12,288 words or 24.5 pages while GPT-4 omni has a context window of roughly 96,000 words or around 192 pages. While there are a few documents in our corpus that have more pages than the context window of GPT-3.5 Turbo, GPT-4 omni´s context window should suffice for most.

A second factor to look at is that it may only be useful to provide *relevant* context to the GPT. In one example document about the training of IT specialists, there is a handwritten remark. ABBYY FineReader was unable to detect that correctly and gave "U&ihn s.z". Using UHHGPT, we got "Um UNIX" as a result which may be a correct detection in the context of IT but it was not what was originally written on the page ("vorher 9.2"). Thus, a middle ground between providing enough context and only providing surely relevant context, which may not suffice, must be found.

UHHGPT tends to correct "mistakes" that were no mistakes

In 1998, the German orthography was subject to significant changes. [7] One example is the usage of the letter "ß" (sharp s). Prior to the reform, it was additionally used as a last letter instead of "ss", e.g. today, the spelling "dass" is used but before 1998 it was "daß" (German word for "that"). Other changes included a change from "ph" to "f", e.g. "Photographie" became "Fotografie". There were also a few other changes that are not discussed here.

Our corpus primarily consists of documents written before the reform. Thus, the old spelling was used instead of the new one after 1998. UHHGPT does however correct the old orthography to the new one which does not change the semantics of the text but it may be problematic if one is interested in a one to one digitisation of the original text.

In a similar way, words may be written in all capital letters on purpose, e.g. as a title, or hyphenation was used in a specific way. UHHGPT will "correct" the words to not be written in all capital letters as well as removing all hyphenation. Depending on the specific task, this may be a problem in later steps.

UHHGPT does not always return what it is asked for

When asking UHHGPT to correct the errors using the prompt shown in Section V, it most of the time did what it is asked for. However, in some instances, UHHGPT answered not only with the correction but also an explanation for what is written in the text. Additionally, it sometimes also provided a translation from German into English.

Some quantitative measures

In addition to the qualitative results mentioned above, we got a few quantitative measures. We have taken a few sample pages from the document describing the vocational training for a "Datenverarbeitungskaufmann" from 1995 (see Fig. 3) [3]. Some results are shown in Table I.

TABLE I
SOME RESULTS FOR THE OCR DETECTION AS WELL AS CORRECTIONS BY UHHGPT (GPT-4 OMNI)

Page no.	No. of errors after OCR	No. of errors corrected	No. of errors added
5	5	5	2
7	75	10	28
17	21	5	20
23	6	4	3
28	2	1	3

While GPT-4 omni was able to correct some OCR errors, it also introduced new errors on several pages. For some examples, UHHGPT added more new errors than it has corrected. When using a combination of OCR and LLMs, it should be considered that the results can also contain incorrect corrections when evaluating the data. It is up to the user doing further research with the results to check whether the new errors are relevant or irrelevant for their specific tasks. We analysed some other documents, too, and came to the overall conclusion that the use of GPT-4 omni has made two-thirds of the corrections of the OCR errors.

VII. CONCLUSION AND OUTLOOK

This paper focuses on a corpus of documents pertaining to vocational education research. These regulations were established in the 1920s and have undergone significant evolution over the course of many decades and through various political regimes, including the Third Reich, the GDR, and the FRG. In order to make these documents digitally available for VET research and data science methods, it is crucial to apply OCR methods and digitised (scanned) documents. However, there are different use cases. For the digital archive (Berufearchiv), it is important to have a one-to-one version of the original documents and also provide digital representatives, e.g., in TEI-XML. In the context of data science methods, it is of greater importance to consider the correct language equivalent, given that, for instance, NLP methods do not typically rely on a specific spelling revision but are influenced by poor OCR quality.

In conclusion, for the first use case, it is evident that a human intervention and revision are still necessary. Consequently, future research should focus on collaborative software or further improvements to OCR. In the second use case, our approaches significantly enhance the quality of the documents, rendering them suitable for NER. However, several critical issues and shortcomings remain. For instance, handwritten artefacts or other alterations to the text are replaced with fictitious texts. With the advancement of OCR, there is a growing need to overcome the challenges presented by historical documents with severe damage. By combining traditional OCR methods with LLMs such as GPT-4 omni, new possibilities for accurately reconstructing and recognising text in damaged documents are emerging. This approach shows promising results in improving text recognition accuracy and preserving the original context of the documents, leading to advances in the preservation and analysis of historical texts.

In our experiments, we have identified the optimal combination for our dataset as ABBYY FineReader with GPT-4 omni (in the form of UHHGPT), which we will further investigate for its effectiveness in the future, in addition, considering other LLMs. As a medium to long-term goal, we can imagine that the approach developed here will be integrated into a RDR (**R**esearch **D**ata **R**epository) in such a way that the combined OCR with LLMs will be offered for the texts in the repository.

REFERENCES

[1] Klassifikation der Berufe, K.: Band 1: Systematischer und alphabetischer Teil mit Erläuterungen (2010)

[2] Bliem, W., Petanovitsch, A., Schmid, K.: Success factors for the Dual VET System. Update (2015)

[3] Bojanowsky, A., Bross, D., Feuerstein, A., Häußler, J., Linde, F., Plattmann, U., Schenk, G., Tumfart, D.: Aktualisierte Ausbildungsplanempfehlungen Datenverarbeitungskaufmann Datenverarbeitungskauffrau. Kuratorium der Deutschen Wirtschaft für Berufsbildung, Adenauerallee 8a, 53113 Bonn (1995)

[4] Bunke, H., Wang, P.S.P. (eds.): Handbook of Character Recognition and Document Image Analysis. World Scientific, Singapore (May 1997). https://doi.org/10.1142/2757

[5] Clausner, C., Antonacopoulos, A., Pletschacher, S.: Efficient and effective OCR engine training. International Journal on Document Analysis and Recognition (IJDAR) **23**, 73–88 (2020)

[6] DeepMind, G.: What is a long context window? Google DeepMind engineers explain (2024), https://blog.google/technology/ai/long-context-window-ai-models/, accessed: 2024-05-18

[7] Dittrich, M.: 25 Jahre Rechtschreibreform: Keiser, Schikoree und Grislibär (2023), https://www.deutschlandfunk.de/rechtschreibreform-deutsche-sprache-100.html, accessed: 2024-05-17

[8] Dobischat, R., Käpplinger, B., Molzberger, G., Münk, D.: Bildung 2.1 für Arbeit 4.0? Springer (2019)

[9] Geis, A.J., Hoffmeyer-Zlotnik, J.H.: Zur Vercodung von Beruf, Branche und Prestige für die DDR, vol. 5. Campus Verl. (1991)

[10] Gessler, M., Howe, F.: From the reality of work to grounded work-based learning in German vocational education and training: Background, concept and tools. International journal for research in vocational education and training **2**(3), 214–238 (2015)

[11] Gewande, W.D.: Historische Entwicklung der staatlich anerkannten Ausbildungsberufe und ihrer Ordnungsmittel von 1934-1999: unter Berücksichtigung der mit deutschen Ausbildungsberufen gleichgestellten österreichischen Lehrberufe und gleichwertigen Facharbeiterberufen aus der ehemaligen DDR. Zentralamt der Bundesanst. für Arbeit, Geschäftsstelle für Veröff. (1999)

[12] Hamann, H.: The German federal courts dataset 1950–2019: From paper archives to linked open data. Journal of empirical legal studies **16**(3), 671–688 (2019)

[13] Harney, K.: Entstehung und Transformation der beruflichen Bildung als Institution–Systemischer Rück-und Ausblick. Bildung und Erziehung **73**(4), 346–357 (2020)

[14] Heliński, M., Kmieciak, M., Parkoła, T.: Report on the comparison of Tesseract and ABBYY FineReader OCR engines. online (2012)

[15] Herkner, V.: Grundzüge der Genese und Entwicklung einer korporatistischen Ordnung von Ausbildungsberufen. Berufsbildung in Wissenschaft und Praxis-BWP **42**(3), 16–19 (2013)

[16] IBM: What Is Optical Character Recognition (OCR)? (2024), https://www.ibm.com/blog/optical-character-recognition/, accessed: 2024-05-17

[17] Koistinen, M., Kettunen, K., Kervinen, J.: How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine. Proc. of LTC pp. 279–283 (2017)

[18] Kuppe, A.M., Lorig, B., Schwarz, H., Stöhr, A.: Ausbildungsordnungen und wie sie entstehen. Bundesinstitut für Berufsbildung (2015)

[19] Leeuwen, M.v., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven UP (2002)

[20] Li, J., Zhou, H., Huang, S., Cheng, S., Chen, J.: Eliciting the translation ability of large language models via multilingual finetuning with translation instructions (2024), https://arxiv.org/abs/2305.15083

[21] Maier, T.: Die Anwendbarkeit des Erlernten in den wandelnden Bildungs-und Arbeitslandschaften der 1970er-bis 2000er-Jahre. Leverkusen: Verlag Barbara Budrich (2021)

[22] Mani, S.: gImageReader: A Gtk/Qt front-end to tesseract-ocr (2024), https://github.com/manisandro/gImageReader, accessed: 2024-05-18

[23] OCR-D project. https://ocr-d.de/en/, accessed: 2024-05-15

[24] OCR-D Glossary (2024), https://ocr-d.de/en/spec/glossary, accessed: 2024-05-17

[25] OCR4all Workflow (2024), https://www.ocr4all.org/guide/user-guide/workflow, accessed: 2024-05-23

[26] Oliver, D.: Complexity in vocational education and training governance. Research in Comparative and International Education **5**(3), 261–273 (2010)

[27] OpenAI: Models (2024), https://platform.openai.com/docs/models, accessed: 2024-05-18

[28] OpenAI: Tokenizer (2024), https://platform.openai.com/tokenizer, accessed: 2024-05-18

[29] Protsch, P.: Zugang zu Ausbildung: Eine historisch vergleichende Perspektive auf den segmentierten Ausbildungsmarkt in (West-) Deutschland. Tech. rep., WZB Discussion Paper (2011)

[30] Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings. Applied Sciences **9**(22), 4853 (2019)

[31] Silva, G.P.e., Lins, R.D.: An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents. In: 2011 International Conference on Document Analysis and Recognition. pp. 553–557 (2011). https://doi.org/10.1109/ICDAR.2011.117

[32] Universität Hamburg: Research Data Repository. Available: https://www.fdr.uni-hamburg.de/ (2022), accessed: 2024-05-14

[33] VD16: VD 16 digital - Verzeichnis der im deutschen Sprachgebiet erschienenen Drucke des 16. Jahrhunderts der Bayerischen Staatsbibliothek. https://www.digitale-sammlungen.de/de/vd-16-digital-verzeichnis-der-im-deutschen-sprachgebiet/about, accessed: 2024-05-15

[34] VD17: VD 17 - Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts. http://www.vd17.de/, accessed: 2024-05-15

[35] VD18: VD 18 digital - Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 18. Jahrhunderts der Bayerischen Staatsbibliothek. https://vd18.gbv.de/viewer/index/, accessed: 2024-05-15

[36] Wang, Q.F., Yin, F., Liu, C.L.: Improving Handwritten Chinese Text Recognition by Unsupervised Language Model Adaptation. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 110–114 (2012). https://doi.org/10.1109/DAS.2012.46

[37] Wolf, S.: Past meets Present–the history of the German Vocational education and training model as a reflection frame to the prospect of the Egyptian model. Social Dimension and Participation in Vocational Education and Training p. 89 (2017)

[38] WordCounter: Words per Page (2024), https://wordcounter.net/words-per-page, accessed: 2024-05-18

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 37–44
DOI: 10.15439/2024F1772
ISSN 2300-5963 ACSIS, Vol. 41

Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation

Carina Geldhauser
0000-0002-9997-6710
Munich Centre for Machine Learning
Technical University Munich
Boltzmannstr. 3,85748 Garching bei München, Germany
and
ETH Zürich, Department of Mathematics
Rämistrasse 101, 8092 Zürich, Switzerland
Email: carina.geldhauser@ma.tum.de

Konstantin A. Malyshev
0009-0009-6338-5941
Saint Petersburg Theological Academy
nab. Obvodnogo kanala 17
Saint Petersburg, 191167, Russian Federation
Email: konstantin.a.malyshev@gmail.com

***Abstract*—We present a prototype for the integration of HTR transcription and semi-automated markup of textual features in the eScriptorium GUI.**

The prototype is designed for scholars working with ancient texts, who desire to perform standardized markup for a larger research project or digital edition. Motivated by research questions in Classics and Theology, we simultaneously investigate upcoming specific transcription challenges arising when working with ancient Greek manuscripts of majuscule type.

***Index Terms*—handwritten text recognition, named entity recognition, annotation, majuscule script**

I. INTRODUCTION

TEXTS play a central role in the work of a humanities scholar. New “distant reading” methods, where huge text corpora are analyzed through queries or statistical methods, open a whole new world of scholarly research questions.

However, not all objects of scholarly inquiry in the humanities are available as digital texts, and non-digital textual data must be, as a first step, extracted from the paper sources through Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR), and then processed further. Many research questions are sensitive to OCR quality, see section II-A1 for some examples.

As a second step, the digital plain text is *annotated*: Annotation[1] or markup, used interchangeably in this work, is the enrichment of a digital (plain) text with tags, categories, or standardized encoding for textual features.

This work is concerned with possibilities to integrate these two steps, transcription and annotation. We explore two possibilities on the example of a critical edition of manuscripts from the graeco-roman antique.

a) Example setting: The critical editions of works from the graeco-roman antique, used frequently by classicists and theologians, have to take into account that it is often not clear what is the actual text: Before the print age, works of popular authors had to be copied manually to be distributed. Both mistakes and deliberate alterations in the texts of such *manuscripts* happened, and *critical editions* have to display the variations between the available manuscripts.

As the variants of the text are a crucial part of the dataset, it is not advisable to use the same OCR post-processing methods as for printed material such as newspapers, see e.g. [1]. Those are often based on the comparison of words with available dictionaries, the algorithms remove hyphens etc, which means a potentially significant piece of information is lost through the post-processing.

Instead, at the current moment, scholars creating such digital editions are either manually correcting and annotating HTR or OCR- digitized manuscripts, or, in cases when the OCR software cannot deal with the used font/handwriting, the desired digital text needs to be established by manual typing, often employing a four-eyes principle to minimize transcription mistakes.

A separate research question is to draw conclusions by comparing the different variants. To automatize that comparison, a flawless *digital edition* (cf section II-B) with relevant annotation is crucial.

b) Motivation: Our work originates in the motivation to ease and facilitate the following task:

Group a set of available manuscripts into different tradition lines, i.e. order them by "similarity". For very significant ancient texts, e.g. Homer’s works or biblical texts (see the discussion in II-B), the aim may be to potentially establishing hypotheses of a common "parent manuscript". This is a very common, but challenging research question, see e.g. [2] for an example in medieval history. Depending on the precise formu-

---

[1] If the text is already available digitally, then annotation may be the only step to prepare for "distant reading" research questions.

©2024, PTI

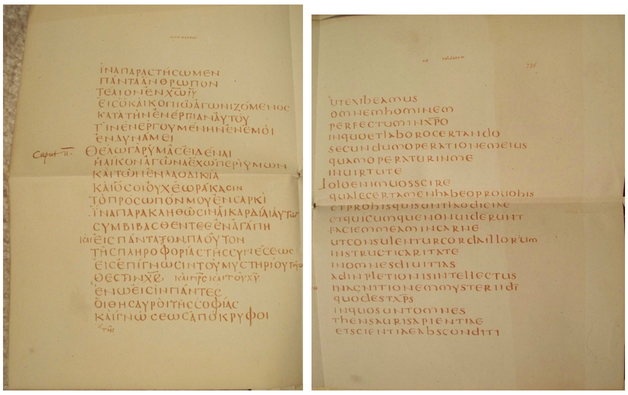


Fig. 1. Sample pages from Codex Claromontanus, taken from [3]. Both pages contain a textual variant of Colossians 1:28–2:3, nomina sacra with their characteristic overline bar, such as visible corrections by later scribes, most notably the insertion of $\kappa\alpha\iota$ in line 15 and diacritics in the Greek manuscript page.

lation, which might be cluster/group identification, similarity measures or the establishment of a proper "family tree" of manuscripts, the complexity of the task may vary, and usually involves non-trivial mathematical algorithms.

Particularly interesting for a comparison task as we described it above are so-called bilingual codices, see figure I-0b: these are manuscripts which contain the same content in two languages, in our case Greek and Latin. Here, a careful annotation of corresponding words and sentences may enable scholars in Classics or Theology to perform more diverse quantitative research tasks on the text at hand, and hence to potentially reach an improved hypotheses of manuscript relationships than if they were to look only at the Greek text.

To this aim, a very accurate transcription of the texts of each manuscript needs to be available, and a fine-grained annotation, with additional elements than what is commonly done, is necessary.

c) Main contribution of this work: The goal of our work was twofold: First, to address certain peculiarities in the ancient manuscripts at hand, listed in section III-B, by targeted HTR training to increase output quality. Second, we provide a prototype for relevant feature annotation, that may contribute to ease or speed up the preparatory work of classicists and theologians, so that they have more time to dedicate on their actual research questions.

A. Paper layout

The paper is structured as follows: In section II, we introduce the necessary background and state of the art. In section III, we describe our "dataset", i.e. Greek manuscripts of majuscule type, their special characteristics, and how the treatment of these characteristics are reflected in our work.

II. BACKGROUND

A. OCR/HTR for the humanities

A lot of effort has been put into the digitization of old texts. The most prominent are OCR methods for printed texts, e.g. old newspapers, collections of letters, or early print editions of major literary works. For input available as text printed on old paper, the output quality of an OCR method may be negatively affected by fading ink or poor paper quality, which could be further worsened by a suboptimal scanning process, leading to additional distortions in the image. Furthermore, the OCR quality is also negatively affected by heterogeneity within the printed text: these may be layout features like changes in fonts or colors of letters within a page, but also text-inherent features like spelling changes (which enlarge the diversity of possible character sequences) or low distinctiveness of characters in se, e.g. the long "s" and "f" in older German texts[2]. Still, for printed texts, significant progress was made, leading to very satisfactory output, in terms of character error rates (CER) lower than 2%, see Ströbel and Clematide for results using Transkribus[3], Wick et al (2018) using OCRopy[4] and Calamari[5] or Martinek et al. [5] using a convolutional and recurrent neural network, combined with suitable preprocessing (e.g. binarization) and data augmentation.

For handwritten text recognition (HTR), the situation is different, and depends very much on the concrete case at hand, therefore, we do not even attempt to give a comprehensive overview here. In general, as text recognition tasks fall in the category of supervised machine learning, their performance depends on the available data, in particular on the number of samples and their variability, and reflects potential biases that are present in the data, see e.g. [6].

Broadly speaking, most progress has been made for contemporary handwriting in widely-used languages and writing systems, especially in the Latin alphabet[6]. Line recognition seems to be essential: most off-the shelf products have issues with recognizing rotated text.

The quality of HTR results depend, among other factors, on the amount of training data with ground truth available. Very good results were achieved e.g. for Manu McFrench [7], a model trained on almost 78.000 lines (more than 4 million characters) of French handwriting[7] from the 17th to 20th century, which reached a character recognition accuracy of 90.56% in version 3.

The situation is different for so-called rare scripts or historical writing styles. As the wording intends, we might have a significantly lower amount of training data available. Other obstacles are built-in assumptions on the nature of the script, i.e. the writing being from left to right, on horizontal lines, etc. We refer to [8] for extensive reflections on the matter.

---

[2] See Ströbel et al [4] for a detailed analysis.

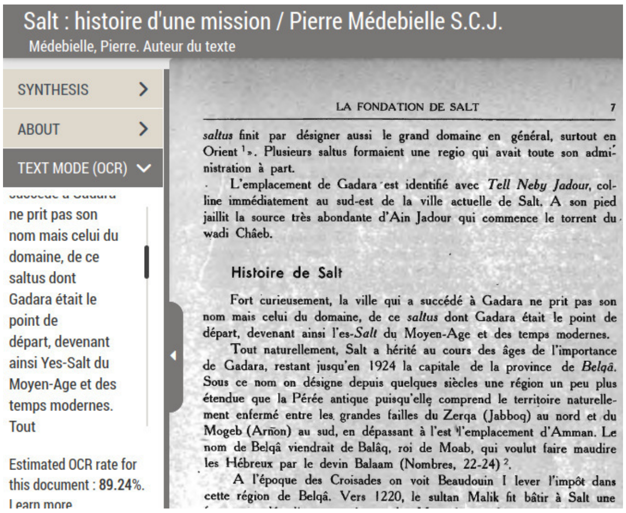
[3] https://transkribus.eu/Transkribus/

[4] https://github.com/tmbdev/ocropy

[5] https://github.com/Calamari-OCR/calamari

[6] We did perform tests of Tesseract, kraken (command-line) and Google's Cloud Vision API for several languages using Cyrillic alphabet, with satisfactory results, however, this is rather anecdotal evidence; we did not aim at a systematic evaluation of OCR tools in this work.

[7] According to the authors [7], the overwhelming majority of training data is in French language, mainly handwriting, but with some percentages of print, and a few thousand lines of Spanish and English handwriting were added in the training dataset.



Salt : histoire d'une mission / Pierre Médebielle S.C.J.
Médebielle, Pierre. Auteur du texte

ne prit pas son nom mais celui du domaine, de ce saltus dont Gadara était le point de départ, devenant ainsi Yes-Salt du Moyen-Age et des temps modernes. Tout

Estimated OCR rate for this document : **89.24%**.

LA FONDATION DE SALT 7

saltus finit par désigner aussi le grand domaine en général, surtout en Orient [1]». Plusieurs saltus formaient une regio qui avait toute son administration à part.

L'emplacement de Gadara est identifié avec *Tell Neby Jadour*, colline immédiatement au sud-est de la ville actuelle de Salt. A son pied jaillit la source très abondante d'Ain Jadour qui commence le torrent du wadi Châeb.

Histoire de Salt

Fort curieusement, la ville qui a succédé à Gadara ne prit pas son nom mais celui du domaine, de ce *saltus* dont Gadara était le point de départ, devenant ainsi l'es-*Salt* du Moyen-Age et des temps modernes.

Tout naturellement, Salt a hérité au cours des âges de l'importance de Gadara, restant jusqu'en 1924 la capitale de la province de *Belqâ*. Sous ce nom on désigne depuis quelques siècles une région un peu plus étendue que la Pérée antique puisqu'elle comprend le territoire naturellement enfermé entre les grandes failles du Zerqa (Jabboq) au nord et du Mogeb (Arnon) au sud, en dépassant à l'est l'emplacement d'Amman. Le nom de Belqâ viendrait de Balâq, roi de Moab, qui voulut faire maudire les Hébreux par le devin Balaam (Nombres, 22-24) [2].

A l'époque des Croisades on voit Beaudouin I lever l'impôt dans cette région de Belqâ. Vers 1220, le sultan Malik fit bâtir à Salt une

Fig. 2. Screenshot of a page in the Gallica collection [10], with estimated OCR accuracy given.

Furthermore, the current restrictions within the Transkribus HTR platform, which does not allow users to export models, even those which they trained themselves on their data, is a great disadvantage to the progress of HTR for rare scripts and historical writing styles.

1) Quick status assessment of DH for OCR: To summarize, Digital Humanities heavily profit from technical developments in OCR, but there is still space for improvement: the accuracy obtained so far may not be enough for reliable results on certain types of research question. For example, Chiron et al. [9] showed that OCR errors lead to significant missing relevant documents, as output of user queries in the OCRed Gallica collection: We see in Figure II-A1 a sample page from the Gallica collection, with an estimated OCR accuracy of below 90%. While this may sound fair, there is a drastic variance, and some infrequent, but highly relevant words may be wrongly ORC-red up to two thirds of their occurrences[8]. This may lead to biased query results, and therefore inaccurate answers to research questions.

Some scholars in the humanities, e.g. Smith and Cordell [11], even argue that the remaining errors in the digitized text is still "impeding advances in Digital Scholarship".

One of the applications of OCR/HTR where an accuracy of 90% is not enough is the establishment of a scholarly edition. As the research questions that motivated our prototype are closely related to the work on critical editions of ancient texts, we describe them briefly in the next paragraph.

B. Digital editions of ancient manuscripts

Digital scholarly editions are said to be the "crown jewels of Digital Humanities" [12], offering a plethora of ways of representing texts[9] and their transmission histories.

Roughly, we may distinguish three steps in digital editorial work of an ancient work: First, the provision of a digital (main) text[10] from the available sources, which usually involves the transcription/OCR/HTR of the raw material. Second, a markup/annotation[11] step, and third, an appropriate visualization[12], which includes tools for scholarly work with the edition. Scholarly editions of ancient manuscripts have certain peculiarities, among else:

a) Complex transmission histories: The transmission histories of centuries-old texts, such as Homer's works or biblical texts, is rich and highly complex, due to partial losses or damage, scribal errors, editorial decisions, and in general the huge spread and impact of these works. Hence, there may be many variants for the text to be presented in a new edition.

To make an illustrative example: there exist about 1000 manuscripts[13] of Homer‘s works, written on papyrus or parchment, and later on paper. These were copied by scribes multiple times over the course of two millennia, resulting in numerous losses and the introduction of many variations in the texts along the way. Which text is closest to the "real" Homer is an ongoing question of scholarly debate.

In the case of biblical manuscripts, the INTF Münster collects, curates and transcribes all available manuscripts of biblical texts. It prepares a scholarly critical edition of the New Testament, called the *Editio Critica Maior*, based on roughly 5800 manuscripts available to them at the moment.

b) Transparency about textual variations: A critical editions needs to be transparent about the different readings that are present in the different manuscripts. A "reading" in this case means an occurrence of a specific string on the paragraph, sentence, word or character level, in one manuscript, which is different from the string occurring in another manuscript. The two strings are then called "variants".

In order for this to be possible, the correctness of the transcription of each manuscript's text is extremely important, and human post-processing and error correction strictly necessary.

c) Less hierarchy of readings: A great opportunity of a digital edition is the possibility, through clever digital presentation and visual effects, to present different readings as equally plausible to the user, so that the user can decide for themselves which reading they want to adopt. This is a huge advantage w.r.t. the classical paper-based editions, which could only display one reading as the main text, and kept all variants in the apparatus.

---

[8] See [9], page 5.

[9] Sahle has argued in [13] for the usage of the word "document" instead of "text", but due to the dominance of "text" also in scholarly literature, we use this intuitive notion also for the purpose of this work.

[10] To establish the "correct text" in presence of variants is a huge field, which we do not want to enter here. To remain neutral in the debate, we use the uncommon wording "main text".

[11] As in the literature, we use "annotation" and "markup" as synonyms.

[12] With visualization, we mean the presentation of the TEI-XML code in a user-friendly environment, which helps scholars to answer their research questions. Available tools are e.g. The Versioning Machine or EVT https://visualizationtechnology.wordpress.com , but many larger projects build their own, customized tools.

[13] Estimate taken from https://www.lib.uchicago.edu/collex/exhibits/homer-print-transmission-and-reception-homers-works/homer-print/, last accessed: 30.09.2023.

C. Annotation

In this section, we describe the usage of annotation or markup of texts in a Digital Humanities context.

The aim of the annotation process is to provide a sufficient data enrichment as to allow adequate tools to answer a specific research question. Very popular is the markup of named entities, e.g. persons or places. Several stand-alone GUI tools[14] designed for scholars already exist, and they allow humanities scholars to find relevant attestations of named entities relevant to them. However, it is not possible to add further texts to the database of these tools, so that a scholar working on a different text has no advantage of them. Machine-Learning based NER tools such as *spaCy*[15] and *greCy*[16] have emerged lately, with accuracy strongly depending on context and the training data used by the developer. For example, as the models were trained on classical Greek, they do not lead to satisfactory results with late ancient Greek texts such as patristics. An overview of Named Entity Recognition models and challenges can be found in [14].

However, for some circumstances, the research question posed implies a need for manual annotation, as no appropriate tool exists. We illustrate our claim here on the example of creating visualizations on the usage of certain word categories, used in [15]: On the one hand, defining tags or categories is an independent research step that requires individual case decisions and therefore must be carried out by qualified personnel. On the other hand, most annotation tools are specialized, require significant time to learn and are limited in their distinctive features [16].

On the positive side, once mastered and provided with suitable annotated data, these tools not only make it possible to conduct detailed text-scientific research, but also to create visual forms of presentation of the text such as graphs, heat-maps and network graphs. Semantic markup of texts has been used for various purposes including categorizing handwritten annotations of an author [17], visualizing collaboration networks [18] and analyzing the lexical variance that occurs in the transmission of a medieval text [19].

As far as the creation of a digital edition is concerned, annotation of textual features in mark-up languages plays an integral role, as to provide functionalities to the user that they could not enjoy in a paper edition. Such annotation can be used to mark a variety of stylistic (such as text breaks and re-inkings) and semantic (such as place names, proper names and lemmatization) features, and we will describe this in detail in section III-B.

a) The text encoding initiative: Ideally, annotation follows a common standard, which allows for a group of scholars to build up upon each other's work. The text encoding initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. It develops and maintains a set of guidelines, the TEI Guidelines, which specify encoding methods, designed for the digital humanities community. According to the initiative[17], the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. The guidelines specify the semantics and interpretation of tags and attributes for basically all different textual components and concepts, from words to glyphs, persons, named entities etc.

In this work, we will follow the TEI guidelines for our automated markup, as it seems to be widely used within the digital classicist and digital theologian community.

III. GREEK MAJUSCULE TEXTS AND THEIR HTR

In this section, we introduce the reader to the main points to consider when dealing Greek Majuscule manuscripts, and how it shaped our work. To keep this paragraph short and readable to a diverse audience, we consciously simplify and may use wordings which are intuitively understandable, yet suboptimal for the experts. We apologize in advance to the well-versed classicist or theologian among our readers.

A. Classical text features to consider

In the following, we emphasize those textual features in Greek majuscule manuscripts that may be unknown to the reader, but which need careful treatment or algorithm adjustments in the HTR step.

a) Scriptio continua: scriptio continua is an inherent feature of majuscule manuscripts. It means that there is no visual gap between two words, but the letters of the text are aligned with uniform distance to each other until a line break. See Figure [3] for an example. There are usually also no punctuation signs, diacritics, or distinguished letter cases used.

A line break does not have to coincide with the end of a word or syllable, but the scribe may decide to break the line at any point. This may be simply, for aesthetic reasons, at a certain, specified distance to the margin, irrespective of the ending of a word. An example is the fourth century Codex Sinaiticus.[18].

Scriptio continua is one of the factors that deteriorate the HTR quality significantly. Indeed, as reported also by Perdiki [20], who used the commercial software Transcribus, most erroneous output was caused by scriptio continua, such as misrecognition of accents, wrong punctuation or wrong word token splitting. In our case, we implemented a separate word-split functionality, described in section IV-C to get a transcription in our modern way of writing ancient Greek.

b) Nomina Sacra: "nomina sacra" are specific abbreviations for frequent words such as "God", "Christ" or "Jerusalem", used in biblical codices. In biblical manuscripts written on papyri or parchment, such as in figure III-A0b or figure I-0b, a nomen sacrum is marked with an overline bar, it is usually two or three letters long, with letters taken from the

---

[14] E.g. Recogito https://recogito.pelagios.org/, or Kima https://data.geo-kima.org/, which is specifically for places in Hebrew script already exist that find the occurrences (attestations) of these named entities in digital texts.

[15] https://spacy.io/universe/project/grecy

[16] https://github.com/jmyerston/greCy

[17] https://tei-c.org/

[18] High-quality photographs of Codex Sinaiticus are openly available on https://codexsinaiticus.org.

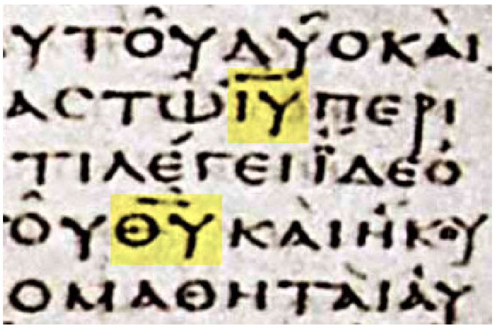


Fig. 3. Nomina Sacra (highlighted in yellow) in Codex Vaticanus, from [21], both are in genitive form, $\iota\upsilon$ represents the declined form of the word "Jesus", $\theta\upsilon$ represents the declined form of the word "God".

word it stems from. The last letter indicates the grammatical form of the abbreviated word. These specific characteristics make nomina sacra differ from generic abbreviations, e.g. on stone inscriptions.

Current research of New Testament textual scholars shows that the abbreviations are not unique, i.e. there might be a multitude of possible abbreviated forms used to encapsulate the same word. However, every nomen sacrum corresponds to a unique word.

c) Multiple hands and corrections: The major biblical codices all underwent changes by later scribes, to different extend or in different forms - diacritica were added, corrections were made, and more. The above-mentioned Codex Sinaiticus has undergone a particularly complex manual editing process over the centuries, and the investigation of the number of scribal hands, see e.g. [22], [23], is still ongoing research.

To our best knowledge, current machine-learning based results on scribe distinction for ancient or medieval handwritten texts are rather scarce, limited to binary classification ("is it scribe A or not?"), need a full page consisting of only one writing hand, and have, up to now, unsatisfactory accuracy. The automatic identification of different scribes in one manuscript remains a highly desired, but challenging feature.

B. Relevant textual features

The following text features are relevant towards a possible quantitative analysis, already within one text, before comparison takes place:

1) line breaks and other breaks in the original manuscript
2) multiple hands, especially corrections by a later scribe
3) re-inking (redrawing of letters)
4) highlighting, e.g. initial letters and flared letters
5) paratextual elements such as titles, marginal glosses, etc.

IV. First results in semi-automatic annotation of textual features

In this section, we first give an overview of our work, followed by subsections on each specific step.

A. Outline and Assumptions

In this work, we focus on the situation of scholars working with ancient manuscripts. Our **test case** and primary example are biblical manuscripts in majuscule style and scriptio continua, see Figure I-0b for an example. We assume that high quality images of the manuscripts are available, and the goal of the digital edition is to provide both a transcription of the text contained in these images, enriched with annotations in TEI-XML, and, in a later step, additional features that allow scholarly work with the text at hand, e.g. an apparatus, a way to compare the texts[19] displayed in different witnesses, or named entity recognition [24]–[26].

B. Overview

In our work, we developed a very first tool that may allow scholars working on digital editions to carry out annotations or other textual enrichment with less effort and without knowledge of a programming language. To be of real use for the humanities community, we aimed to make these tools available to the public in an easy-to-use form, i.e. inside a tool that is already in use, and with a visual frontend to avoid usage of scripts, codes or the opening of the command-line.

Taking into account the needs of our humanities colleagues, we decided to contribute to eScriptorium, a digital text production pipeline for print and handwritten texts using machine learning techniques [27]. The advantages of eScriptorium, from our perspective, are described in section IV-F.

The fork we created includes a couple of extra functions for annotation in TEI-XML standard (see section II-C) in a semi-automated manner, i.e. the user uploads a text file which contains the words, names or places they wish to be annotated. Our fork allows to export this skeleton annotation as “custom XML”, see figure IV-E, which can be further enriched manually in any other tool, e.g. Oxygen XML editor.

C. Word splitting

For word-splitting tasks we decided to use the *SymSpell* library[20]. It is based on so called "Symmetric Delete" spelling correction algorithm[21]. The crucial advantage of this library is that it can not only split the continuous non-space string into words, but also can correct errors to some degree. This is an important case, because the OCR/HTR step never delivers 100% recognition accuracy. For example, the string "cnebigeIefant" has two errors: "c" instead of "o", and "I" instead of "l", and if they are fixed and the words are splitted, we get "one big elefant". The SymSpell library must be provided with the words dictionary, sorted by their assumed frequency in the given text. In addition, it can also use a bigram or trigram dictionary, where occurrences of two or three words

---

[19] Here, we take a viewpoint similar to Sahle’s [13] “Text als Fassung”, where we define the “text” as the “reading”, i.e. the ensemble of words presented in a particular physical witness.

[20] https://github.com/wolfgarbe/SymSpell

[21] We refer the interested reader to the introductory notes in https://seekstorm.com/blog/fast-word-segmentation-noisy-text/ and https://seekstorm.com/blog/fast-approximate-string-matching/.

are sorted according to their frequency in the language. The most probable words candidates are defined according to the frequency dictionaries and the Damerau-Levenshtein distance between the given string and the candidate. Note that this dictionary should also include nomina sacra (see III-A) or other abbreviations, in order to increase accuracy.

At the moment, two unsolved problems remain: First, each text has its own distribution of words frequencies, and this can vary a lot. This means that the standard language dictionary should not be used, but an additional custom-dictionary-generation step is required, which takes into account the peculiarities of the given text. The second unsolved question is: how to adjust the maximum Damerau-Levenshtein spelling correction distance to the optimal value? As far as we oversee the issue, this step is dependent on the accuracy of the HTR step.

D. Semi-automated annotation

Though machine-learning based NER tools are available, see the discussion in section II-C, we decided for a rule-based implementation, as also other recent work, e.g. the *Opera Graeca Adnotata* [28]. Our decision relied on the following thoughts:

First, there are currently no good greCy models for majuscule Greek on which we could rely on. This means also that annotation will not be available to a user with datasets of different type than what an external tool can handle. A rule-based implementation is independent on external tools.

Second, despite necessary post-processing might always be done for delicate tasks related to digital editions, our users prefer reliable output, instead of having to deal with "false-positives": Therefore, we prefer to annotate less, but to do that with the highest precision possible, avoiding ML biases.

Third, a direct linkage with external tools is both delicate to implement and vulnerable to break down, due to software changes from version to version.

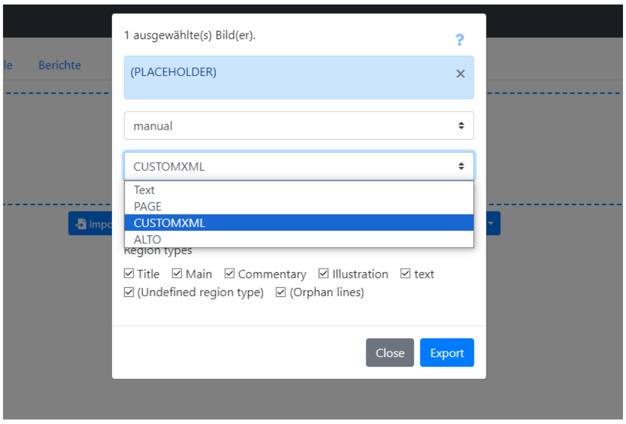
Therefore, we decided for the following approach: the user creates a list of relevant words for annotation, using a tool of their choice. This list, in plain text format, can then be uploaded into our eScriptorium fork, where the annotation algorithm creates TEI-standard markup of all words in the list with one click. This means that all grammatical forms of a word have to be provided in the word list.

E. Annotated features

We coded and incorporated into our fork of eScriptorium a semi-automated TEI mark-up functionality for Hebrew and Greek personal names and place names, numerals, nomina sacra,(see section III-A), punctuation signs and line and page breaks in the original manuscript.

F. Why eScriptorium?

After preliminary tests using eScriptorium, Transkribus, but also tools without GUI, such as Tesseract, kraken and Google Cloud Vision API's Document OCR tool, we realized that very few transcription tools are able to deal with Greek majuscule manuscripts: some were simply not designed for ancient texts, and therefore gave terrible results, probably due to the uncommon font and the scriptio continua.



1 ausgewählte(s) Bild(er).

(PLACEHOLDER)

manual

CUSTOMXML

Region types

☑ Title ☑ Main ☑ Commentary ☑ Illustration ☑ text
☑ (Undefined region type) ☑ (Orphan lines)

Fig. 4. Screenshot of our prototype front-end: exporting the annotated text XML file.

As a GUI was important to us due to its friendliness to the relevant scholarly community, the decision stood between eScriptorium and Transkribus.

As our tests found no significantly better performance of Transkribus w.r.t. eScriptorium we decided to take advantage of the open source nature of eScriptorium, which made us more flexible, saved resources, and hopefully allows for an easy adaptation of our fork in the relevant community.

During the work with eScriptorium, we came to enjoy its additional advantages: The core code base appeared to be clean, understandable and well-designed. The internal architecture of the tool is modular, well-structured and easy-to-extent. This makes it ideal for an expandable open-source project. It was not complicated to integrate our custom code into the eScriptorium core.

G. Annotation of nomina sacra

The current fork of our project[22] contains a function that expands abbreviated nomina sacra from the transcribed text. For this, the user needs to provide a list of abbreviations used, or adapt our default, provided for Greek and Latin letters.

The advantage of this approach is its versatility towards different writing styles in Greek, Latin, or any other language the user works with inside eScriptorium: This way of expanding a nomen sacrum works as soon as the transcription obtained by eScriptorium's transcription step is accurate, it is independent of the model parameters used or trained. As such, it can be used also to expand and annotate nomina sacra in texts written in minuscule style, or even printed editions, e.g. by 19th century scholars like K. von Tischendorf [29].

Another method, dependent on the used model, is to link transcription and annotation of nomina sacra and other "relevant" features directly. With "relevant" we intend annotation that can be used to answer research questions or to serve the

---

[22] Available at https://gitlab.com/archtype/escriptorium/-/branches

reader of a critical edition in their exploration and understanding of the text. With "direct linking" we intend an integrated HTR and annotation pipeline, namely to recognize a nomen sacrum from the overline bar in the handwritten document, and then to suggest the correct expansion directly in an annotation, without a human-created list.

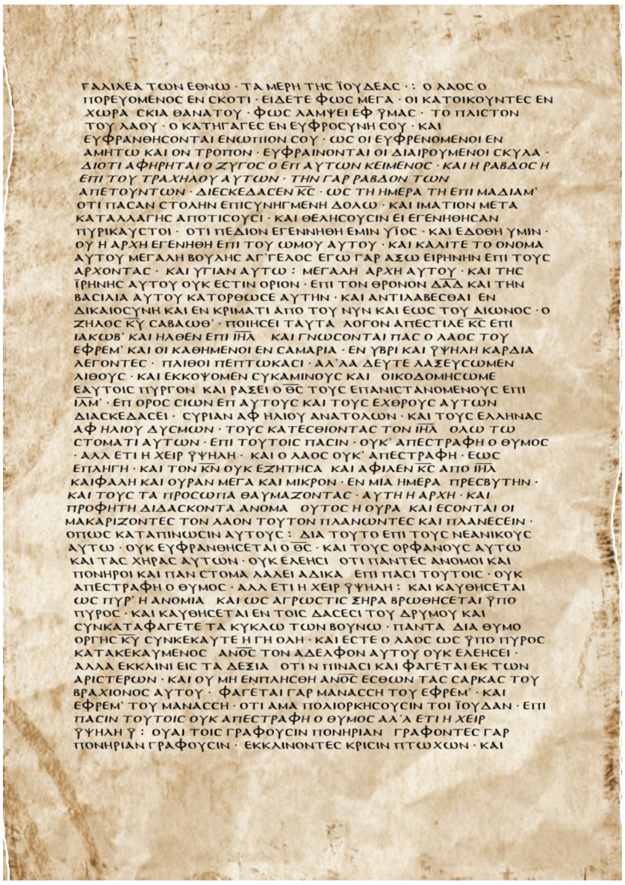
This implies training of the underlying machine learning model to an extremely high accuracy, in order to recognize nomina sacra by the overline bar used by the scribe on the abbreviated version (recall Figure III-A0b). We discuss this in section V.

H. Annotation of numerals

The annotation of numerals is less common, as only relevant for very specific research questions. We included it both as a 'placeholder annotation', i.e. it may be replaced by the user, to instead annotate something else, and in order to show the limitations of the algorithmic side: Our algorithm technically checks all words in their order of appearance in the text, linearly going through the text word-by-word. This linear processing will recognize a numeral and annotate it immediately. However, if this numeral is part of a compound number word, the linear processing will not be successful.

I. Towards an integrated HTR and annotation pipeline

As discussed above, an integrated HTR and annotation pipeline could use certain image features directly for "relevant" annotation. To achieve the necessary high accuracies and density of appearance of nomina sacra, we used a version of “data augmentation”: we created 50 pages of artificial digital manuscripts [30] containing all possible grammatical forms of the nomina sacra, available on Zenodo[23]. See figure IV-I for an example. Our set of augmented images uses genuine biblical unicial fonts[24] and incorporates a variety of visual characteristics that deteriorates the quality of a scanned manuscript page, such as distortions, heterogeneities in the background color, damaged or partially degraded paper ("dark spots") etc.



ΓΑΛΙΛΕΑ ΤΩΝ ΕΘΝΩ · ΤΑ ΜΕΡΗ ΤΗϹ ΪΟΥΔΕΑϹ · : Ο ΛΑΟϹ Ο
ΠΟΡΕΥΟΜΕΝΟϹ ΕΝ ϹΚΟΤΙ · ΕΙΔΕΤΕ ΦΩϹ ΜΕΓΑ · ΟΙ ΚΑΤΟΙΚΟΥΝΤΕϹ ΕΝ
ΧΩΡΑ ϹΚΙΑ ΘΑΝΑΤΟΥ · ΦΩϹ ΛΑΜΨΕΙ ΕΦ ΫΜΑϹ · ΤΟ ΠΛΙϹΤΟΝ
ΤΟΥ ΛΑΟΥ · Ο ΚΑΤΗΓΑΓΕϹ ΕΝ ΕΥΦΡΟϹΥΝΗ ϹΟΥ · ΚΑΙ
ΕΥΦΡΑΝΘΗϹΟΝΤΑΙ ΕΝΩΠΙΟΝ ϹΟΥ · ΩϹ ΟΙ ΕΥΦΡΕΝΟΜΕΝΟΙ ΕΝ
ΑΜΗΤΩ ΚΑΙ ΟΝ ΤΡΟΠΟΝ · ΕΥΦΡΑΙΝΟΝΤΑΙ ΟΙ ΔΙΑΙΡΟΥΜΕΝΟΙ ϹΚΥΛΑ ·
ΔΙΟΤΙ ΑΦΗΡΗΤΑΙ Ο ΖΥΓΟϹ Ο ΕΠ ΑΥΤΩΝ ΚΕΙΜΕΝΟϹ · ΚΑΙ Η ΡΑΒΔΟϹ Η
ΕΠΙ ΤΟΥ ΤΡΑΧΗΛΟΥ ΑΥΤΩΝ · ΤΗΝ ΓΑΡ ΡΑΒΔΟΝ ΤΩΝ
ΑΠΕΤΟΥΝΤΩΝ · ΔΙΕϹΚΕΔΑϹΕΝ Κ̅Ϲ̅ · ΩϹ ΤΗ ΗΜΕΡΑ ΤΗ ΕΠΙ ΜΑΔΙΑΜ’
ΟΤΙ ΠΑϹΑΝ ϹΤΟΛΗΝ ΕΠΙϹΥΝΗΓΜΕΝΗ ΔΟΛΩ · ΚΑΙ ΙΜΑΤΙΟΝ ΜΕΤΑ
ΚΑΤΑΛΛΑΓΗϹ ΑΠΟΤΙϹΟΥϹΙ · ΚΑΙ ΘΕΛΗϹΟΥϹΙΝ ΕΙ ΕΓΕΝΗΘΗϹΑΝ
ΠΥΡΙΚΑΥϹΤΟΙ · ΟΤΙ ΠΕΔΙΟΝ ΕΓΕΝΝΗΘΗ ΕΜΙΝ ΥΪΟϹ · ΚΑΙ ΕΔΟΘΗ ΥΜΙΝ ·
ΟΥ Η ΑΡΧΗ ΕΓΕΝΗΘΗ ΕΠΙ ΤΟΥ ΩΜΟΥ ΑΥΤΟΥ · ΚΑΙ ΚΑΛΙΤΕ ΤΟ ΟΝΟΜΑ
ΑΥΤΟΥ ΜΕΓΑΛΗ ΒΟΥΛΗϹ ΑΓ’ΓΕΛΟϹ ΕΓΩ ΓΑΡ ΑΞΩ ΕΙΡΗΝΗΝ ΕΠΙ ΤΟΥϹ
ΑΡΧΟΝΤΑϹ · ΚΑΙ ΥΓΙΑΝ ΑΥΤΩ : ΜΕΓΑΛΗ ΑΡΧΗ ΑΥΤΟΥ · ΚΑΙ ΤΗϹ
ΪΡΗΝΗϹ ΑΥΤΟΥ ΟΥΚ ΕϹΤΙΝ ΟΡΙΟΝ · ΕΠΙ ΤΟΝ ΘΡΟΝΟΝ Δ̅Α̅Δ̅ ΚΑΙ ΤΗΝ
ΒΑϹΙΛΙΑ ΑΥΤΟΥ ΚΑΤΟΡΘΩϹΕ ΑΥΤΗΝ · ΚΑΙ ΑΝΤΙΛΑΒΕϹΘΑΙ ΕΝ
ΔΙΚΑΙΟϹΥΝΗ ΚΑΙ ΕΝ ΚΡΙΜΑΤΙ ΑΠΟ ΤΟΥ ΝΥΝ ΚΑΙ ΕΩϹ ΤΟΥ ΑΙΩΝΟϹ · Ο
ΖΗΛΟϹ Κ̅Υ̅ ϹΑΒΑΩΘ’ · ΠΟΙΗϹΕΙ ΤΑΥΤΑ ΛΟΓΟΝ ΑΠΕϹΤΙΛΕ Κ̅Ϲ̅ ΕΠΙ
ΙΑΚΩΒ’ ΚΑΙ ΗΛΘΕΝ ΕΠΙ Ι̅Η̅Λ̅ ΚΑΙ ΓΝΩϹΟΝΤΑΙ ΠΑϹ Ο ΛΑΟϹ ΤΟΥ
ΕΦΡΕΜ’ ΚΑΙ ΟΙ ΚΑΘΗΜΕΝΟΙ ΕΝ ϹΑΜΑΡΙΑ · ΕΝ ΥΒΡΙ ΚΑΙ ΫΨΗΛΗ ΚΑΡΔΙΑ
ΛΕΓΟΝΤΕϹ · ΠΛΙΘΟΙ ΠΕΠΤΩΚΑϹΙ · ΑΛ’ΛΑ ΔΕΥΤΕ ΛΑΞΕΥϹΩΜΕΝ
ΛΙΘΟΥϹ · ΚΑΙ ΕΚΚΟΨΟΜΕΝ ϹΥΚΑΜΙΝΟΥϹ ΚΑΙ ΟΙΚΟΔΟΜΗϹΩΜΕ
ΕΑΥΤΟΙϹ ΠΥΡΓΟΝ ΚΑΙ ΡΑΞΕΙ Ο Θ̅Ϲ̅ ΤΟΥϹ ΕΠΑΝΙϹΤΑΝΟΜΕΝΟΥϹ ΕΠΙ
Ι̅Λ̅Μ̅’ · ΕΠ ΟΡΟϹ ϹΙΩΝ ΕΠ ΑΥΤΟΥϹ ΚΑΙ ΤΟΥϹ ΕΧΘΡΟΥϹ ΑΥΤΩΝ
ΔΙΑϹΚΕΔΑϹΕΙ · ϹΥΡΙΑΝ ΑΦ ΗΛΙΟΥ ΑΝΑΤΟΛΩΝ · ΚΑΙ ΤΟΥϹ ΕΛΛΗΝΑϹ
ΑΦ ΗΛΙΟΥ ΔΥϹΜΩΝ · ΤΟΥϹ ΚΑΤΕϹΘΙΟΝΤΑϹ ΤΟΝ Ι̅Η̅Λ̅ ΟΛΩ ΤΩ
ϹΤΟΜΑΤΙ ΑΥΤΩΝ · ΕΠΙ ΤΟΥΤΟΙϹ ΠΑϹΙΝ · ΟΥΚ’ ΑΠΕϹΤΡΑΦΗ Ο ΘΥΜΟϹ
· ΑΛΛ ΕΤΙ Η ΧΕΙΡ ΫΨΗΛΗ · ΚΑΙ Ο ΛΑΟϹ ΟΥΚ’ ΑΠΕϹΤΡΑΦΗ · ΕΩϹ
ΕΠΛΗΓΗ · ΚΑΙ ΤΟΝ Κ̅Ν̅ ΟΥΚ ΕΖΗΤΗϹΑ ΚΑΙ ΑΦΙΛΕΝ Κ̅Ϲ̅ ΑΠΟ Ι̅Η̅Λ̅
ΚΑΙΦΑΛΗ ΚΑΙ ΟΥΡΑΝ ΜΕΓΑ ΚΑΙ ΜΙΚΡΟΝ · ΕΝ ΜΙΑ ΗΜΕΡΑ ΠΡΕϹΒΥΤΗΝ ·
ΚΑΙ ΤΟΥϹ ΤΑ ΠΡΟϹΩΠΑ ΘΑΥΜΑΖΟΝΤΑϹ · ΑΥΤΗ Η ΑΡΧΗ · ΚΑΙ
ΠΡΟΦΗΤΗ ΔΙΔΑϹΚΟΝΤΑ ΑΝΟΜΑ ΟΥΤΟϹ Η ΟΥΡΑ ΚΑΙ ΕϹΟΝΤΑΙ ΟΙ
ΜΑΚΑΡΙΖΟΝΤΕϹ ΤΟΝ ΛΑΟΝ ΤΟΥΤΟΝ ΠΛΑΝΩΝΤΕϹ ΚΑΙ ΠΛΑΝΕϹΕΙΝ ·
ΟΠΩϹ ΚΑΤΑΠΙΝΩϹΙΝ ΑΥΤΟΥϹ : ΔΙΑ ΤΟΥΤΟ ΕΠΙ ΤΟΥϹ ΝΕΑΝΙΚΟΥϹ
ΑΥΤΩ · ΟΥΚ ΕΥΦΡΑΝΘΗϹΕΤΑΙ Ο Θ̅Ϲ̅ · ΚΑΙ ΤΟΥϹ ΟΡΦΑΝΟΥϹ ΑΥΤΩ
ΚΑΙ ΤΑϹ ΧΗΡΑϹ ΑΥΤΩΝ · ΟΥΚ ΕΛΕΗϹΙ ΟΤΙ ΠΑΝΤΕϹ ΑΝΟΜΟΙ ΚΑΙ
ΠΟΝΗΡΟΙ ΚΑΙ ΠΑΝ ϹΤΟΜΑ ΛΑΛΕΙ ΑΔΙΚΑ ΕΠΙ ΠΑϹΙ ΤΟΥΤΟΙϹ · ΟΥΚ
ΑΠΕϹΤΡΑΦΗ Ο ΘΥΜΟϹ · ΑΛΛ ΕΤΙ Η ΧΕΙΡ ΫΨΗΛΗ : ΚΑΙ ΚΑΥΘΗϹΕΤΑΙ
ΩϹ ΠΥΡ’ Η ΑΝΟΜΙΑ ΚΑΙ ΩϹ ΑΓΡΩϹΤΙϹ ΞΗΡΑ ΒΡΩΘΗϹΕΤΑΙ ΫΠΟ
ΠΥΡΟϹ · ΚΑΙ ΚΑΥΘΗϹΕΤΑΙ ΕΝ ΤΟΙϹ ΔΑϹΕϹΙ ΤΟΥ ΔΡΥΜΟΥ ΚΑΙ
ϹΥΝΚΑΤΑΦΑΓΕΤΕ ΤΑ ΚΥΚΛΩ ΤΩΝ ΒΟΥΝΩ · ΠΑΝΤΑ ΔΙΑ ΘΥΜΟ
ΟΡΓΗϹ Κ̅Υ̅ ϹΥΝΚΕΚΑΥΤΕ Η ΓΗ ΟΛΗ · ΚΑΙ ΕϹΤΕ Ο ΛΑΟϹ ΩϹ ΫΠΟ ΠΥΡΟϹ
ΚΑΤΑΚΕΚΑΥΜΕΝΟϹ Α̅Ν̅Ο̅Ϲ̅ ΤΟΝ ΑΔΕΛΦΟΝ ΑΥΤΟΥ ΟΥΚ ΕΛΕΗϹΕΙ ·
ΑΛΛΑ ΕΚΚΛΙΝΙ ΕΙϹ ΤΑ ΔΕΞΙΑ ΟΤΙ Ν ΠΙΝΑϹΙ ΚΑΙ ΦΑΓΕΤΑΙ ΕΚ ΤΩΝ
ΑΡΙϹΤΕΡΩΝ · ΚΑΙ ΟΥ ΜΗ ΕΝΠΛΗϹΘΗ Α̅Ν̅Ο̅Ϲ̅ ΕϹΘΩΝ ΤΑϹ ϹΑΡΚΑϹ ΤΟΥ
ΒΡΑΧΙΟΝΟϹ ΑΥΤΟΥ · ΦΑΓΕΤΑΙ ΓΑΡ ΜΑΝΑϹϹΗ ΤΟΥ ΕΦΡΕΜ’ · ΚΑΙ
ΕΦΡΕΜ’ ΤΟΥ ΜΑΝΑϹϹΗ · ΟΤΙ ΑΜΑ ΠΟΛΙΟΡΚΗϹΟΥϹΙΝ ΤΟΙ ΪΟΥΔΑΝ · ΕΠΙ
ΠΑϹΙΝ ΤΟΥΤΟΙϹ ΟΥΚ ΑΠΕϹΤΡΑΦΗ Ο ΘΥΜΟϹ ΑΛ’Λ ΕΤΙ Η ΧΕΙΡ
ΫΨΗΛΗ Ϋ : ΟΥΑΙ ΤΟΙϹ ΓΡΑΦΟΥϹΙΝ ΠΟΝΗΡΙΑΝ ΓΡΑΦΟΝΤΕϹ ΓΑΡ
ΠΟΝΗΡΙΑΝ ΓΡΑΦΟΥϹΙΝ · ΕΚΚΛΙΝΟΝΤΕϹ ΚΡΙϹΙΝ ΠΤΩΧΩΝ · ΚΑΙ

Fig. 5. Sample page from our data augmentation set. The dataset is available on Zenodo [30].

V. SUMMARY AND OUTLOOK: INTEGRATING HTR AND ANNOTATION

The presented prototype provides a semi-automated markup functionality to the HTR transcription step of eScriptorium. We started with a couple of exemplary features to annotate through a rule-based algorithm. While a rule-based approach provides maximum accuracy, it needs well-prepared input of the user. Also, the annotation of multi-word numerals turned out to be difficult.

a) Current work: At the moment, we work on the improvement of the performance of the new transcription model in terms of the recognition of nomina sacra. While our team developed a bounding box model (via Kraken) achieving high accuracy levels (>90%), the training of a baseline model (done directly in eScriptorium) resulted in lower accuracy rates. The improvement of our transcription model in a baseline format will continue through the feeding of additional training data. We envision that the complete automated annotation of nomina sacra will be possible through these improvements.

b) Discussion: This work was motivated by the work of our colleagues on digital editions and manuscript comparison.

We are happy that our prototype saves their time in the annotation step of their work: the most common or basic annotations are already taken care of algorithmically, and the researcher can correct both OCR errors and missing annotations in the same round of manual interventions.

Note, however, that this is only a small part of the time that our colleagues invest in building a good dataset / a good critical edition: For these aims, an OCR accuracy of 90 % is barely enough to make OCR-transcription and subsequent manual correction as fast as they were when they transcribed completely manually.

Hence, in order for a digital tool to be truly useful for them, a much higher OCR accuracy is needed.

c) Further Steps: A few open questions and improvement points have been pointed out in the various sections above, e.g. in IV-C we pointed out the need of custom-dictionary-generation and other open issues in the word split-

---

[23] URL: https://zenodo.org/records/12755706

[24] Available at http://individual.utoronto.ca/atloder/uncialfonts.html

ting step. With regard to further developments in the project, we aim to explore other annotation options for our prototype, to accommodate a larger variety of research questions. One idea is adding GPS information for annotated place names. In view of scholarly research questions on manuscript transmission history, scribal habits and cultural heritage questions, the annotation of a variety of more subtle visual features, such as different scribal hands, deletions and re-inkings, are also envisioned.

ACKNOWLEDGMENT

This work was partially supported by the German Federal and Bavarian Ministries of Education and Research, through its financing of the Munich Center for Machine Learning. We thank J. Heilmann and N. Müller for supplying us ground truth data, and our students J. Ehness, N. Dominko Kobilica, P. Kumar and I. Tuncel for their help in the code development and testing phase.

REFERENCES

[1] B. Alex, C. Grover, E. Klein, and R. Tobin, "Digitised historical text: Does it have to be mediocre?." in *KONVENS*, 2012, pp. 401–409.

[2] P. Roelli and D. Bachmann, "Towards generating a stemma of complicated manuscript traditions: Petrus alfonsi's dialogus," *Revue d'histoire des textes*, vol. 5, pp. 307–321, 2010.

[3] Wikimedia, "Codex claromontanus, the greek text of colossians 1:28-2:3," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://en.wikipedia.org/wiki/Codex_Claromontanus#/media/File:Claromontanus_2_greek.jpg

[4] P. B. Ströbel, S. Clematide, and M. Volk, "How much data do you need? about the creation of a ground truth for black letter and the effectiveness of neural ocr," *Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, 2020*, 2020.

[5] J. Martínek, L. Lenc, and P. Král, "Training strategies for ocr systems for historical documents," in *Artificial Intelligence Applications and Innovations: 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15*. Springer, 2019, pp. 362–373.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] A. Chagué, T. Clérice, J. Norindr, M. Humeau, B. Davoury, E. Van Kote, A. Mazoue, M. Faure, and S. Doat, "Manu mcfrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets," in *Digital Humanities 2023: Collaboration as Opportunity*, 2023.

[8] P. Stokes and B. Kiessling, "Sharing data for handwritten text recognition (htr)," *Digital Humanities in Practice*, 2024.

[9] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J.-P. Moreux, "Impact of ocr errors on the use of digital libraries: towards a better access to information," in *2017 ACM/IEEE joint conference on digital libraries (JCDL)*. IEEE, 2017, pp. 1–4.

[10] T. G. Collection, "Pierre médebielle s.c.j. gallica (auteur); salt: Histoire d'une mission (texte)," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://gallica.bnf.fr/ark:/12148/bpt6k91248315/f7.item#

[11] D. A. Smith and R. Cordell, "A research agenda for historical and multilingual optical character recognition," *NUlab, Northeastern University.@ https://ocr. northeastern. edu/report*, p. 36, 2018.

[12] E. Pierazzo, "A rationale of digital documentary editions," *Literary and linguistic computing*, vol. 26, no. 4, pp. 463–477, 2011.

[13] P. Sahle, "What is a scholarly digital edition?" *Digital scholarly editing: Theories and practices*, vol. 1, pp. 19–39, 2016.

[14] K. Pakhale, "Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges," *arXiv preprint arXiv:2309.14084*, 2023.

[15] W. Riess, "Prolegomena zu einer digitalen althistorischen Gewaltforschung: Gewaltmuster bei Solon, Alkibiades und Arat im Vergleich," *Klio*, vol. 102, no. 2, pp. 445–473, 2020.

[16] A. Przepiórkowski, "Tei p5 as an xml standard for treebank encoding," in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, 2009, pp. 149–160.

[17] S. A. and A. M. Del Grosso, "Giorgio bassani's notes between tradition and innovation." *Digital Humanities 2023: Book of Abstracts*, 2023.

[18] M. A. Cipolla, A. Cappellotto, M. Rospocher *et al.*, "Collaboration practices between people and tools: the case of" snorra edda. a collaborative bibliography (snecb)"," in *Digital Humanities 2023: Book of Abstracts*, 2023, pp. 93–94.

[19] S. Moors, "Constrained. a computational study of the influence of formal characteristics on the transmission of the middle dutch martijn trilogy by jacob van maerlant." *Digital Humanities 2023: Book of Abstracts*, 2023.

[20] E. Perdiki, "Preparing big manuscript data for hierarchical clustering with minimal htr training," *Journal of Data Mining & Digital Humanities*, no. Sciences of Antiquity and digital humanities, 2023.

[21] Wikimedia, "Nomina sacra in codex vaticanus john 1," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://commons.wikimedia.org/wiki/File:Nomina_Sacra_in_Codex_Vaticanus_John_1.jpg

[22] D. Jongkind, *Scribal Habits of Codex Sinaiticus*. Gorgias Press, 2013.

[23] A. Wilson, "Scribal habits in greek new testament manuscripts," *Filología neotestamentaria*, vol. 24, pp. 95–126, 2011.

[24] R. Hanslo, "Deep learning transformer architecture for named-entity recognition on low-resourced languages: State of the art results," in *PROCEEDINGS OF THE 2022 17TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2022, pp. 53–60.

[25] R. Sharma, D. Chauhan, and R. Sharma, "Named entity recognition system for the biomedical domain," in *PROCEEDINGS OF THE 2022 17TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2022, pp. 837–840, 17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, BULGARIA, SEP 04-07, 2022.

[26] R. Hanslo, "Evaluation of neural network transformer models for named-entity recognition on low-resourced languages," in *PROCEEDINGS OF THE 2021 16TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2021, pp. 115–119, 16th Conference on Computer Science and Intelligence Systems (FedCSIS), ELECTR NETWORK, SEP 02-05, 2021.

[27] B. Kiessling, "Kraken-an universal text recognizer for the humanities," in *Proceedings of the DH2019 Conference*, 2019.

[28] G. G. Celano, "Opera graeca adnotata: Building a 34m+ token multilayer corpus for ancient greek," *arXiv preprint arXiv:2404.00739*, 2024.

[29] C. v. Tischendorf, *Novum Testamentum graece*. Leipzig, 1841.

[30] C. Geldhauser, "Artificially created image files resembling ancient greek manuscripts in majuscule script (version v1) [data set], zenodo." 2024, [Online; accessed July 23, 2024]. [Online]. Available: https://doi.org/10.5281/zenodo.12755706

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 45–52
DOI: 10.15439/2024F2752
ISSN 2300-5963 ACSIS, Vol. 41

AI in the Workplace: Who Is Using It and Why? A Look at the Driving Forces Behind Artificial Intelligence in German Companies

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

Myriam Baum
0000-0003-1230-3984
Federal Institute for Vocational Education and Training
Friedrich-Ebert Allee 114-116,
53113 Bonn, Germany
Email: baum@bibb.de

***Abstract*—We examine the spread and usage of artificial intelligence (AI) in German companies. The study analyzes company characteristics that favor or inhibit the adoption of AI. Hypotheses are developed that include factors such as the type of job tasks, level of innovation, degree of digitalization, company size, and industry affiliation. Empirical quantitative data from the BIBB Training Panel shows that AI usage is slowly but steadily increasing, particularly in larger companies. The multivariate analysis highlights that an advanced digital infrastructure and an innovative corporate culture are crucial for the usage of AI applications. The findings aim to support political and business strategic decision-making processes and to promote the implementation of AI in companies while considering ethical considerations.**

***Index Terms*—Artificial Intelligence (AI), company level data, AI at the labor market, job tasks.**

I. INTRODUCTION

WHEN talking about technological transformation and digital technologies, artificial intelligence (AI) is of central importance, as it is a key technology for the digital transformation. The progressive integration of AI into the world of work supposedly marks an upheaval in the way companies operate and compete. The development of AI is not only a technical innovation, but also influences the structure of labor markets, the design of jobs and consequently, the qualification requirements for employees [1] - [4].

AI innovations have changed the perception on which job tasks are substitutable by technologies [1], [2]. In earlier work on changes in job tasks due to technology mostly routine manual tasks were labeled as substitutable [5]. Nowadays the debate also includes the possible substitutability of more complex or analytic tasks [6], as well as the emerging of new job tasks [1]. This shift in perception highlights the importance of analyzing the impact of digital technologies on the economy and society.

The fundamental difference between AI and other digital technologies is that AI technologies, particularly through the use of machine learning, focus on automating and optimizing tasks by learning from data, which can significantly enhance efficiency and decision-making in both physical and non-physical work processes. Deep learning is a subset of AI techniques that uses layered neural networks to analyze various levels of data.

So far, there is no clear definition of AI. AI is a collective term that is filled differently depending on the considered time period and technologies.

The concept of AI began with the first computers [7, pp. 529]. Creativity, self-improvement, and language use were quickly identified as important criteria for defining AI [8, p. 18]. In this sense, AI aims to mimic and replace human (job) tasks. The goal is to make problems mathematically computable. The term AI has been expanded through psychology, cognitive sciences, knowledge modelling, and expert systems to include learnable skills and competencies. Developing AI requires expertise in mathematics and computer science, as well as knowledge about the field the AI should be applied to [9]. In general, a distinction should be made between the development and the application of AI [10, p. 16]. The objective of this article is to examine the application and usage of AI in physical and non-physical work processes in companies and not the development of AI.

Despite the increasing presence of AI in discourses on the future of work, there is still relatively little empirically based knowledge about the spread of AI on the labor market and which specific types of companies actually utilize AI technologies. This is accompanied by a lack of knowledge about which company characteristics favor the adoption of AI and why. This research gap limits the understanding of how AI is used in companies as a tool to increase productivity and innovation, as well as how AI might change the company structures and the working conditions of employees.

Against this background, our study focuses on the question: Which company characteristics foster the use of AI? By answering this question, we aim to draw a differentiated picture of AI use in companies and to understand which factors pro-

The BIBB Qualification Panel is a study conducted by the German Federal Institute for Vocational Education and Training (BIBB). BIBB is a German federal institution under public law with legal capacity that is financed from German federal budget funds and is subject to the legal supervision of the Federal Ministry of Education and Research (BMBF).

©2024, PTI

mote or inhibit the integration of AI applications into everyday business life.

The relevance of this research arises from the rapid development and diffusion of AI technologies and their profound impact on the global economy and society. A comprehensive understanding of the operational use of AI in companies is crucial for shaping the future direction of labor market strategies, educational needs and technology policy measures.

Our findings could help to support political and strategic decision-making processes in companies, in order to utilize the potential of AI and to support companies on their way to implement AI applications in the work context while considering ethical considerations.

II. State of Research And Theoretical Background

The discussion (and the research) on the use of AI applications in the world of work has increased in recent years, particularly in light of the rapid developments in the field of machine learning and the associated potential for changing work processes. However, the discussion about AI is by no means new and dates back to the 1950s. Early research primarily focused on the technological foundations and theoretical possibilities of AI [11]. Over the years, interest has as well shifted to the practical implications of AI applications for the world of work, particularly in the context of digitalization [1], [3].

A major focus of current research, alongside the innovation of AI, is the substitutability of human labor by machine systems. Previous studies suggested that extensive displacement of human labor by technology is unlikely in the near future, even in advanced economies such as Germany [6], [12], [13]. However, the public discussion has taken a new direction, especially since the introduction of more advanced AI systems such as ChatGPT-3.5 at the end of 2022. Current debates emphasize that AI is increasingly able to take on more complex cognitive tasks, which were previously considered less automatable or substitutable [2], [14].

Empirical research on the spread of AI in companies shows that although the usage of AI applications is increasing, it remains overall at a relatively low level. According to the Mannheim Innovation Panel, the use of AI applications in manufacturing and business-related services was approximately 6 percent in 2019 [15]. Initial analyses from the BIBB Training Panel showed that only 3 to 4 percent of all companies in Germany used AI applications in 2019 and 2020 [3].

An employee survey from 2019 found that the non-use of AI applications in the workplace is declining, but that approximately 90 percent of employees still do not use AI applications or use them rarely [3].

The theoretical foundation of our analysis on the use of AI applications in German companies is based on the interplay of organization-specific characteristics and their reactions to technological innovations.

David Autor's and colleagues [5] research on the impact of technological change on the labor market, known as the task-based approach, which was later transformed to the concept of Routine-Biased Technological Change (RBTC) [16]. RBTC is a theory that assumes that technological innovations usually are able to replace programmable tasks which are referred to as routine and complement more analytic tasks. Manual tasks usually are not heavily affected by technologies.

According to the RBTC, this means that for companies, the decision to implement AI applications depends not only on the availability of the technology, but also on the types of tasks that exist in the company. Companies in sectors that are heavily characterized by routine tasks may be more inclined to use AI application to substitute certain tasks, while companies in sectors that require complex decision-making and human interaction may be more likely to use AI in a complementary way. Companies with more manual tasks might be less likely to use AI applications [2]. Moreover, advancements in AI have redefined the boundaries of replaceable tasks. So, as well certain complex and analytical tasks might be substitutable by AI.

In the following, on the basis of organizational theory and innovation economics [17] we developed hypotheses to explain the differences in the use of AI in companies according to certain company characteristics.

1. Types of job tasks

The type of tasks in a company influences the suitability of AI applications. Tasks that require analytical thinking and decision-making could be complemented and enhanced from AI, whereas routine tasks might be replaced by AI. Companies with a high degree of manual task should use AI less frequently. Therefore, we assume:

> **Hypothesis 1:** Companies whose activities require a high level of analytical and communication tasks as well as those with many routine tasks, use (or plan to use) AI more frequently than companies whose activities are mainly manual tasks.

2. Level of innovation

Companies that cultivate a culture of innovation and regularly introduce new products and services may be more willing to adopt AI applications. Thus:

> **Hypothesis 2:** Companies with a high degree of innovation are more inclined to use AI or plan to introduce it.

3. Level of digitalization

A higher degree of digitalization may indicate a greater willingness and ability to integrate AI applications, as existing digital systems can be more easily supplemented with new technologies. Therefore:

> **Hypothesis 3:** Companies with a higher level of digitalization are more likely to use or plan to use AI.

The importance of tasks and the potential for innovation and the level of digitalization must also always be considered in the context of economic resources and the institutional background of organizations. (Sociological) neo-institutionalism emphasizes the role of culture, social norms and education as driving forces behind the structure and behavior of organizations [17]. This approach can be used to explain differences in AI utilization based on cultural norms within different industrial sectors.

4. Company size

Larger companies generally have more resources and a greater capacity to spread risk, which enables them to adopt new technologies such as AI applications more quickly. Thus, we hypothesize:

> **Hypothesis 4:** The larger a company is, the more likely it is to use or plan the use of AI.

5. Industry-specific differences

The applicability and benefits of AI applications vary greatly between different industries, depending on the specific requirements and technological maturity of the industry.

> **Hypothesis 5:** Companies in technology-intensive sectors such as medical services and business-related services use AI (or plan to use AI) more frequently than companies in less technology-intensive sectors such as construction.

6. Chamber membership

Chamber membership can serve as a proxy for industry-specific norms and the degree of formalization of business practices. Companies in more modern and formally organized chambers may be more inclined to adopt new technologies. Additionally, companies affiliated with the chamber of crafts often perform tasks that require manual dexterity and are (so far) less replaceable by AI.

> **Hypothesis 6:** Companies that belong to chambers of crafts are less likely to use AI.

III. Data Basis, Operationalization and Model

A. Data basis

The BIBB Training Panel, short for BIBB Establishment Panel on Training and Competence Development, forms the basis for analyzing the use of AI in German companies. The panel has been conducted annually since 2011 and is representative of all companies in Germany with at least one employee who subjects to social security contributions. It comprises a sample of at least 3,500 companies, with the number of companies surveyed varying between 3,500 and 4,000. The survey focuses on vocational education and training as well as continuing training and, since 2016, increasingly on digitalization in companies. Moreover, each year there are changing focus modules [18], [19]. For the descriptive results we use the waves 2020-2023, the multivariate analysis only focuses on the newest wave from 2023.

B. Operationalization

In order to measure and analyze the use of AI in companies, the variables were operationalized as follows for this study.

Dependent variable: AI utilization

As initially stated, AI is more than just deep learning and pattern recognition. However, our focus is data driven: To remain comprehensible to the broad range of companies in a general establishment survey, we use the term artificial intelligence and supplement it with typical but broad examples. AI utilization was measured using two items from a large item battery focusing on digital technology use in the company: Use of artificial intelligence and machine learning for 1) physical work processes (e.g. deep learning and pattern recognition in production, maintenance, building management or care); and 2) for non-physical work processes (e.g. deep learning and pattern recognition in marketing, procurement or human resources). In our question, we do not limit AI to deep learning and pattern recognition, but rather mention these as examples that are understandable and relevant to many companies.

The items were surveyed with three answer possibilities: 1) No, the technology is not currently being used in operations and there are no plans to purchase it. 2) No, the technology is not currently used in operations, but a purchase is planned. 3) Yes, the technology is currently being used in operations. The answers to both AI items were then combined to measure the overall use of AI in a company. For this we created a variable, with three categories: No AI use, if neither of the AI items are used (0), planed AI use, if at least one AI item is planned and none is used (1), active AI use, if at least one AI item is used (2).

Independent variables

The independent variables for the hypotheses are operationalized as follows:

- **Types of job tasks:** The survey asked about the frequency with which employees engage in job tasks categorized by the skill-level of their jobs: simple, medium, and highly skilled. Specifically, the tasks investigated included tasks that; a) where all details are prescribed, b) where involving repetitive processes down to the minutest details, c) require the use of tools or machinery, d) necessitate manual dexterity and craftsmanship, e) where involving informing or advising customers or patients, f) involve persuading others and negotiating compromises, g) where related to organizing processes or conducting research, h) improve or innovate procedures and processes.

  These tasks of all skill levels were grouped together and were then summarized into three dimensions using factor analysis: Routine tasks (a & b), dexterity (manual tasks; c & d) and communicative-analytical tasks (e, f, g & h).
- **Degree of innovation:** Recorded by asking about the introduction of new or significantly improved products or services in the last three years. The values are summarized in an index that reflects the degree of innovation: 0 = no innovations, 1 = improvements, 2 = new products, 3 = improvements and new products.
- **Level of digitalization**: In terms of digitization, companies were presented with a range of digital technologies that could be used in their operations. These technologies were added together to create a digitalization index ranging from 0 to 11. AI technologies were not included.
- **Company size:** Divided into four size categories: 1 to 19 employees, 20-99 employees, 100-199 employees and 200 and more employees. Employees in this case are only those, who are subjected to social security contributions.
- **Industries:** Eight categories: 1. Primary Sector (Agriculture/Mining/Energy), 2. Manufacturing, 3. Construction, 4. Trade and repair, 5. Business related services, 6. Other personal services, 7. Medical services, 8. Public services

and education. Those are based on the 2-digit NACE Rev. 2 classification but are summarized to only 8 categories.
- **Chamber affiliation:** Four categories, including no chamber (0), chamber of commerce and industry (1), chamber of crafts (2) and other chambers (3).

Additional control variables, which categorize the company environment, were included in the model:
- **Proportion of continuing training participants** Measures the proportion of employees (without apprentices) who have participated in continuing training measures in the past year (2022) financed by the employer (between 0 and 1).
- **Proportion of employees with simple task jobs:** Employees who carry out jobs that do not usually require vocational education and training (between 0-1).
- **Location:** Measures whether a company is located in Eastern (2) or Western Germany (1).
- **Training company:** Measures whether a company offer apprenticeships (1) or not (0).

C. Model

For the statistical analysis, an **ordered logit model**, which analyzes the probability of an ordered response variable, as it is the case for our dependent variable (No AI use (0), planned AI use (1) and active AI use (2)).

The ordered logit model is particularly suitable for analyzing ordinal response categories, as here the order of the categories is meaningful, but no equal distance between the categories is assumed. The model estimates the probability that an observation falls into a particular category or a lower category, given the explanatory variables. The coefficients in the model are interpreted as the change in the log odds of a higher category of the response variable when the explanatory variable is increased by one unit.

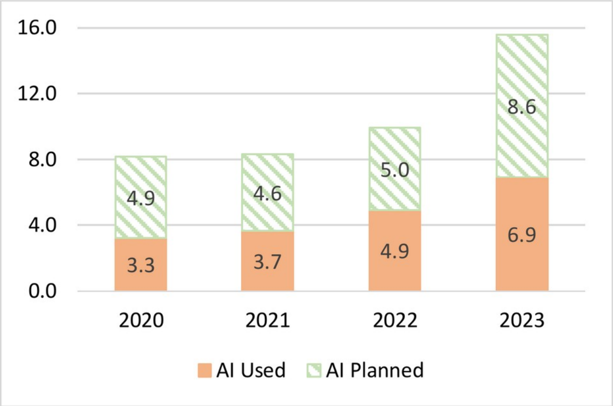
After estimating the ordered logit model, average marginal effects (AMEs) were calculated. These indicate the average change in the probability of the different AI application use categories when an independent variable is changing by one unit. AMEs provide a direct interpretation of the impact of the independent variables on the probability of each response category and are particularly useful for interpreting the results of a non-linear model such as the ordered logit.

IV. RESULTS

D. Descriptive results

The descriptive analysis of the development of AI use in German companies from 2020 to 2023 shows a steady, albeit small, increase in both the actual use and the plans to implement AI technologies (cf. Fig. 1).

In 2020, 3.3% of companies actively used AI, while approximately 5% of companies planned to use AI in the near future. By 2021, the proportion of companies using AI had risen to 3.7%, while the proportion of companies planning to use AI remained roughly constant around 5%. A further increase was observed in 2022, with 4.9% of companies stating that they use AI and still 5.0% planning to do so. A steeper increase in both use and planning can be seen for 2023. 6.9% of companies actually use AI, while now 8.6% state that they are planning to introduce AI. These figures illustrate that the willingness to integrate AI into business processes continues to grow. In the last four years, AI use in companies has doubled.



Source: BIBB Training Panel 2019-2023, weighted data, n_{2020} =4,097, n_{2021} =3,981, n_{2022} =3,527, n_{2023} =3,002.

Fig. 1 Usage and planning of the use of AI by companies 2020 to 2023

This increasing trend in the usage and planning of using AI applications could reflect a growing acceptance and confidence in AI technology, as well as the growing awareness of the benefits AI can offer in various business areas. The data also emphasizes the need to continuously monitor developments in the field of AI application, as rapid increases within four years are possible. Still, one has to keep in mind, that the spread of AI applications in German companies is still low, as only one in fourteen companies is using AI in 2023.

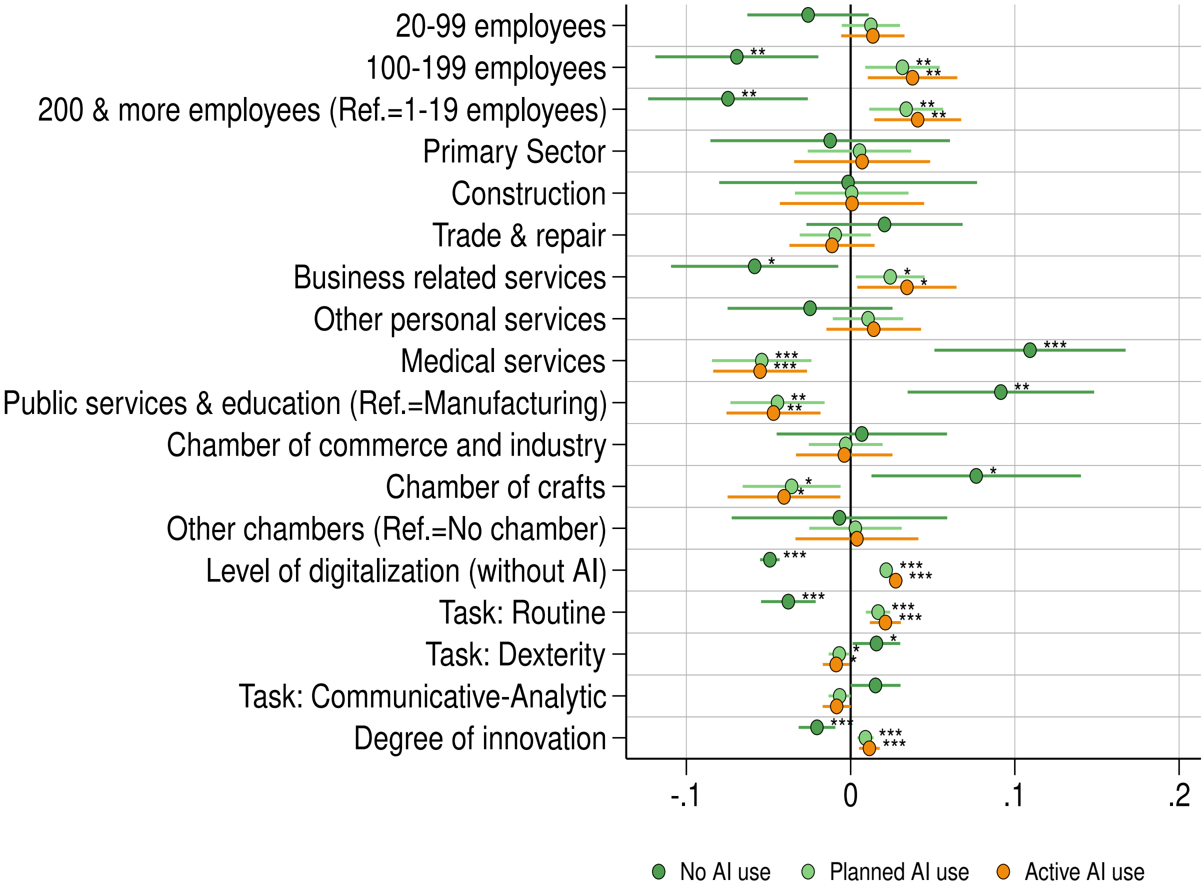
E. Multivariate results

The multivariate analysis of the use of AI applications in German companies in 2023 shows a clear differentiation according to company size, sector, chamber affiliation, level of digitalization, type of tasks and degree of innovation (cf. Fig. 2 and Table I). The analysis divides companies into those that use AI or plan to use AI, which show similar results, and those that do not use AI.

Types of tasks: Companies that require a high level of analytical and communication skills have a lower probability to use or plan to use AI. Though, this is not significant. This indicates that AI is not particularly used in areas where it can help to support complex decision-making processes.

In companies with a high level of routine tasks companies have a significantly higher probability to use and plan to use AI more often. Whether this means that those with routine tasks use AI more often and that AI complements their jobs or that companies with a lot of routine tasks use AI to substitute for such routine tasks must be explored in further research.

However, it appears that companies with a high level of tasks that are highly dexterous tend to have a lower probabil-



Source: BIBB Training Panel 2023, ordered logit model, n=3,002, Pseudo-R² = 0.1425. Average Marginal Effects (AMEs). ***>0.001, ** > 0.01, *>0.05. Also controlled for are: Training company (yes/no), Proportion of continuing training participants, Proportion of employees with simple task jobs and location in West or East Germany. Ref. = Reference Category. The entire model, including control variables, is listed in the appendix.

Fig. 2 AI use in German companies in 2023 according to company structure characteristics

ity of using AI applications. As well their probability of planning to use AI is lower.

Degree of innovation: Companies that have a high degree of innovation have a higher probability of using AI as well as of planning to use AI

Level of digitalization without AI: Companies that have a high level of technology use (level of digitalization without AI) also have a higher probability of using or planning on using AI applications. This could emphasize the importance of an existing digital infrastructure as a basis for the introduction of more advanced technologies such as AI.

Company size: Larger companies (200+ employees and 100-199 employees) show a higher probability to use or to plan the use of AI compared to smaller companies (1-19 employees). This suggests economies of scale effects and larger pools of resources in large companies, that favor the introduction of AI.

Industry-specific differences: Industries such as business-related services have a higher probability to use and plan AI applications in comparison to manufacturing. In contrast, the probability for medical services and public services and education are lower for using or planning to use AI applications. AI. For the other sectors no significant differences arise. This reflects the different digitalization potentials and needs of the sectors.

Chamber membership: Companies that are members of the chamber of crafts have a lower probability to use or plan the use of AI compared to companies without a chamber. This could be due to the more traditional business models and processes in many manual tasks. There are no significant differences for the other chambers.

V. CONCLUSION

Descriptively, our data show that the use of AI in German companies is slowly but steadily increasing. The results of the multivariate analysis demonstrate that company size, sector, level of digitalization and degree of innovation are important predictors of AI use. Moreover, the results emphasize the types of job tasks within a company as fundamental for the introduction of AI.

In summary, the following hypotheses cannot be rejected:

- H2: Companies with a high degree of innovation are more likely to use or plan to use AI.
- H3: A higher degree of digitalization of a company correlates with a higher probability of AI use or the planning of it.
- H4: Larger companies are more likely to use or plan to use AI.
- H6: Companies that are members of chambers of crafts are less likely to use AI.

Against H1 companies with a high intensity of communicative-analytic tasks show a negative albeit non-significant correlation with the probability of using or the planning of using AI. However, in accordance with H1 companies with a high level of routine tasks use or plane to use AI more often. Companies with a high level of dexterity tasks seem to use AI less often. So, H1 only has to be (partly) rejected. Companies, whose activities require a high level of routine tasks use AI more frequently, while companies with a high level of dexterity tasks use less AI.

For H5 we see an ambiguous picture. Most sectors are not significantly different in their AI use in comparison to manufacturing. In comparison to manufacturing, public services and education have a lower probability of AI use or planning of it, while business related services have a higher probability of AI use or planning of it. These results fit to the hypothesis that companies in technology-intensive sectors use (or plan to use) AI more frequently than those in less technology-intensive sectors.

However, as well medical services show a lower probability of AI use ore planning of it in comparison to manufacturing. This seems to be against H5 as certain branches of the medicine field seem to be technology-intense. An explanation might be the heterogeneity of this sector, as it as well incorporates smaller medical practices (e.g. general practitioners) and nursing services. Also, the results just indicate lower probability in comparison to manufacturing and not a generally low use. Still, H5 cannot be fully accepted.

Our findings build on the existing state of research and expand our understanding of where AI is being used in German companies. Previous quantitative studies for Germany have shown that the adoption of AI in companies is progressing but remains at a relatively low level [3], [15]. Our research supports these findings and provides detailed insights into the company characteristics that correlate with AI use, emphasizing in particular the role of job tasks next to structural company characteristics. Further research should explore those finding more in depth and as well could focus more on the (adaption) processes within the companies with quantitative as well as qualitative data (e.g. in-depth interviews, observations).

A limitation of the study is the restriction to quantitatively recordable data at the company level, which does not consider the subjective perceptions and attitudes of the interviewed decision-makers as well as their knowledge about the usage of AI in all company areas or by all employees (e.g. usage of AI software by employees without official introduction by the company). This might lead to respondent biases and potential inconsistencies in the reporting across different companies. A further limitation is that the measurement of AI is somewhat approximate and may not be fully comprehensible to all interviewees, nor does it encompass all applications that fall within the definition of AI. In addition, the dynamics of the AI market are so fast that the data can quickly become outdated, which limits the generalizability of the results.

The results of this study offer starting points for future research that could deal with the implementation of AI in specific industrial contexts or with the effects of AI on the quality of work. As well, these results could be mirrored with qualitative data, to gain deeper insights in the workplace use of AI and the factors affecting AI use.

For policymakers, the findings can provide a basis for formulating guidelines that could promote a broader and more effective use of AI in the German economy, within ethical limits. The application of AI should be used to improve human working conditions and enrich their job tasks and not lead to displacement of jobs or worsening working conditions (e.g. more routine tasks, surveillance, clock or click work). Moreover, privacy and security concerns should be acknowledged in this regard and AI should not be used for extensive surveillance of the employees. Furthermore, it is not uncommon for biases and discriminatory patterns to be embedded in the training data. It is imperative that such biases are identified and subsequently avoided, as they have the potential to influence crucial decisions such as hiring or performance assessments.

The application of AI in the workplace and in society gives rise to a number of further ethical concerns. These include the need for transparency and accountability in AI systems, which may be perceived as opaque and unaccountable (i.e. AI as a black box). There is also a need to establish ownership of AI-created work, and to consider the potential for manipulation and misinformation through the use of AI (i.e. deep fakes).

In practice, companies can use the results of this analysis to enhance their strategic planning with regard to the introduction of AI technologies.

In the context of digital transformation and its impact on society and the economy, the findings of this study emphasize the necessity to proactively shape technological change and complement it with tailored education and labor market strategies in order to fully capitalize on the benefits of AI while minimizing potential risks for employees and society as a whole.

REFERENCES

[1] D. Acemoglu und P. Restrepo, “Artificial intelligence, automation, and work. An agenda.”, *The Economics of Artificial Intelligence*, S. 197–233, 2019, doi: 10.7208/chicago/9780226613475.003.0008.

[2] D. Acemoglu und P. Restrepo, “The wrong kind of AI? Artificial intelligence and the future of labour demand”, *Cambridge J Re-*

APPENDIX

TABLE I:
RESULTS OF THE ORDINAL REGRESSION

	No AI use	Planned AI use	Active AI use
1 - 19 employees	Ref.	Ref.	Ref.
20 - 99 employees	-0.026	0.012	0.014
100 - 199 employees	-0.069**	0.032**	0.038**
200 & more employees	-0.075**	0.034**	0.041**
Primary Sector	-0.012	0.005	0.007
Manufacturing	Ref.	Ref.	Ref.
Construction	-0.002	0.001	0.001
Trade & repair	0.021	-0.009	-0.011
Business related services	-0.058*	0.024*	0.034*
Other personal services	-0.025	0.011	0.014
Medical services	0.109***	-0.054***	-0.055***
Public services & education	0.091**	-0.045**	-0.047**
No chamber	Ref.	Ref.	Ref.
Chamber of commerce and industry	0.007	-0.003	-0.004
Chamber of crafts	0.076*	-0.036*	-0.041*
Other chambers	-0.007	0.003	0.004
Level of digitalization (without AI)	-0.049***	0.022***	0.027***
Task: Routine	-0.038***	0.017***	0.021***
Task: Dexterity	0.016*	-0.007*	-0.009*
Task: Communicative-Analytic	0.015	-0.007	-0.008
Degree of innovation	-0.020***	0.009***	0.011***
Proportion of continuing training participants	0.036	-0.016	-0.02
Proportion of employees with simple task jobs	-0.010	0.004	0.006
Location: Eastern Germany	-0.008	0.003	0.004
Training company	0.011	-0.005	-0.006
N	3,002	3,002	3,002
R²	0.1425	0.1425	0.1425

Notes: Ordered logit model. n = 3,002. Pseudo-R² = 0.1425. Average Marginal Effects (AMEs). ***>0.001, ** > 0.01, *>0.05.
Ref. = Reference Category. Source: BIBB Training Panel 2023

gions Econ Soc, Jg. 13, Nr. 1, S. 25–35, 2020, doi: 10.1093/cjres/rsz022.

[3] U. Sevindik, “Verbreitung und Einsatz von Künstlicher Intelligenz in Deutschland – Auswirkungen auf berufliche Anforderungen und Strukturen”, Bundesinstitut für Berufsbildung, Bonn, 2022. [Online]. Verfügbar unter: https://res.bibb.de/vet-repository_780476.

[4] D. Acemoglu, D. Autor, J. Hazell und P. Restrepo, “Artificial intelligence and jobs: Evidence from online vacancies”, *Journal of Labor Economics*, Jg. 40, S1, S293-S340, 2022, doi: 10.1086/718327.

[5] D. H. Autor, F. Levy und R. J. Murnane, “The skill content of recent technological change: An empirical exploration”, *The Quarterly Journal of Economics*, Jg. 118, Nr. 4, S. 1279–1333, 2003.

[6] K. Dengler und B. Matthes, “Folgen des technologischen Wandels für den Arbeitsmarkt: Auch komplexere Tätigkeiten könnten zunehmend automatisiert werden”, IAB, IAB-Kurzbericht 13, 2021.

[7] H. Wußing, *6000 Jahre Mathematik. Eine kulturgeschichtliche Zeitreise – 2. Von Euler bis zur Gegenwart*. Heidelberg, 2009.

[8] S. J. Russell und P. Norvig, *Artificial intelligence. A modern approach*. Upper Saddle River, 2009.

[9] J. Dörpinghaus und M. Tiemann, “Künstliche Intelligenz – aktueller Treiber der Transformation?”, BIBB, Bonn, Datenreport zum Berufsbildungsbericht 2024, 2024.

[10] W. Ertel, *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung,* 4. Aufl. Wiesbaden, 2016.

[11] P. Buxmann und H. Schmidt, *Künstliche Intelligenz Mit Algorithmen zum wirtschaftlichen Erfolg: Mit Algorithmen zum wirtschaftlichen Erfolg*, 2019.

[12] R. Helmrich *et al.,* “Digitalisierung der Arbeitslandschaften. Keine Polarisierung der Arbeitswelt, aber beschleunigter Strukturwandel und Arbeitsplatzwechsel”, BIBB, Bonn, Wissenschaftliche Diskussionspaper, 2016.

[13] C. Schneemann *et al.,* “Aktualisierte BMAS-Prognose „Digitalisierte Arbeitswelt“”, Forschungsbericht 526/3, 2021.

[14] M. Webb, “The Impact of Artificial Intelligence on the Labor Market”, Stanford University, 2020.

[15] BMWi, “Einsatz von Künstlicher Intelligenz in der Deutschen Wirtschaft: Einsatz von Künstlicher Stand der KI-Nutzung im Jahr 2019”, Berlin, 2020.

[16] D. Acemoglu und D. Autor, *Skills, tasks and technologies: Implications for employment and earnings*. Elsevier, 2011.

[17] Agnes Dietzen, Christian Gerhards, Mortimer Schlieker und Klaus Troltsch, *Handlungslogiken in der betrieblichen Qualifikationsbedarfsdeckung: Entwicklung eines organisations- und institutionssoziologischen Theorierahmens und empirische Exploration von Einflussfaktoren*, 2023. [Online]. Verfügbar unter: https://res.bibb.de/vet-repository_781520

[18] A. Friedrich, C. Gerhards, S. Mohr, K. Troltsch und K. Weis, “BIBB Training Panel – An Establishment Panel on Training and Competence Development 2011 to 2021 long. GWA_1.0”, Research Data Center at BIBB (ed., data access); Federal Institute for Vocational Education and Training, Bonn, 2023.

[19] A. Friedrich und F. Lukowski, “BIBB Establishment Panel on Training and Competence Development. The longitudinal data set”, *SozW*, Jg. 74, Nr. 2, S. 273–293, 2023, doi: 10.5771/0038-6073-2023-2-273.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 53–59
DOI: 10.15439/2024F1407
ISSN 2300-5963 ACSIS, Vol. 41

Gradient Boosting Trees and Large Language Models for Tabular Data Few-Shot Learning

Carlos Huertas
Amazon Research
Email: carlohue@amazon.com

***Abstract*—Large Language Models (LLM) have brought numerous of new applications to Machine Learning (ML). In the context of tabular data (TD), recent studies show that TabLLM is a very powerful mechanism for few-shot-learning (FSL) applications, even if gradient boosting decisions trees (GBDT) have historically dominated the TD field. In this work we demonstrate that although LLMs are a viable alternative, the evidence suggests that baselines used to gauge performance can be improved. We replicated public benchmarks and our methodology improves LightGBM by 290%, this is mainly driven by forcing node splitting with few samples, a critical step in FSL with GBDT. Our results show an advantage to TabLLM for 8 or fewer shots, but as the number of samples increases GBDT provides competitive performance at a fraction of runtime. For other real-life applications with vast number of samples, we found FSL still useful to improve model diversity, and when combined with ExtraTrees it provides strong resilience to overfitting, our proposal was validated in a ML competition setting ranking first place.**

I. INTRODUCTION

TABULAR data in real-world applications is the most common type of data [1], this continues to be true since relational databases are still pretty common in all sort of domains from social to natural sciences [2]–[6]. Deep Learning (DL), or in general, Neural Network based architectures have shown tremendous potential in tasks like Natural Language Processing (NLP) with developments like transformers [7] and large-scale pre-trained models like DeBERTa [8] have pushed the state-of-the-art (SOTA) and gave DL a top spot in performance. The same can be observed for Computer Vision (CV) with developments like convolutional neural networks (CNN) opening the door for more advanced designs like EfficientNets [9] and more recently Vision Transformers (ViT) have found their way into CV as well [10] with Next-ViT [11] aiming to bridge the gap that still separates ViT from CNN in terms of efficiency in the latency/accuracy trade-off.

Despite all the success from DL, tabular data continues to be omnipresent [12], [13], and to the best of our knowledge, we have not found a consistent DL-based approach that can outperform Gradient Boosted Decision Trees (GBDT) [14]–[16] over a *wide variety* of tasks and conditions, even though it is possible to find specific niche setups where this happens [17]–[19].

Recently, the introduction of Large Language Models (LLM) [20] demonstrated a whole new level of performance for several tasks [21], [22], from traditional NLP to even code generation [23]. The concept of revisiting the qualities of DL-based techniques, in particular LLM for tabular data surged again [6], due to some of the key properties over GBDT [24], such as: representation learning, sequential processing and generalization. Even though DL provides some advantages, if maximum performance is desired, GBDT continues to be the SOTA [25] even with amazing advances in DL, some of the most notable attempts to outperform GBDT with DL methods include: Wide&Deep [26], DeepFM [27], SDTR [28], DeepGBM [29], TabNN [30], BGNN [31], TabNet [32], TransTab [33], TabTransformer [34], SAINT [35] and NPT [36], none of them providing enough evidence to actually be able to beat GBDT over a wide variety of tasks, most of the time, it has been demonstrated the claimed improvements are only present in very specific cases or datasets [17].

There are however, some situations where LLM based solutions seem to have an edge [6], this is when data is limited, and LLM have the capacity to perform both zero-shot (ZSL) and few-shot learning (FSL) [37]. While there is no doubt current SOTA in GBDT will show random-performance for zero-shot learning, recent studies [38] show that even under a few-shot schema, LLM can outperform Xgboost [14], one of the most popular GBDT algorithms.

In this work, we will further explore the performance of GBDT under a FSL schema in order to provide strong baselines. Since previous studies [17] have demonstrated bias in claims of DL outperforming GBDT in other tasks, we look to enhance experiments to confirm SOTA results in the new trend of results regarding FSL and the superiority of LLM over GBDT.

II. RELATED WORK

The main concept behind ZSL or FSL by definition implies the evaluated classifier has either (a) never seen the data samples before (ZSL), or only a few samples (FSL), however, this can only be proven true if we were to train a model (LLM for the purpose of this research) from scratch. Any sort of pre-trained architecture could, in theory, already seen the dataset, hence showing incredible performance. This particular problem has been studied before [39], where both GPT-3.5 and GPT-4 are proven to have seen common datasets in the past, like *Adult Income* and *FICO* [40], in some cases, even proven LLM have literally memorized the datasets verbatim [41] as samples can be extracted out. With this in mind, the fair

TABLE I
GPT-3/4 VS TRADITIONAL ALGORITHMS FOR FEW-SHOT-LEARNING PERFORMANCE (AUC)

Algorithm	**Kaggle Titanic**	**OpenML Diabetes**	**Adult Income**	**FICO**	**Spaceship Titanic**	**Pneumonia**
GPT-4	**0.98**	0.75	0.82	0.68	0.69	0.81
GPT-3.5	0.82	0.74	0.79	0.65	0.63	0.54
GBDT (Xgboost)	0.84	0.75	**0.87**	**0.72**	**0.80**	**0.90**
Logistic Regression	0.79	**0.78**	0.85	**0.72**	0.77	**0.90**

evaluation of LLM vs GBDT under a truly FSL schema is very challenging, while we can guarantee GBDT has never seen the data, the same cannot be said for many LLM applications.

The results from Bordt et al. [39], using a 20-shot-learning are shown in Table I, in this work authors study LLM memorization.

Although LLM results are far from bad, the performance still shows gaps to match GBDT. On top of this, GBDT is a much simpler and faster model, essentially being a more efficient and more powerful option. For the Kaggle Titanic dataset, the power of GPT-4 might look impressive, until authors have proven this is due to memorization and not any particular useful learning. This problem is not particular to tabular data, as LLM have been proved to do so as well for other domains [42]. Nonetheless, authors have found that there is some learning happening, for datasets with no memorization LLM can still provide some performance, especially in very few shot-learning, which leads to the work of Hegselmann et al. [43], where LLM are shown to actually outperform GBDT.

In such work, authors present TabLLM, a very innovative solution to use LLMs for few-shot classification on tabular data, in principle, first running a serialization-step, to turn tabular into a natural language representation. An extensive analysis is done to benchmark multiple serialization techniques. Surprisingly, one of the simplest approaches resulted to be very effective, *"Text Template"* is a compact representation of the form: *"The <column name> is <value>"*. This followed by a task-specific prompt, that can later be fined-tuned for FSL.

TabLLM has been benchmarked for both binary and multiclass problems, from datasets identified in key literature for this task [19], [25], [44]. For simplicity, we will focus on the binary tasks, as to ensure all tasks are of the same objective, and metrics are comparable, e.g. AUC. A summary of their benchmarking results is presented in Table II. For full details refer to Table 12, 13 and 14 in [43].

The results show NN-based solutions, both TabPFN [47] and TabLLM [43], substantially outperform LightGBM for FSL, the improved performance by these techniques is such that the minimal delta observed comes in the *Bank* dataset where TabLLM shows an average (over 4 to 64 FSL) advantage of 163% relative improvement $[(0.642 - 0.5)/(0.554 - 0.5)]$ vs the GBDT solution. On the other extreme, the superiority of TabLLM goes further to outperform LightGBM for as much as 745% $[(0.686 - 0.5)/(0.522 - 0.5)]$ for the *Credit-g* experiment.

In the next section, we present our analysis regarding the extreme underperformance from LightGBM, and our recommendations to establish a fair baseline for a FSL application. Increasing its performance to a more competitive level, and hoping this serves as reference for future benchmarks in the field.

III. PROPOSED SOLUTION

The process of FSL might have slightly different interpretations depending on the field, but the core concept remains, the usage of only a few samples to train a model. This concept holds for the tabular data use-case. Knowing this, is imperative to understand how algorithms like LightGBM work in order to build an effective FSL solution. The LightGBM algorithm is a boosting approach using decisions trees (DT) to learn a function from the input space X^s to the gradient space G [15], the splitting criteria is reviewed below.

Given a training set with n i.i.d. instances $\{x_1, ..., x_n\}$, where each x_i is a vector with dimension s in space X^s. For each boosting iteration, the negative gradients of the loss function with respect to the output of the model are denoted as $\{g_1, ..., g_n\}$. The DT model splits each node to maximize information gain, which is measured by the variance *after* splitting. For a training set O on a fixed node, the gain of splitting (V) feature j at point d is defined as:

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\left(\Sigma_{\{x_i\in:x_{ij}\leq d\}}g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\Sigma_{\{x_i\in:x_{ij}>d\}}g_i\right)^2}{n^j_{r|O}(d)}\right) \quad (1)$$

The problem however arises in practice since the optimization is constrained so that the left $n^j_{l|O}(d)$ and right $n^j_{r|O}(d)$ nodes have a minimum sample size. A segment of LightGBM implementation is shown in Algorithm 1.

The *minimum samples per leaf* then becomes a blocker for FSL, causing the algorithm to stall. Unable to perform any split until training samples exceeds the *min_samples_leaf* parameter. Although previous works [43] have explored parameter tuning based on literature recommendations [12], [19], this is not being addressed, and as a result LightGBM shows *random-guess* performance (e.g. 0.5 AUC) in most experiments, since the default value for *min_samples_leaf* is set to 20.

In this work we propose a LightGBM configuration specifically for FSL applications. We identified key parameters needed as shown in Table III.

The most important parameter for FSL is, without a doubt, *min_data_in_leaf*, as otherwise optimization cannot happen.

TABLE II
TabLLM Experiments Results: LightGBM (GBDT) vs NN-Based (AUC)

Dataset	Method	4-shot	8-shot	16-shot	32-shot	64-shot	Average
Bank [44]	LightGBM	0.50	0.50	0.50	0.50	0.77	0.554
	TabPFN	0.59	0.66	0.69	0.76	0.82	**0.704**
	TabLLM	0.59	0.64	0.65	0.64	0.69	0.642
Blood [44]	LightGBM	0.50	0.50	0.50	0.50	0.69	0.538
	TabPFN	0.52	0.64	0.67	0.70	0.73	**0.652**
	TabLLM	0.58	0.66	0.66	0.68	0.68	0.652
Credit-g [44]	LightGBM	0.50	0.50	0.50	0.50	0.61	0.522
	TabPFN	0.58	0.59	0.64	0.69	0.70	0.640
	TabLLM	0.69	0.66	0.66	0.72	0.70	**0.686**
Diabetes [45]	LightGBM	0.50	0.50	0.50	0.50	0.79	0.558
	TabPFN	0.61	0.67	0.71	0.77	0.82	**0.716**
	TabLLM	0.61	0.63	0.69	0.68	0.73	0.668
Heart [46]	LightGBM	0.50	0.50	0.50	0.50	0.91	0.582
	TabPFN	0.84	0.88	0.87	0.91	0.92	**0.884**
	TabLLM	0.76	0.83	0.87	0.87	0.91	0.848
Income [12]	LightGBM	0.50	0.50	0.50	0.50	0.78	0.556
	TabPFN	0.73	0.71	0.76	0.80	0.82	0.764
	TabLLM	0.84	0.84	0.84	0.84	0.84	**0.840**

TABLE III
Proposed Parameters for FSL Applications in LightGBM

Parameter	Description	Default	Recommended
extra_trees	use extremely randomized trees	false	True
num_leaves	max number of leaves in one tree	31	4
eta	shrinkage rate	0.1	0.05
min_data_in_leaf	**minimal number of data in one leaf**	**20**	**1**
feature_fraction	subset of features on each tree	1.0	0.5
bagging_fraction	select part of data without resampling	1.0	0.5
bagging_freq	frequency for bagging	0	1
min_data_per_group	number of data per categorical group	100	1
cat_l2	L2 regularization in categorical split	10	0
cat_smooth	reduce noise-effect in categoricals	10	0
max_cat_to_onehot	one-vs-other algorithm control	4	100
min_data_in_bin	minimal number of data inside one bin	3	3

The same concept applies to any other parameter that relies on counting of samples, such as *min_data_per_group*. In general, it is required to minimize the restrictions here, this is however a very bad practice for Non-FSL applications, leading to overfitting, and should be used with care in any other types of problems.

Due to the partition mechanism of DT, small sample-size will generate a very constrained histogram, and a greedy partition threshold is not desirable, to enhance this, the usage of extremely randomized trees is required to ensure partition splits are over represented in the tree structure.

In the next section we provide experimental results to demonstrate the ability of LightGBM to do few-shot learning.

IV. Experiments

Our experiment design covers two folds. First, we replicate previous work [43], but apply our recommended methodology to enable efficient FSL for LightGBM. Second, we bring a practical application to incorporate FSL into larger-scale data, this serves as reference that even if samples are vast, FSL can provide benefits.

A. TabLLM Experiment Replication

Both TabPFN and TabLLM show similar performance in average. Only a marginal improvement of 1% in favor of TabPFN, however, both of those solutions outperform LightGBM over 343% in average, with extreme cases such as Credit-g where the relative performance of TabLLM is 745% better. While we were able to validate these numbers are correct, our results show this extreme underperformance is driven due to incorrect parameters.

We have replicated the binary problems. For the sake of simplicity, our LightGBM does not include hyperparameter tuning and instead executed with our fixed recommended parameters as shown in Table III. This leads to intentional underoptimization to disregard the effect of better tuning in the results. We found LightGBM much more competitive as seen in Table IV.

Algorithm 1 LightGBM: feature_histogram Implementation

```
is_splittable_ = false;

//...
const auto grad = GET_GRAD(data_, t);
const auto hess = GET_HESS(data_, t);

sum_left_gradient += grad;
sum_left_hessian += hess;

left_count += cnt;

if (left_count < min_data_in_leaf) {
        continue;
}

right_count = num_data - left_count;
if (right_count < min_data_in_leaf) {
        break;
}
//...

is_splittable_ = true;
```

Our methodology improved the performance of LightGBM by 290%, essentially reducing both TabLLM and TabPFN claimed advantage by 84.5%.

LightGBM can outperform or meet TabLLM for 64-shot performance in 5 out of 6 datasets, only missing for Income dataset, where TabLLM performance is constant regardless the number of shots. This is an interesting problem to review for memorization.

For extreme low FSL, like 4 and 8 shot, we found LightGBM to be competitive, yet falling generally behind, this can further be improved with parameter tuning, but gaps are large to close still. Over 16-shots there is considerable performance parity and as the shots increase LightGBM consistently starts to take over. When enough samples are available, no performance advantages were found from TabLLM or TabPFN, yet both solutions are considerably slower to LightGBM.

B. *FedCSIS 2024 Data Science Challenge*

To further review performance and applications of FSL, we applied our findings to the FedCSIS 2024 Data Science Challenge hosted in the KnowledgePit platform, a web system designed for ML competitions helping to bring collaboration between industry and academia [48].

The challenge: *Predicting Stock Trends*, provides stock-tickers and their performance as measured by 116 financial-markers, such as: *Dividend Payout Ratio*, *Gross Profit Margin*, and *Price to Total Revenue per Share*. The information is provided for current Trailing Twelve Months (TTM), these are static features, named *I1* to *I58*. Another set, known as relative-features, named *dI1* to *dI58* provide the relative 1-yr change for such indicators.

This is a competition event that promotes an objective evaluation of performance. Participants were asked to predict the optimal investment strategy of securities among 3 actions: *buy*, *hold* or *sell*. An in-depth review of the competition is detailed in [49].

Initial Model: In order to establish a baseline we started our simplest possible solution directly with DT, this due to its usual superiority over other algorithms for tabular data that has not been deeply feature engineered [50]. A LightGBM regression model using all features as-is and the original discrete target *"Class"* achieves 0.8439 mean absolute error (MAE). The first insight came from feature importance, which suggests the relative (*dI**) variables far dominate the static set (*I**) as seen in Table V, taking 4 out of the top 5 spots. This inspired further review to enhance generalization given the limited data size.

Sample and Feature Selection: Following Occam's Razor principle, we challenged the value of the static features (*I**). When using all variables it's possible to get 0.6018 AUC, an alternate variant for diversification would be to use relative-features (*dI**) only, this proves to be quite competitive, retaining 95% of predictive power (0.5963 AUC), with a 50% reduction of features. This is important since the large feature mismatch promotes orthogonal decisions boundary for subsequent ensembling techniques.

Another diversification technique comes from instance sampling. We studied the sample-size vs performance in the same binary case to determine the right number of shots to use, ideally the smaller the better for diversification in further stages. Results are provided in Table VI where we can observe even after a 40% sample size reduction (6864 to 4118) there is zero impact in performance, and reducing further brings minimal degradation, this provides an ideal framework for FSL, as the ability to use few samples allows for stacking level-0 models with non-overlapping samples.

Stacking Level-0 Models: Based on previous insights, we determined that FSL is a viable strategy to enable multiple orthogonal models. Although previous analysis was done in a binary setting, these new models are built with the *Perform* target in the dataset. Unlike the discrete buy/hold/sell, this continuous representation allows the model to understand the impact of each action, e.g. not all *"buys"* are equal, since they provide different levels of financial gain/loss. Using a 3k shot-approach per model we forced diversification in the sample space. In order to improve generalization, we used the learnings that ExtraTrees outperforms GBDT in most FSL settings. We did not create any feature engineering, but our *Base Feature Set* is a concatenation of existing features over multiple years for stock-tickers that are present more than once in the dataset, only relative features (*dI**) are used. The details of each model and their respective performance is shown in Table VII. Note that because we switched to *Perform* as target, MAE is no longer optimal, so we optimized for the mean squared error (MSE) instead.

Final Blend: Our Level-1 Meta model is fed with the five different L0 configurations. *MLPRegressor* from *sklearn* was selected for simplicity, architecture is 2 hidden-layers of 10

TABLE IV
UNTUNED LIGHTGBM IMPROVED BASELINE PERFORMANCE (AUC)

Dataset	Method	4-shot	8-shot	16-shot	32-shot	64-shot	Average
Bank	Our LightGBM	0.54	0.62	0.65	0.70	0.77	**0.656**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.77	0.554
Blood	Our LightGBM	0.50	0.63	0.67	0.70	0.71	**0.642**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.69	0.538
Credit-g	Our LightGBM	0.60	0.64	0.62	0.65	0.70	**0.642**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.61	0.522
Diabetes	Our LightGBM	0.50	0.62	0.65	0.71	0.78	**0.652**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.79	0.558
Heart	Our LightGBM	0.78	0.85	0.88	0.90	0.91	**0.864**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.91	0.582
Income	Our LightGBM	0.60	0.68	0.77	0.81	0.83	**0.738**
	LightGBM [43]	0.50	0.50	0.50	0.50	0.78	0.556

TABLE V
COMPETITION: TOP FINANCIAL INDICATORS AS DETERMINED BY LGBM BASELINE MODEL

Feature	Description	Importance
dI58	1-year Absolute Change of Price to Cash Flow from Operations per Share	1.000
I57	Cash Flow from Operations Pct of Capital Expenditures	0.725
dI52	1-year Absolute Change of Cash Ratio	0.675
dI43	1-year Absolute Change of Dividend Yield - Common - Net - Issue - %	0.613
dI56	1-year Absolute Change of Book Value Percentage of Market Capitalization	0.537
I5	Excess Cash Margin - %	0.536
dI57	1-year Absolute Change of Cash Flow from Operations Pct of Capital Expenditures	0.521
Group	Industry sector	0.520
I24	Accounts Receivable Turnover	0.471
dI17	1-year Absolute Change of Debt - Total to EBITDA	0.404
dI44	1-year Absolute Change of PE Growth Ratio	0.377

TABLE VI
COMPETITION: SAMPLE SIZE EFFECT IN PERFORMANCE

Sample Size	AUC
6,864	0.6018
6,178	0.6098
5,491	0.6027
4,118	0.6055
1,373	0.5835
686	0.5887

and 5 neurons with ReLU activation [51]. Optimization is still using *Perform* target, with a 10% validation sample size and adam optimizer [52]. Early stopping is based on R2 score with 64 max epochs.

Since the competition requires discrete actions (buy/hold/sell) instead of expected performance, we optimize the performance-to-action thresholds by ensuring the same action-distribution between train and test. This solution has ranked 1st place in the event, with a MAE score of 0.772, which represents a 3.66% and 7.12% relative improvement against 2nd and 10th place respectively.

V. CONCLUSIONS

When the merit of a proposal is measured by its relative performance to a baseline, the baseline itself is equally, or even more important than the proposal. It is trivial to show a solution is good by simply selecting a weak reference point to compare with. Efforts invested in a new proposal can also be applied to improve a baseline. In this work we have improved LightGBM FSL performance found in literature by 290%. Improvements of this magnitude are unusual with just parameter optimization.

Our results show GBDT can perform few-shot-learning (FSL) with surprising performance with as little as 8-shots. And when data is available, FSL can be used to force diversification between individual models in ensemble or stacking architectures.

While global optimum is too expensive to reach, its imperative to learn the inner caveats of algorithms to exploit their strengths to reasonable levels. Our solution in FedCSIS

TABLE VII
LEVEL-0 MODELS FOR FEDCSIS24: STOCK PREDICTION COMPETITION

Model	**Features**	**Target**	**MSE**
ExtraTrees	Base Feature Set	Original	0.020039
GBDT	Base Feature Set	Original	0.020051
ExtraTrees	Base Feature Set	Quantile(0.5%,99.5%)	0.019609
ExtraTrees	Base with Categorical Removed	Original	0.020088
ExtraTrees	Base with static features added back	Original	0.020055

competition shows the importance of understanding your algorithms to maximize performance, both the FSL approach for diversity and ExtraTrees to fight overfitting proved to be very successful in our experiments to achieve 1st place.

REFERENCES

[1] Ravid Shwartz-Ziv and Amitai Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[2] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang, "Deep & cross network for ad click predictions," 2017.

[3] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar, "Vime: Extending the success of self- and semi-supervised learning to tabular domain," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 11033–11043, Curran Associates, Inc.

[4] Yixuan Zhang, Jialiang Tong, Ziyi Wang, and Fengqiang Gao, "Customer transaction fraud detection using xgboost model," in *2020 International Conference on Computer Engineering and Application (ICCEA)*, 2020, pp. 554–558.

[5] Zifeng Wang and Suzhen Li, "Data-driven risk assessment on urban pipeline network based on a cluster model," *Reliability Engineering & System Safety*, vol. 196, pp. 106781, 2020.

[6] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos, "Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey," 2024.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, "Deberta: Decoding-enhanced BERT with disentangled attention," *CoRR*, vol. abs/2006.03654, 2020.

[9] Mingxing Tan and Quoc V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[11] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan, "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022.

[12] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci, "Deep neural networks and tabular data: A survey," *CoRR*, vol. abs/2110.01889, 2021.

[13] Dugang Liu, Pengxiang Cheng, Hong Zhu, Xing Tang, Yanyu Chen, Xiaoting Wang, Weike Pan, Zhong Ming, and Xiuqiang He, "Diwift: Discovering instance-wise influential features for tabular data," 2022.

[14] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.

[15] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[16] Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev, "Fighting biases with dynamic boosting," *CoRR*, vol. abs/1706.09516, 2017.

[17] Ravid Shwartz-Ziv and Amitai Armon, "Tabular data: Deep learning is not all you need," 2021.

[18] Tomaso Poggio, Andrzej Banburski, and Qianli Liao, "Theoretical issues in deep networks," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30039–30045, 2020.

[19] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 2022.

[20] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian, "A comprehensive overview of large language models," 2024.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[22] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of artificial general intelligence: Early experiments with gpt-4," 2023.

[23] Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao, "Self-planning code generation with large language models," 2023.

[24] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016, http://www.deeplearningbook.org.

[25] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–21, 2024.

[26] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah, "Wide and deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, New York, NY, USA, 2016, DLRS 2016, p. 7–10, Association for Computing Machinery.

[27] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, Xiuqiang He, and Zhenhua Dong, "Deepfm an end-to-end wide and deep learning framework for ctr prediction," 2018.

[28] Haoran Luo, Fan Cheng, Heng Yu, and Yuqi Yi, "Sdtr: Soft decision tree regressor for tabular data," *IEEE Access*, vol. 9, pp. 55999–56011, 2021.

[29] Guolin Ke, Zhenhui Xu, Jia Zhang, Jiang Bian, and Tie-Yan Liu, "Deepgbm: A deep learning framework distilled by gbdt for online prediction tasks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2019, KDD '19, p. 384–394, Association for Computing Machinery.

[30] Guolin Ke, Jia Zhang, Zhenhui Xu, Jiang Bian, and Tie-Yan Liu, "TabNN: A universal neural network solution for tabular data," 2019.

[31] Sergei Ivanov and Liudmila Prokhorenkova, "Boost then convolve: Gradient boosting meets graph neural networks," 2021.

[32] Sercan O. Arik and Tomas Pfister, "Tabnet: Attentive interpretable tabular learning," 2019.

[33] Zifeng Wang and Jimeng Sun, "Transtab: Learning transferable tabular transformers across tables," 2022.

[34] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin, "Tab-transformer: Tabular data modeling using contextual embeddings," 2020.

[35] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein, "Saint: Improved neural networks for tabular data via row attention and contrastive pre-training," 2021.

[36] Jannik Kossen, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal, "Self-attention between datapoints: Going beyond individual input-output pairs in deep learning," 2022.

[37] Omurhan A. Soysal and Mehmet Serdar Guzel, "An introduction to zero-shot learning: An essential review," in *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2020, pp. 1–4.

[38] Ruiyu Wang, Zifeng Wang, and Jimeng Sun, "Unipredict: Large language models are universal tabular classifiers," 2024.

[39] Sebastian Bordt, Harsha Nori, and Rich Caruana, "Elephants never forget: Testing language models for memorization of tabular data," 2024.

[40] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang, "An interpretable model with globally consistent explanations for credit risk," 2018.

[41] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel, "Extracting training data from large language models," 2021.

[42] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang, "Quantifying memorization across neural language models," 2023.

[43] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag, "Tabllm: Few-shot classification of tabular data with large language models," 2023.

[44] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka, "Well-tuned simple nets excel on tabular datasets," 2021.

[45] Jack Smith, J. Everhart, W. Dickson, W. Knowler, and Richard Johannes, "Using the adap learning algorithm to forcast the onset of diabetes mellitus," *Proceedings - Annual Symposium on Computer Applications in Medical Care*, vol. 10, 11 1988.

[46] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano, "Heart Disease," UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C52P4X.

[47] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second," 2023.

[48] Sebastian Stawicki Andrzej Janusz, Dominik Slezak and Mariusz Rosiak, "Data-driven risk assessment on urban pipeline network based on a cluster model," *Proceedings of the 24th International Workshop on Concurrency, Specification and Programming*, 2015.

[49] Ivana T. Dragovic Ana M. Poledica Milica M. Zukanovic Andrzej Janusz Dominik Slezak Aleksandar M. Rakicevic, Pavle D. Milosevic, "Predicting stock trends using common financial indicators: A summary of fedcsis 2024 data science challenge held on knowledgepit.ai platform," *Proceedings of FedCSIS 2024*, 2024.

[50] C. Huertas and Q. Zhao, "On the irrelevance of machine learning algorithms and the importance of relativity," in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Los Alamitos, CA, USA, jul 2023, pp. 16–21, IEEE Computer Society.

[51] Abien Fred Agarap, "Deep learning using rectified linear units (relu)," 2019.

[52] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 61–68
DOI: 10.15439/2024F3164
ISSN 2300-5963 ACSIS, Vol. 41

Towards Evolvable APIs through Ontological Analysis

Nikolas Jíša
0009-0005-1551-2740
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic
Email: jisaniko@fit.cvut.cz

Robert Pergl
0000-0003-2980-4400
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic
Email: robert.pergl@fit.cvut.cz

***Abstract*—In recent times, the growth of technology toward decentralized solutions and microservice architecture has made Application Programming Interfaces (APIs) crucial for connecting different parts of business software systems. Although the technologies for developing and using APIs are quite stable, the fast-changing business world demands that APIs be easy to maintain and adapt. Currently, changes in APIs made by API providers often imply required updates on the side of API consumers, which can be costly and prone to mistakes. This paper analyzes the types of changes in APIs and uses this analysis to build a detailed model that shows the relationships between API consumers and API providers. This model helps visualize these relationships and can serve as a stepping stone for further automation. As a means of possible evolvable realization, we discuss the Normalized System Theory and implementation.**

I. INTRODUCTION

OVER the past few decades, there has been an extraordinary surge in technological progress, particularly evident in the widespread incorporation of the Internet into various aspects of our daily routines, such as social networking, e-commerce and banking [1]. In [2], Kurzweil proposes an extension to Moore's Law to apply the exponential growth of hardware progress to also include software and other technological areas. One such technological area could be the Internet, which incorporates a multitude of distributed systems, where services frequently depend on one another to function effectively. Among these distributed systems, there are client-server applications that can be realized through Application Programming Interfaces (APIs). Naturally, APIs evolve and are subject to changes, which is also acknowledged by Lamonthe et al. in [3]. Changes in an API on the side of the API Provider can imply other necessary changes on the side of API consumer for given API to continue to function as expected. However, keeping up with all the changes of all consumed APIs in a system, to ensure that the expected behavior matches the actual behavior, can be challenging [3]. This further highlights the critical importance of research and inquiries aimed at optimizing API development, especially if Kurzweil's Law of Accelerating Returns is right.

The focus of this paper is the management of changes on the side of the API consumer implied by updates of the consumed API with the goal of introducing more automation and reducing the manual labor needed on the side of API consumer. We delve into an analysis of this problem and then put together categories of conditions that must be satisfied for the consumption of an API to work as expected. These categories of conditions are closely related to changes on the side of API provider, implying changes on the side of API consumer. Finally, we establish an API ontological model for both the API provider side and the API consumer side, with an emphasis put on the API changes. Then use this model to argue about API evolvability challenges and their possible addressing.

The remainder of this paper is structured as follows. Section II describes the research methodology and the research goal. Section III explains relevant terms used in this paper and section IV lists existing related research works. In section V, the relevant changes in the APIs are delineated and the ontological model based on these changes is described. Section VI discusses usability of the model. Section VII evaluates the model using the identified change drivers. Section VIII concludes this paper and also mentions some ideas on follow-up research.

II. RESEARCH OVERVIEW

A. *Research Methodology*

The research presented in this paper adheres to Design Science Research Methodology (DSRM) [4] comprising three interconnected cycles: **1)** Relevance Cycle, which kicks off the research by linking it to real-world needs, outlining what needs to be studied, and setting clear standards for judging the outcomes **2)** Rigor Cycle, which ensures research innovation by assembling existing knowledge as the foundation for the study **3)** Design Cycle, which involves constructing an artifact, evaluating it, and incorporating feedback. This central cycle is based on the other two.

The relevance cycle for this paper is described in section II-B. The rigor cycle is described in section IV. The design cycle and its results are described in section V and their evaluation is given in section VII.

B. *Research Goal*

The goal of this paper is to contribute to machine actionability (i.e. automation) of changes on the side of API consumer

©2024, PTI

Topical area: Software, System and Service Engineering

implied by updates of given consumed API as depicted in Figure 1. The research questions are as follows.

RQ1: What are the changes of APIs on the side of API providers, that might imply necessity of changes on the side of API consumers, for the calling of API methods to continue working in the desired manner?

RQ2: What ontological model could describe APIs from the point of view of these changes?

As an evaluation of the model, we represent an API using the designed API model to illustrate it and visualise the identified changes in the model.

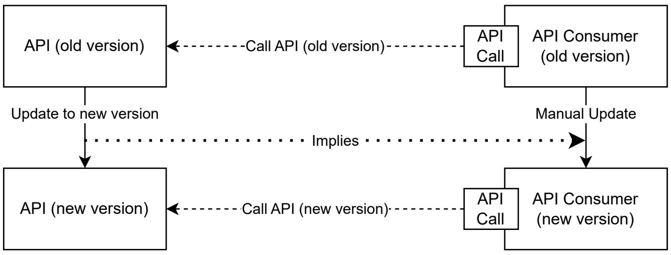


Fig. 1: API update schema

III. Theoretical Background

A. Application Programming Interface (API)

According to Reddy in [5]: "*An Application Programming Interface (API) provides an abstraction for a problem and specifies how clients should interact with software components that implement a solution to that problem.*" In this paper, we distinguish the side of API provider, which is the side that provides functionality in the form of API methods, and the side of API consumer, which is the side that consumes functionality by calls to API methods provided by the side of API provider.

B. Ontology

In this article, the term `ontology` has the meaning of `(Computational) ontology` defined in [6] as follows: "*Computational ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes. An example of such a system can be a company with all its employees and their interrelationships.*"

C. Static Analysis

According to Rival et al. in [7]: "*Static analysis is an automatic technique for program-level analysis that approximates in a conservative manner semantic properties of programs before their execution.*"

D. Normalized Systems Theory

Normalized Systems Theory (NST) [8] is a theory based on fine-grained modularity with goal to make systems more evolvable and stable, specifically by elimination of Combinatorial Effects. Compliance to fine-grained modularity ensures that complex system is broken down into small components that together form the system. NST involves four theorems as follows: **T1**: Separation of Concerns: A processing function can only contain a single task in order to achieve stability **T2**: Action Version Transparency: A processing function that is called by another processing function, needs to exhibit version transparency in order to achieve stability **T3**: Data Version Transparency: A structure that is passed through the interface of a processing function needs to exhibit version transparency in order to achieve stability **T4**: Separation of States: Calling a processing function within another processing function needs to exhibit state keeping in order to achieve stability.

There exists an implementation of NST developed by NSX[1] as described in [9] that has been successfully applied to multiple real-world projects. This implementation introduces five types of elements aligned with basic software concepts as follows:

- Data Elements for data variables and structures
- Task Elements for instructions and functions
- Flow Elements for flows and orchestrations
- Connector Elements allowing input/output commands
- Trigger Elements allowing to setup triggers.

In addition to these elements, and NST Theorems, this implementation is based on code generation via expanders. Expanders are used to create instances of the elements and allow isolation of cross-cutting concerns in most cases; situations that would be difficult to cover with expanders can be handled with custom code.

E. Combinatorial Effect

According to [8], combinatorial effect is characterized by a change whose significance is influenced not only by the nature of the alteration itself, but also by the scale or scope of the system that undergoes the change.

F. FAIR Principles

`FAIR` principles are rules for scientific data management and stewardship, which were established in [10]: "*Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data.*". Each letter of the `FAIR` acronym represents a group of principles:

F: Findability; **F1**: (Meta)data are assigned a globally unique and persistent identifier **F2**: Data are described with rich metadata (defined by **R1** below) **F3**: Metadata clearly and explicitly include the identifier of the data it describes **F4**: (Meta)data are registered or indexed in a searchable resource

A: Accessibility; **A1**: (Meta)data are retrievable by their identifier using a standardized communications protocol **A1.1**: The protocol is open, free, and universally implementable **A1.2**: The protocol allows for an authentication and authorization procedure, where necessary **A2**: Metadata are accessible, even when the data are no longer available

---

[1] https://normalizedsystems.org/about-us/

I: Interoperability; **I1**: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation **I2**: (Meta)data use vocabularies that follow FAIR Principles **I3**: (Meta)data include qualified references to other (meta)data

R: Reusability; **R1**: Meta(data) are richly described with plurality of accurate and relevant attributes **R1.1**: (Meta)data are released with a clear and accessible data usage license **R1.2**: (Meta)data are associated with detailed provenance **R1.3**: (Meta)data meet domain-relevant community standards.

Although originally formulated for scientific data, they have been adopted to generally drive advances in the management of all types of digital objects [11]. In our work, we do not address the full scope of FAIR principles, just the parts important for machine-actionability of API evolvability.

IV. RELATED WORK

First, we reviewed state-of-the-art of API evolvability research, and later we tried to find related ontologies to our desired API model.

A. API Evolvability Related Work

In [3] Lamonthe et al. review API Evolvability literature and, among other things, list open challenges and gaps in the research area. From these open challenges, we identified the following as being the most relevant for our paper (we also included labels from [3]):

1) EC-2 Providing a commercially viable API migration solution
2) EC-10 More tools to help with Web APIs
3) EC-15 Automatically identify factors driving API Changes
4) UC-9 Tools to help API developers deal with API migration, not just users.

For EC-2 and UC-9, there exist approaches that attempt to resolve these challenges. We identified most of the publications mentioned in this paragraph by applying snowballing [12] to [3]. In [13], Brito et al. introduce AppDiff system, which can identify breaking and non-breaking changes between two versions of a Java library based on similarity heuristics and static analysis and in [14], Dagenais et al. introduce SemDiff tool which recommends replacements for framework methods that were accessed by a client program and deleted during the evolution of the framework. However, these tools have not yet provided a commercially viable solution [3]. In [15] Ramos et al. introduce the MELT system, which can extract transformations for the API Consumer side based on the analysis of pull requests on the API Provider side based on static analysis and natural language processing of descriptions in pull requests and comments. In [16] Deshpande et al. address problem of API migration with multi-objective evolutionary algorithms without being limited to scenarios of source method getting transformed always only to single target method (one-to-one mappings), which makes this approach applicable also to scenarios of one or multiple source methods getting mapped to multiple target methods (one-to-many and many-to-many mappings). In [17], Lamothe et al. introduce system A3 for API migration of Android applications based on generation of migration code from code examples. There are also other approaches based on program synthesis that utilize examples of mapping of API calls from one version to another in order to generate transformation procedure. One such example is APIFIX tool introduced in [18], another example is ReFazer tool introduced in [19]. In [20] Beuer-Kellner et al. introduces an API Migration approach based on a service handling conversion of data structures between different versions of APIs. In [21] Huang et al. propose API Mapping approach MATL which leverages transfer learning technique to automatize API mapping without necessarily having knowledge of underlying source code of concerned APIs. One other idea is to have developers on the side of API Providers create transformation scripts for API Consumers to help update API calls from one version of API to another [3]. Similarly to our goal, the mentioned techniques deal with improving API migration. However, our focus is on providing an analytical method based on APIs modeling to identify change drivers. Our approach can then be used in combination with these techniques and other implementation technologies, such as the Normalized Systems we discuss here.

For EC-10, according to [22] Web API providers also control runtime of APIs and can do changes anytime with severe consequences to Web API consumers as opposed to Library APIs. Additionally, Web APIs often lack machine-undrestandable specifications, and data are often passed over strings. Our paper focuses on APIs in general, and the proposed API Model can later be used as a basis for contribution to Web API tooling.

For EC-15, according to Granli et al. in [23], the largest driving force for API changes is the desire for new functionality with changes occurring sporadically rather than continuously, and the Law of Conservation of Organization Stability [24] is not a considerable factor. The case study by Hou et al. in [25] shows the reasons for the changes during the evolution of the AWT[2]/Swing[3] library. The case study by Zarras et al. in [26] shows detection of evolution patterns and regularities based on Lehman's laws of software evolution. In contrast to the studies mentioned above, in this paper, we distinguish changes by the entities concerned.

B. Related Work on API Ontologies

In [27], Karavisileiou proposes a reference ontology based on OpenAPI Specification[4] for Representational State Transfer (REST) services and a procedure to convert OpenAPI Specification description into this ontology. This ontology is somewhat different from what we aim to do here, since it does not place an emphasis on combinatorial effects.

---

[2] https://docs.oracle.com/javase/7/docs/api/java/awt/package-summary.html

[3] https://docs.oracle.com/javase/7/docs/api/javax/swing/package-summary.html

[4] https://spec.openapis.org/oas/latest.html

In [28], Togias proposes a ontology for social network API. Although the ontology is specifically meant for social network APIs rather than APIs in general, multiple parts of the model are also applicable to APIs in general. In our ontological model, we use multiple entities that have similarities in this ontology.

In [29], Androces introduces Platform as a Service (PaaS) ontology. Although this ontology contains some entities that are applicable to our desired API ontological model (such as `Operation` or `API`), most of the entities are specific to the PaaS area.

There are also multiple API description formats, which can also be considered as types of ontologies on their own, because these formats support structured metadata that describe APIs semantically and therefore conform to ontology as defined in section III-B. Examples are OpenAPI Specification[4], API Blueprint [5] and SmartAPI[6]. Although all of these API description formats contain rich structured metadata, they do not focus on API change management.

V. IMPLEMENTATION AND RESULTS

A. Problem Analysis

From what Wilkinson et al. describe in [10], compliance of a system with FAIR Principles helps to introduce machine actionability in general. Therefore, compliance of APIs with FAIR Principles could be a starting point to make changes on the side of the API consumer implied by updates on the side of API provider machine actionable. However, structured rich metadata (from **F2** and **I1**) must contain relevant data for API changes on the side of API provider which impact the side of API consumer. Furthermore, the API version before an update and the API version after the update should have metadata in the same format, so that the metadata after the update can be compared with metadata before the update to get the semantic representation of the API update. This semantic representation of the API update shall then serve as an input for the automation of changes on the side of the API consumer implied by the update of API on the side of API provider.

To make APIs comply with FAIR Principles, generally, a structured API description conforming to an ontology such as one of ontologies mentioned in section IV would suffice. However, since the ontologies mentioned in section IV are not meant for managing API changes on both the API provider side and the API consumer side, we need to create our own ontology. The changes are analyzed in section V-B and based on these changes a ontological API model is created in section V-C.

B. API Change Drivers

Based on analysis of changes in multiple APIs, in order to better understand change drivers (i.e. reasons to change), we put together three categories of conditions that must be satisfied between the call to the API consumer method and the API method itself on the side of the API provider. If any condition in any of these categories changes on either side, changes on the other side may also be required for the API consumer to be able to call the given API method with the desired behavior. The conditions are as follows: **C1**: Correspondence of data transfer settings such as protocol settings and endpoint settings **C2**: Correspondence of API method signature and its meaning (on side of API provider) with API method call and its expected meaning (on side of API consumer) **C3**: Correspondence of API method behavior on side of API provider with expected behavior on side of API consumer. Examples of changes affecting given conditions are:

E1: Change to different communication protocol (affects **C1**)

E2: Change of communication protocol settings such as change of authentication, encryption, encoding, serialization ... (affects **C1**)

E3: Change of signature of API method (affects **C2**):

- Change of API method name or return type
- Creation or deletion of a API method parameter
- Change of order of API method parameters
- Change of API method parameter name or parameter type

E4: Deletion of API method (affects **C2**, **C3**)

E5: Change of meaning of API method return value or parameter values (such as change of expected units from kilometers to meters) (affects **C3**)

E6: Change of API method pre-conditions / post-conditions (affects **C3**)

E7: Change of API method behavior (affects **C3**)

E8: Change of API method mechanism resulting in drastic decrease of performance (affects **C3**)

C. API Ontological Model

We decided to create a new model from scratch inspired by these existing API ontologies rather than using any of the mentioned models (such as, for example, OpenAPI) as a starting point, because we wanted our model to focus on change drivers for modifications on API consumer side implied by modifications on API provider side and also to be more abstract than the existing API ontologies mentioned in section IV-B. We utilized Visual Paradigm Community[7] for the creation of models and diagrams.

1) API Provider Side Model: Based on our analysis of the change drivers in the previous section, we first created the API model of the side of API provider in the UML class diagram, which is shown in Figure 2. We represent the ontology in plain UML notation instead of using a formal ontology modeling framework (such as OntoUML [30]) because given the technical terms involved in modeling APIs, using such frameworks would add unnecessary complexity without providing benefits for our goal. The idea of this model is the following. **1)** `API Provider` provides `APIs` **2)** `API` contains `API Methods` **3)** `API Method` contains `API Method Signature`, `API`

---

[5] https://apiblueprint.org/documentation/specification.html

[6] https://smart-api.info/

[7] https://www.visual-paradigm.com/

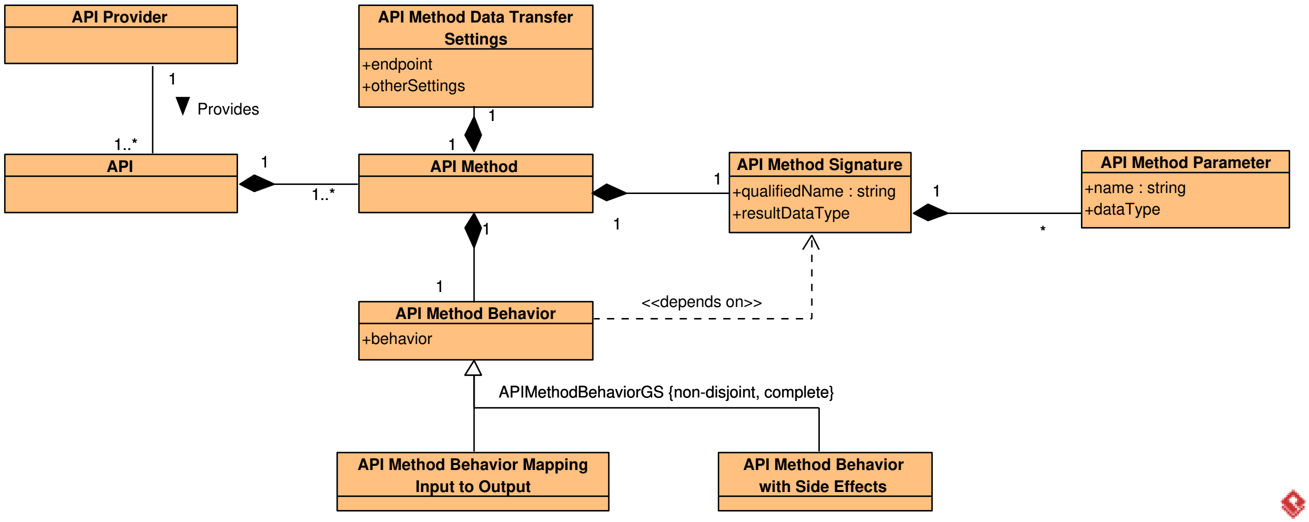


Fig. 2: The suggested API ontological model

`Method Behavior` and `API Method Data Transfer Settings` **4)** `API Method Signature` contains `name`, `resultDataType` and `API Method Parameters` **5)** `API Method Parameter` contains `name` and `dataType` **6)** `API Method Behavior` represents the behavior of the method. It can be `API Method Behavior Mapping Input to Output` and it can also be `API Method Behavior with Side Effects`; this model does not specify exactly how the `behavior` is defined - it could be, for example, natural text description, pseudocode, mapping function, or source code **7)** `API Method Data Transfer Settings` involves the configuration of communication protocol, including serialization settings, authentication settings, encryption settings, endpoint settings, ...; this model does not specify how exactly the `endpoint` and `otherSettings` are defined - `endpoint` could be, for example, defined by Uniform Resource Locator (URL) and `otherSettings` could be defined in another structured object.

2) API Consumer Side Model: Next, we extend the model by incorporating the API consumer side and highlight the dependencies between the API consumer side and the API provider side with red dashed lines in Figure 3. The idea of the entities in this model is the following: **1)** `API` provider side elements are in orange color and have the same meaning as in Figure 2 **2)** `API` consumer side elements are in red color **3)** `API Consumer` represents the entity consuming an `API` **4)** `API Method Call` represents a call of `API Method` by `API Consumer`; it consists of `name` and `API Method Call Parameter Assignments`; it has `API Method Call Data Transfer Settings` and it can also have `API Method Call Result Value Syntactic and Semantic Processing` **5)** `API Method Call Data Transfer Settings` involves configuration of communication protocol, including serialization settings, authentication settings, encryption settings, endpoint settings, ...; this model does not specify how exactly the `endpoint` and `otherSettings` are represented **6)** `API Method Call Parameter Assignment` represent what should be assigned to `API Method Parameter` identified by `name` and `dataType` in `API Method Call`; this model does not specify exactly how the `assignment` is defined - it could be, for example, natural text description, pseudocode, mapping function, or source code **7)** `API Method Call Result Value Syntactic and Semantic Processing` represents processing of the result value of the `API Method Call`; this model does not specify exactly how the `processing` is defined - it could be, for example, natural text description, pseudocode, mapping function, or source code.

As depicted in Figure 3, the API consumer side entities directly depend on the API provider side entities as follows: **1)** `API Method Call` directly depends on `API Method Signature` **2)** `API Method Call Data Transfer Settings` directly depends on `API Method Data Transfer Settings` **3)** `API Method Call Result Value Syntactic and Semantic Processing` directly depends on `API Method Behavior` and `API Method Signature` **4)** `API Method Call Parameter Assignment` directly depends on `API Method Behavior` and `API Method Parameter`.

VI. DISCUSSION

Analysis of potential changes on the side of the API provider that could imply changes on the side of API consumer answers **RQ1** in section V-B and we also pointed out the meaning of these changes and some of their possible sources. For **RQ2**, the answer is the ontological model described in section V-C. The model serves to clarify the ontological aspects

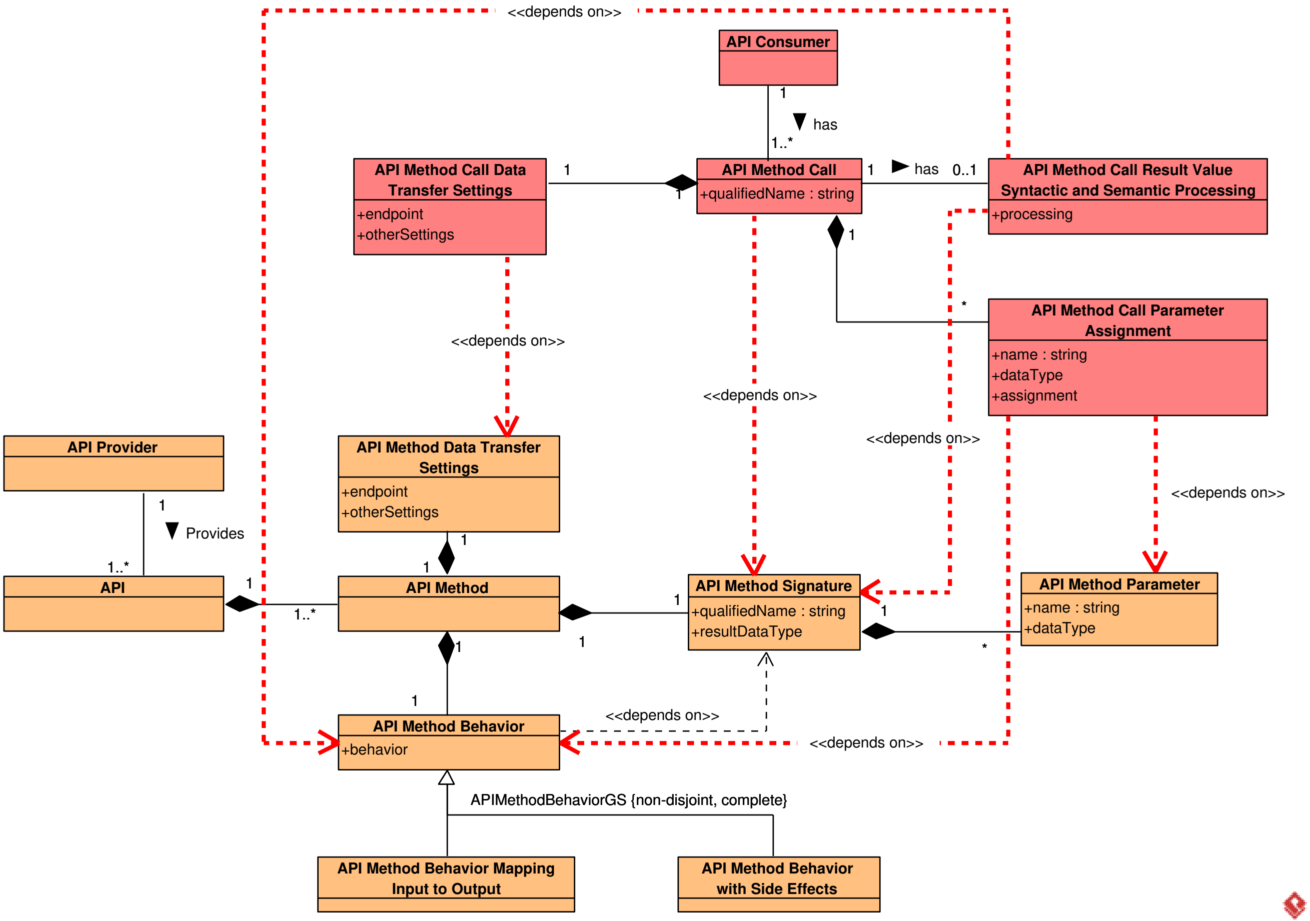


Fig. 3: API ontological model with both provider side and consumer side

of API, API provider, API consumer, and their relationships. We opt for a model that is sufficiently generic to describe most APIs without being limited by any specific domain in context of APIs such as [29] or [28]. Nevertheless, our model focuses on evolvability and change drivers, therefore, some API components (for OpenAPI for example License Object) do not have explicit semantic constructs in our model, and some other API components are defined only in an abstract manner (for OpenAPI for example Server Object, which is abstractly defined as part of API Method Data Transfer Settings in our model). Although the model is relatively simple, it can represent and be used for change analysis of any size of a real-world API (just the number of instances grows). The only current limitation is that it does not cover inter-instance dependencies, i.e. changes of one endpoint causing changes in another one. The identified change drivers could serve as a basis for applying NST, which could help mitigate combinatorial effects that cause a change made to an API system to require the same effort and scope as making the same change to a future evolved version of the API system, even if it were a thousand times larger.

To detect changes on the side of API Consumers implied by changes on the side of API Providers, artifacts such as source code, documentation artifacts, or other artifacts generated from source code or documentation artifacts could be used. One of the ways would be to compare the artifacts for a new API version with the artifacts for the previous API version. This comparison could be automatically activated as soon as a new version of API is detected and the result of this comparison could trigger a notification or even an automated script, which could update API consumption calls automatically or with possibly minimal manual intervention in the form of confirmation.

In this paper, we have considered API on an abstract level and have not covered areas in lower levels of abstraction such as security settings, licensing, or areas of concrete protocols and technologies used with APIs. Also, we have not done detailed investigation of cases when an API called (directly or indirectly) by API provider gets updated and indirectly implies changes to the side of API consumer. We have considered these cases to be the same as direct modifications on the side of API provider implying changes to the side of API consumer. We also have not covered the options concerning implementation beyond suggesting the Normalized Systems,

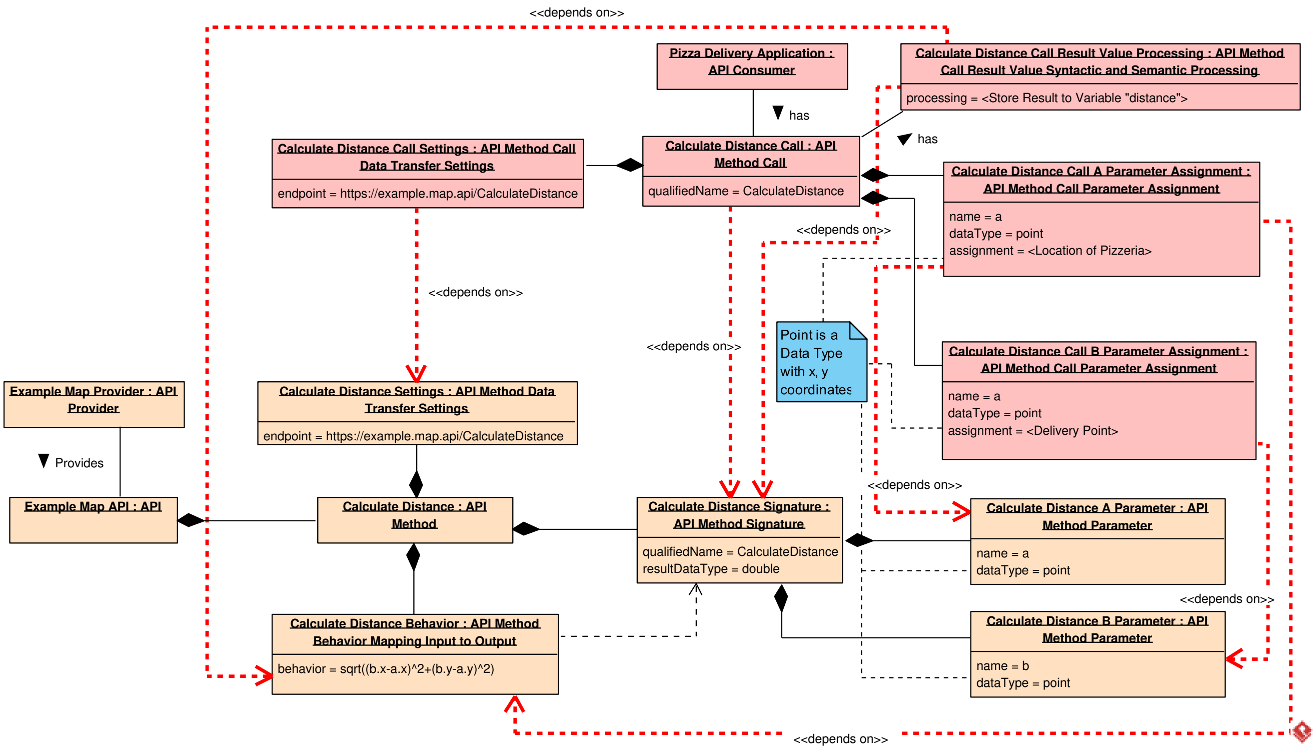

<<depends on>>
Pizza Delivery Application : API Consumer
Calculate Distance Call Result Value Processing : API Method Call Result Value Syntactic and Semantic Processing
processing = <Store Result to Variable "distance">
has
has
Calculate Distance Call Settings : API Method Call Data Transfer Settings
endpoint = https://example.map.api/CalculateDistance
Calculate Distance Call : API Method Call
qualifiedName = CalculateDistance
Calculate Distance Call A Parameter Assignment : API Method Call Parameter Assignment
name = a
dataType = point
assignment = <Location of Pizzeria>
<<depends on>>
<<depends on>>
<<depends on>>
Point is a Data Type with x, y coordinates
Calculate Distance Call B Parameter Assignment : API Method Call Parameter Assignment
name = a
dataType = point
assignment = <Delivery Point>
Example Map Provider : API Provider
Calculate Distance Settings : API Method Data Transfer Settings
endpoint = https://example.map.api/CalculateDistance
Provides
Example Map API : API
Calculate Distance : API Method
Calculate Distance Signature : API Method Signature
qualifiedName = CalculateDistance
resultDataType = double
<<depends on>>
Calculate Distance A Parameter : API Method Parameter
name = a
dataType = point
<<depends on>>
Calculate Distance Behavior : API Method Behavior Mapping Input to Output
behavior = sqrt((b.x-a.x)^2+(b.y-a.y)^2)
Calculate Distance B Parameter : API Method Parameter
name = b
dataType = point
<<depends on>>

Fig. 4: Example of an API representation in API Ontological Model

which would be out of the scope of this paper.

VII. EVALUATION

To verify that our model is applicable and that it can expicitly vizualize the relevant change drivers, we represent an example API with our model and later demonstrate the readability of the relevant change drivers.

Let us consider an `Example Map API` which has method `CalculateDistance` accepting parameters `a` and `b` which are of type `Point`, which has members `x` and `y` of type `double`, and the return value is of type `double`. The `Point` type could be defined in programming language C as `struct Point { double x; double y; }` and the `CalculateDistance` method would have signature `double CalculateDistance(struct Point a, struct Point b)`. Let us also consider an application for pizza delivery that consumes the API. Both the API consumer side and the API provider side could be represented in API Ontological Model as demonstrated in the UML Object Diagram[8] in Figure 4 which can be considered an instance of the model in Figure 3.

The model makes change drivers explicit to see, which we demonstrate on examples listed in section V-B as follows:

- Changes of `API Method Data Transfer Settings` in **E1** and **E2** are changes of `API Method Data Transfer Settings` on the side of API provider and imply changes of `API Method Call Data Transfer Settings` on the side of API consumer.
- Changes of `API Method Signature` in **E3** and **E4** signify changes of `API Method Signature` and `API Method Parameter` on the side of API provider and imply changes of `API Method Call`, `API Method Call Result Value Syntactic and Semantic Processing` and `API Method Call Parameter Assignment` on the side of API consumer.
- **E5**, **E6**, **E7** and **E8** are changes of `API Method Behavior` on the side of API provider and imply changes of `API Method Call Parameter Assignment` and `API Method Call Result Value Syntactic and Semantic Processing` on the side of API consumer.

VIII. CONCLUSION

The main contribution of this paper is the API ontological model, which makes it easier to see how changes on the side of the API provider affect the side of API consumer. Problem Analysis, among other things, suggests the idea of using the API ontological model as a basis for the automation of changes on the side of the API consumer implied by changes on the side of API provider. This automation could also involve an application of NST. The model could also be extended by introducing a structure to unstructured data (such as `behav-`

---

[8] https://www.omg.org/spec/UML

`ior`, `assignment`, `endpoint` or `otherSettings`) and also the idea of API SDK libraries could be applied to our model. Another interesting idea for further research would be to implement conversions between our model and other API ontology models or API description formats such as OpenAPI Specification or API Blueprint, which would make it easier to apply our model to existing APIs.

Statement on the use of AI

AI technologies (Writefull and ChatGPT) were used but only to improve the language of the paper.

REFERENCES

[1] L. Rainie and B. Wellman, *The Internet in Daily Life: The Turn to Networked Individualism*, 07 2019, pp. 27–42. ISBN 9780198843498

[2] R. Kurzweil, *The Law of Accelerating Returns*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 381–416. ISBN 978-3-662-05642-4. [Online]. Available: https://doi.org/10.1007/978-3-662-05642-4_16

[3] M. Lamothe, Y.-G. Guéhéneuc, and W. Shang, "A systematic review of api evolution literature," *ACM Comput. Surv.*, vol. 54, no. 8, oct 2021. doi: 10.1145/3470133. [Online]. Available: https://doi.org/10.1145/3470133

[4] A. Hevner, "A three cycle view of design science research," *Scandinavian Journal of Information Systems*, vol. 19, 01 2007.

[5] M. Reddy, *API Design for C++*. Elsevier Science, 2011. ISBN 9780123850041. [Online]. Available: https://books.google.cz/books?id=IY29LylT85wC

[6] N. Guarino, D. Oberle, and S. Staab, *What Is an Ontology?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–17. ISBN 978-3-540-92673-3. [Online]. Available: https://doi.org/10.1007/978-3-540-92673-3_0

[7] X. Rival and K. Yi, *Introduction to static analysis: an abstract interpretation perspective*. Mit Press, 2020.

[8] H. Mannaert, J. Verelst, and P. De Bruyn, *Normalized Systems Theory: From Foundations for Evolvable Software Toward a General Theory for Evolvable Design*. nsi-Press powered bei Koppa, 2016. ISBN 9789077160091. [Online]. Available: https://books.google.cz/books?id=0rA_tAEACAAJ

[9] G. Oorts, K. Ahmadpour, H. Mannaert, J. Verelst, and A. Oost, "Easily evolving software using normalized system theory-a case study," *Proceedings of ICSEA*, pp. 322–327, 2014.

[10] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[11] K. De Smedt, D. Koureas, and P. Wittenburg, "Fair digital objects for science: From data pieces to actionable knowledge units," *Publications*, vol. 8, no. 2, 2020. doi: 10.3390/publications8020021. [Online]. Available: https://www.mdpi.com/2304-6775/8/2/21

[12] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, 2014, pp. 1–10.

[13] A. Brito, L. Xavier, A. Hora, and M. T. Valente, "Apidiff: Detecting api breaking changes," in *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2018. doi: 10.1109/SANER.2018.8330249 pp. 507–511.

[14] B. Dagenais and M. P. Robillard, "Semdiff: Analysis and recommendation support for api evolution," in *2009 IEEE 31st International Conference on Software Engineering*, 2009. doi: 10.1109/ICSE.2009.5070565 pp. 599–602.

[15] D. Ramos, H. Mitchell, I. Lynce, V. Manquinho, R. Martins, and C. L. Goues, "Melt: Mining effective lightweight transformations from pull requests," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2023. doi: 10.1109/ASE56229.2023.00117 pp. 1516–1528.

[16] N. Deshpande, M. W. Mkaouer, A. Ouni, and N. Sharma, "Third-party software library migration at the method-level using multi-objective evolutionary search," *Swarm and Evolutionary Computation*, vol. 84, p. 101444, 2024. doi: https://doi.org/10.1016/j.swevo.2023.101444. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221065022300216X

[17] M. Lamothe, W. Shang, and T.-H. P. Chen, "A3: Assisting android api migrations using code examples," *IEEE Transactions on Software Engineering*, vol. 48, no. 2, pp. 417–431, 2022. doi: 10.1109/TSE.2020.2988396

[18] X. Gao, A. Radhakrishna, G. Soares, R. Shariffdeen, S. Gulwani, and A. Roychoudhury, "Apifix: output-oriented program synthesis for combating breaking changes in libraries," *Proceedings of the ACM on Programming Languages*, vol. 5, no. OOPSLA, pp. 1–27, 2021.

[19] R. Rolim, G. Soares, L. D'Antoni, O. Polozov, S. Gulwani, R. Gheyi, R. Suzuki, and B. Hartmann, "Learning syntactic program transformations from examples," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 2017. doi: 10.1109/ICSE.2017.44 pp. 404–415.

[20] L. Beurer-Kellner, J. von Pilgrim, C. Tsigkanos, and T. Kehrer, "A transformational approach to managing data model evolution of web services," *IEEE Transactions on Services Computing*, vol. 16, no. 1, pp. 65–79, 2023. doi: 10.1109/TSC.2022.3144613

[21] Z. Huang, J. Chen, J. Jiang, Y. Liang, H. You, and F. Li, "Mapping apis in dynamic-typed programs by leveraging transfer learning," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 4, apr 2024. doi: 10.1145/3641848. [Online]. Available: https://doi.org/10.1145/3641848

[22] E. Wittern, "Web apis - challenges, design points, and research opportunities: invited talk at the 2nd international workshop on api usage and evolution (wapi '18)," in *Proceedings of the 2nd International Workshop on API Usage and Evolution*, ser. WAPI '18. New York, NY, USA: Association for Computing Machinery, 2018. doi: 10.1145/3194793.3194801. ISBN 9781450357548 p. 18. [Online]. Available: https://doi.org/10.1145/3194793.3194801

[23] W. Granli, J. Burchell, I. Hammouda, and E. Knauss, "The driving forces of api evolution," in *Proceedings of the 14th International Workshop on Principles of Software Evolution*, ser. IWPSE 2015. New York, NY, USA: Association for Computing Machinery, 2015. doi: 10.1145/2804360.2804364. ISBN 9781450338165 p. 28–37. [Online]. Available: https://doi.org/10.1145/2804360.2804364

[24] M. Lehman, J. Ramil, P. Wernick, D. Perry, and W. Turski, "Metrics and laws of software evolution-the nineties view," in *Proceedings Fourth International Software Metrics Symposium*, 1997. doi: 10.1109/METRIC.1997.637156 pp. 20–32.

[25] D. Hou and X. Yao, "Exploring the intent behind api evolution: A case study," in *2011 18th Working Conference on Reverse Engineering*, 2011. doi: 10.1109/WCRE.2011.24 pp. 131–140.

[26] A. V. Zarras, P. Vassiliadis, and I. Dinos, "Keep calm and wait for the spike! insights on the evolution of amazon services," in *ADVANCED INFORMATION SYSTEMS ENGINEERING (CAISE 2016)*, ser. Lecture Notes in Computer Science, S. Nurcan, P. Soffer, M. Bajec, and J. Eder, Eds., vol. 9694, 2016. doi: 10.1007/978-3-319-39696-5_27. ISBN 978-3-319-39696-5; 978-3-319-39695-8. ISSN 0302-9743 pp. 444–458, 28th International Conference on Advanced Information Systems Engineering (CAiSE), Ljubljana, SLOVENIA, JUN 13-17, 2016.

[27] A. Karavisileiou, N. Mainas, and E. G. Petrakis, "Ontology for openapi rest services descriptions," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020. doi: 10.1109/ICTAI50040.2020.00016 pp. 35–40.

[28] K. Togias and A. Kameas, "An ontology-based representation of the twitter rest api," vol. 1, 11 2012. doi: 10.1109/ICTAI.2012.85 pp. 998–1003.

[29] D. Androcec and N. Vrcek, "Platform as a service api ontology," in *PROCEEDINGS OF THE 12TH EUROPEAN CONFERENCE ON EGOVERNMENT, VOLS 1 AND 2*, M. Gasco, Ed., 2012. ISBN 978-1-908272-42-3 pp. 47–54, 12th European Conference on eGovernment (ECEG), ESADE, Inst Publ Governance & Management, Barcelona, SPAIN, JUN 14-15, 2012.

[30] G. Guizzardi, G. Wagner, J. P. Andrade Almeida, and R. S. S. Guizzardi, "Towards ontological foundations for conceptual modeling: The unified foundational ontology (ufo) story," *APPLIED ONTOLOGY*, vol. 10, no. 3-4, pp. 259–271, 2015. doi: 10.3233/AO-150157

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 69–76
DOI: 10.15439/2024F9975
ISSN 2300-5963 ACSIS, Vol. 41

Job Shop Scheduling with Integer Programming, Shifting Bottleneck, and Decision Diagrams: A Computational Study

Brannon B. King
0000-0002-9269-6206
Computer Science Dept.
Virginia Tech in Blacksburg, VA
Email: brannonking@vt.edu

Robert Hildebrand
0000-0002-2730-0084
Industrial Systems and Engineering Dept.
Virginia Tech in Blacksburg, VA
Email: rhil@vt.edu

***Abstract*—We study heuristic algorithms for job shop scheduling problems. We compare classical approaches, such as the shifting bottleneck heuristic with novel strategies using decision diagrams. Balas' local refinement is used to improve feasible solutions. Heuristic approaches are combined with Mixed Integer Programming and Constraint Programming approaches. We discuss our results via computational experiments.**

I. INTRODUCTION

THE job shop scheduling problem (JSP) has long been a challenging area in operations research, historically tackled through disjunctive integer programming formulations that often yield poor linear programming relaxations. Constraint programming (CP) has emerged as a more effective approach for these problems, outperforming traditional mixed-integer programming (MIP) models. Heuristics, including scheduling and dispatching rules, have been extensively studied and applied to provide feasible solutions that can be further refined through local search algorithms. This paper explores the integration of problem-specific heuristics into modern solvers, with a focus on compound heuristic approaches utilizing Decision Diagrams (DDs) and Balas's local search methods.

Scheduling problems are known for having poor linear programming relaxations. To quote [1]: "In spite of a great deal of effort, the disjunctive integer programming formulation of the job-shop problem appears to be of little assistance in solving instances of even moderate size; furthermore, its natural linear programming relaxation has been shown to give very poor lower bounds for the problem." Solving the problem for a subset of machines/jobs seems to be the go-to mechanism for finding a better lower bound, and the solvers rely mostly on branching to improve the bounds.

On the other hand, primal heuristics for the problem abound [2]–[4]. They are often referred to as scheduling or dispatching "rules". And, as these provide feasible solutions, it's recommended to follow the heuristic with a local minimization algorithm. Also, approximation algorithms exist [5]. Balas [6] gave an inspiring local search algorithm many years ago based upon the critical path in a disjunctive graph representation of a scheduling problem. This heuristic is easily conjoined to modern MIP solvers.

Recent studies comparing constraint programming (CP) to mixed-integer programming (MIP) models show that CP clearly outperforms MIP in the realm of scheduling problems. See [7], [8], and [9]. The formulas in those papers are fairly standard and build on long-existing formulations [10]. We find, though, that tools/mechanisms for mixing heuristics with CP tools are lacking.

Recently, *Decision Diagrams* have shown to be a useful approach to some types of optimization problems. These perform a decomposition of the problem based on a sequential decision-making process. Bergman et al's book [11] on Decision Diagrams (DDs) is the starting point for our work. We assume the reader's familiarity with said book's content. We also assume familiarity with scheduling problems and their triplet notation (as in [12]). At first sight, DDs appear to be nothing more than a formal method for enumerating all possible solutions to a problem, with the detection of duplicated intermediates. However, their power is found in two separate mechanisms: 1) the restricted form that uses fixed memory to generate multiple feasible solutions, and in 2) how they can provide a relaxed form of the problem. This is done by intelligently merging nodes as the tree of solutions grows too wide. Solutions that go through one or more of these relaxed nodes provide a useful dual bound.

A. Contributions

We develop heuristic approaches based on Balas's work and on Decision Diagrams. Our novel Decision Diagram models for the JSP encourage minimizing stored symmetry, and thus reducing computational effort.

We evaluate the effectiveness of Restricted Decision Diagrams compared to traditional heuristics for JSPs. Lastly, we investigate the impact of warm-starting modern solvers with a heuristic solution. We then discuss the conclusions of

B. King and R. Hildebrand were partially funded by ONR Grant N00014-20-1-2156 and AFOSR grant FA9550-21-1-0107. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research or the Air Force Office of Scientific Research.

©2024, PTI

our computational experiments. We use GUROBI [13] and CPLEX [14] for MIP solvers and also CPLEX's constraint programming solver when testing CP versions of the JSP.

II. FORMULATION BACKGROUND

A. Job shop scheduling

The scheduling problem denoted as $Jm||C_{max}$ refers to a specific class of job shop scheduling problems (JSPs) characterized by the goal of minimizing the makespan across multiple machines. Formulas for it are common in existing literature, e.g [10]. Formal description follows:

Let there be a set of n jobs $\{J_1, J_2, \ldots, J_n\}$ and a set of m machines $\{M_1, M_2, \ldots, M_m\}$. Each job J_i consists of a sequence of operations $\{O_{i1}, O_{i2}, \ldots, O_{im}\}$, where each operation O_{ij} must be processed on a specific machine $M_{\pi_i(j)}$ for some permutation map $\pi_i \colon [m] \to [m]$ and for a predetermined duration p_{ij} without interruption. Each machine can process only one operation at a time, and each operation can be processed on exactly one machine as per its predefined sequence in a job. The objective is to find a schedule, i.e., an allocation of operations to time slots on each machine, that minimizes the makespan C_{max}, which is the time when the last job completes processing.

1) Mathematical Optimization Formulation: We will focus on the disjunctive MIP formulation, which generally solves the quickest using MIP solvers [10]. The idea is to model the problem using binary variables to represent the sequencing decisions between operations on the same machine. Here's a step-by-step formulation:

Variables:

- S_{ij}: Nonnegative start time of operation O_{ij}.
- C_{ij}: Completion time of operation O_{ij} and easily collapsed into $S_{ij} + p_{ij}$.
- $C_{\max}$: Maximum completion time (makespan).
- x_{ijkl}: Binary variable indicating operation O_{ij} follows O_{kl}, only existing if both are on the same machine.

Constraints:

- Precedence Constraints: Ensure the correct order of operations within each job.
- Disjunctive Constraints (Eqs. 1c, 1d): Ensure that no two operations on the same machine overlap by enforcing that one must precede the other.

Mixed Integer Programming Model:

$$\min C_{\max} \quad \text{subject to} \tag{1a}$$

$$S_{ij} \geq C_{i(j-1)} \quad \forall i \in [n], \forall j \in J_i, j > 1 \tag{1b}$$

$$S_{ij} \geq C_{kl} - \overline{M}(1 - x_{ijkl}) \quad \forall ij, kl : \pi_i(j) = \pi_k(l) \tag{1c}$$

$$S_{kl} \geq C_{ij} - \overline{M} x_{ijkl} \quad \forall ij, kl : \pi_i(j) = \pi_k(l) \tag{1d}$$

$$S_{ij} \geq 0 \quad \forall i, j \tag{1e}$$

$$C_{ij} = S_{ij} + p_{ij} \quad \forall i, j \tag{1f}$$

$$C_{\max} \geq C_{ij} \quad \forall i, j : O_{ij} \text{ is final task of job } i \tag{1g}$$

$$x_{ijkl} \in \{0, 1\} \quad \forall i, j, k, l : \pi_i(j) = \pi_k(l) \tag{1h}$$

where $\overline{M}$ is a big-M multiplier, typically set to the one plus the sum of the delays: $1 + \sum_{ij} p_{ij}$. Those big-M constraints may also be formulated using the solver's indicator constraint functionality.

2) Constraint programming formulation: See the full explanation in [9].

Variables:

- I_{ij}: Interval variable containing the start and end of operation O_{ij} with width as the duration$= p_{ij}$.

CP Formula:

$$\min \max_{i \in [n]}(\text{end\_of}((I_{im}))) \quad \text{subject to:} \tag{2a}$$

$$\text{no\_overlap}(\{I_{ij} \mid i \in [n]\}) \quad \forall j \in [m] \tag{2b}$$

$$\text{end\_before\_start}(I_{ij}, I_{i(j+1)}) \quad \forall i \in [n], j \in [m-1] \tag{2c}$$

$$I_{ij} = \text{IntervalVar}(p_{ij}) \quad \forall i \in [n], j \in [m] \tag{2d}$$

B. Existing Work regarding scheduling via DD

Bergman et al. [11] discuss how a DD can solve scheduling problems in general and give this example: the single-machine makespan minimization with sequence-dependent delays: $1|p_{ij}|C_{max}$. They don't use a binary expansion tree; instead, they represent the problem as a permutation of possible orderings, known as a multivalue decision diagram (MDD), as shown in Table I.

TABLE I: Bergman's simple state operators

State	S_j holds the j jobs already done
Transition	$S_j \cup x$ where $x \in [n]$ and $x \notin S_j$
Cost	the delay from S_j to S_{j+1} via x

They do not give merge and split definitions for their scheduling example. They do give merge operations for other problems, namely, maximum independent set, maximum cut, and maximum 2-SAT. They also discuss how some merge operations might be possible for a scheduling problem if we separate the jobs already done into two groups: those that are surely done no matter what path arrives at a given state, and those that are done in at least one path arriving at a given state. This latter group is the "maybes". In [11]'s terms, the two groups are “All” and “Some”. The maybes doesn't exist in a full expansion because, in that context, we don't have any nodes that represent more than one unique solution.

Hooker, in [15], expands on the above ideas for a more complicated scheduling problem shown in Table II, minimizing tardiness given release dates and due dates $1|r_j, d_j|\sum T_j$.

TABLE II: Hooker's tardiness operations

State	a tuple $S_j = (V, U, f)$: V holds up to j jobs surely done, U holds jobs done on some route, f is the running completion time
Transition	$(V \cup x, U \cap x, \max(r_x, f) + p_x)$ where $x \in [n]$ and $x \notin V$
Cost	the delay from S_j via x, the value of p_x
Merge	$(V \cap V', U \cup U', \min(f, f'))$

His paper demonstrates that these operations are sufficient to ensure that the relaxed tree contains a path that represents a dual (in this case, lower) bound. He also gives a mechanism for proving any merge operation to be sufficient. He expands his effort with a later paper, [16], where he includes merge operations for sequencing with time windows, time-dependent processing times, sequence-dependent processing times, and state-dependent processing times. They all follow a pattern very similar to the one given above.

Moreover, in [16], Hooker suggests optimizing the Lagrangian dual where he incorporates an additional penalty for sequences that repeat operations – a common infeasibility in a relaxed DD representation of a schedule. It was inspired by [17]. The paper also includes a table of timed runs on the CPW and Biskup-Feldman public datasets containing tardiness problems. (We were unable to make this succeed in our context. Perhaps it was due to some failure of our model to meet the necessary assumptions, or it required an extremely high number of iterations for convergence, or coding error.)

Building on Hooker's work, [18] provides two recent papers tackling multi-machine scheduling. In [18], they focus on a tardiness problem with substantial state; $(V, U, f, t, f^u, t^u, g)$. f, t, f^u, t^u are all vectors, where f refers to running completion times per machine, t refers to running release times, and the superscript u implies the same for the maybes (the items included in some ancestral lines but not all). They show that their merge operation is suitable using Hooker's rules. It's given here:

$$\begin{gathered}(V \cap V',\ U \cup U',\ \min(f, f'),\ \min(t, t'),\\ \max(f^u, f^{u\prime}),\ \max(t^u, t^{u\prime}),\ \min(g, g'))\end{gathered}$$

In their later work, [19], they build rules for uniform scheduling over total tardiness, or $Um|r_j, st_{jj'}, d_j|\sum T_j$. They track the current machine as part of the state. This leads to a notable limitation; their merge operation is only allowed to merge nodes where the current machine matches. They generally follow the patterns given above for tardiness problems.

C. Decision Diagrams

See [20] for a recent review of decision diagrams (DDs) for discrete optimization. We follow that survey for some details.

III. DD OPERATORS FOR THE JSP

In this section we present several different models that store and transition state in DDs. All solve the JSP, but not at the same efficiency. They are not the only possible models. For merging state, additional information must be stored, and that is covered in a separate section.

If we assign each operation a unique identifier, we have all feasible solutions as the permutations of those identifiers. Of course, most permutations are infeasible in that some job's operations may be out of order. Moreover, if we have any permutation, either complete or partial, we can trivially compute its completion times, feasibility, and, for partial sets, a lower bound on the completion max. See Algorithm 1.

Algorithm 1 Cost from partial solution (CFP)

```
Require: $X$ is a list of to-be-done operations.
Require: $\mathbf{f}^M$ is machine completion times, 0 by default.
Require: $O$ operations already done with times $\mathbf{f}^O$.
  for $x \in X$ do
    $s \leftarrow \mathbf{f}^M_{mach(x)}$
    if $\exists\ pre(x)$ then
      if $pre(x) \notin O$ then
        return Error: Missing Prerequisite
      $s \leftarrow \textsc{max}(s, \mathbf{f}^O_{pre(x)})$
    $\mathbf{f}^O_x \leftarrow s + \textsc{delay}(x)$
    $\mathbf{f}^M_{mach(x)} \leftarrow s + \textsc{delay}(x)$
  return $\textsc{max}(\mathbf{f}^M)$
```

Given a partial ordering and the next operation to be concatenated to it, we can trivially compute the change in completion times brought about by the additional operation. We can also update the running C_{max} if it is changed by this additional operation.

A. The basic permutation model

Since we're just storing a partial ordering, all we need is a list. We make use of some helpers such as $pre(x)$, which returns the operation required right before x if there is one. $mach(x)$ returns the required machine for operation x. $delay(x)$ is the delay for operation x, commonly referred to as p_{ij}. $trailer(x)$ contains the sum of the operation times that must follow operation x, where this could be either the amount of work left on the job of x or the machine required for x (or the maximum of both of those). Given a list of operations with labels in $[n]$, we get Table III.

TABLE III: JSP-for-MDD Model 0

State	$S := V$, an ordered list of ops. done so far
Cost	$c(S, x) :=$ CFP$([V, x])$
Transition	$\phi(S, x) := [V, x] : x \in [n] \backslash V,\ pre(x) \in V$

If we make our model slightly more advanced, we can cache the completion times for each operation (in $\mathbf{f}^O$) and each machine (in $\mathbf{f}^M$), we get Table IV.

TABLE IV: JSP-for-MDD Model 1

State	a tuple $S := (V, \mathbf{f}^O, \mathbf{f}^M)$
Cost	$c(S, x) := \max(\mathbf{f}^M_{mach(x)}, \mathbf{f}^O_{pre(x)}) + delay(x)$
Transition	$\phi(S, x) := (V \cup \{x\},$ $\mathbf{f}^O_x \leftarrow c(S, x), \mathbf{f}^M_{mach(x)} \leftarrow c(S, x))$

It's implied that $(V, \mathbf{f}^O, \mathbf{f}^M)$ are independent for every state – copied from the parent state and then modified/extended. When storing the states in each layer, it's useful to store duplicate states only once. When comparing these states, all fields of the tuple must be compared. Notice that only states within a given layer will have equal cardinality for V (unless states are merged in some way that violates that, as discussed in the merge section below).

B. A model to detect more symmetry

When growing a tree of possible solutions, such as is done by decision diagrams, one may arrive at equivalent solutions through differing paths. In this context such solutions are symmetrical. With a goal of keeping the tree of possible solutions as small as possible, DDs benefit from any reduction in state space. We want to avoid symmetry in the storage of our accumulated state.

The above model is not bad, but notice (Table V) that many of the stored completion times can never eclipse any machine's finish time. Sometimes those may happen in some other order and still produce the same state; we redesign the state to capture that symmetry.

TABLE V: JSP-for-MDD Model 2

State	a tuple $S := (V, V_L, \mathbf{f}^O, \mathbf{f}^M)$;
Cost	$c(S, x) := \max(\mathbf{f}^M_{mach(x)}, \mathbf{f}^O_{pre(x)}) + delay(x)$
Transition	$\phi(S, x) := (V \cup \{x\} \backslash pre(x), V_L \cup pre(x),$ $\mathbf{f}^O_x \leftarrow c, \mathbf{f}^M_{mach(x)} \leftarrow c)$

V_L refers to those operations that are long-done, those that will never need to be used as an immediate prerequisite. *For hash and comparison we ignore the completion times of operations in* V_L. If $pre(x) \notin \mathbf{f}^O$ you can return zero, although tracking some completion time for items in V_L can be handy for generating the final schedule at the end.

In Table VI, we show the number of nodes for full expansion averaged over 10 random problems to demonstrate that the number of nodes expanded is reduced by capturing more symmetry:

TABLE VI: Node expansion demonstration

	Model 1	**Model 2**
Valid nodes, 4x5 JSP	1245k	793k
Duplicates	1017k	776k
Valid nodes, 3x10 JSP	726k	528k
Duplicates	638k	504k

C. Modeling disjunctives directly

While the MILP formulation described above relies on real variables for start and completion times, the values for these variables can be fully determined from a fixed set of binary variables, denoted as x. This is common in scheduling problems. Consequently, this can be directly modeled as a binary decision diagram (BDD), as opposed to the multivalue or MDDs mentioned earlier. For each disjunction, a binary variable x_i determines whether the path is forward or reverse. In this context, forward implies that $i < j$ in an adjacency matrix representation of the disjunctive graph of the problem, as explained in [6]. Following the approach in that paper, it is assumed that the process begins with a feasible set of disjunctives, all oriented forward.

This approach encompasses a significantly larger number of variables and consequently necessitates many more layers compared to MDDs. For example, the well-known JSP problem *abz5* involves 100 operations but 900 disjunctives. (The number of dijunctives matches the number of binary variables shown in the Gurobi log, but it can also be easily computed from the number of bidirectional arcs in 10 cliques of 10 nodes each: $10n(n-1)$.) Grouping these bits into bytes could reduce the number of layers by a factor of eight while simultaneously increasing the width of each layer by the same factor. Additionally, this method might potentially reduce the 256 possible expansions by eliminating common 3- or 4-cycle problems from the possible values, although this approach has not yet been explored.

Given a disjunctive model, including invalid models with cycles, we can always find the longest path through that model in polynomial time or less. That's done trivially with a flow LP or an adjacency matrix selection LP. The challenge is to reduce the number of LP calls needed, as one for each state (or every other state as shown below) gets expensive. This approach was generally unhelpful although it did work correctly; it's included here in Table VII for contrast.

TABLE VII: Basic BDD-with-LP for layer j

State	V holds the reversed disjunctives' index
Transition	$V' = [V, j]$ if $x = 1$ else $V' = V$
Cost	LP(V)

With this approach, each node is unique; there is no duplication and no reduction in possible values as you progress through the layers. You cannot cull states that have cycles in their graph as later reversals may eliminate those cycles. As with all DD approaches, it would be worth it to cull via some known primal bound, heuristically determined.

D. Merging state

Both Model 1 and Model 2 support merge operations. Following [15] and [18], we add additional fields to the state: V_s and $\mathbf{f}^s$, where s stands for "some", meaning that some ancestral path covered the items in this set. For merging Model 1 state, $S' \leftarrow S \oplus \overline{S}$:

$$\mathbf{f}^{M\prime}_i \leftarrow \min(\mathbf{f}^M_i, \overline{\mathbf{f}^M_i}) \quad \forall i \in M \tag{3a}$$

$$V' \leftarrow V \cap \overline{V} \tag{3b}$$

$$\mathbf{f}^{O\prime}_x \leftarrow \max(\mathbf{f}^O_x, \overline{\mathbf{f}^O_x}) \quad \forall x \in V' \tag{3c}$$

$$V'_s \leftarrow (V \cup \overline{V}) - V' \tag{3d}$$

$$\mathbf{f}^{s\prime}_x \leftarrow \max(\mathbf{f}^O_x, \overline{\mathbf{f}^O_x}, \mathbf{f}^s_x, \overline{\mathbf{f}^s_x}) \quad \forall x \in V'_s \tag{3e}$$

Hooker [15] gives two criteria for a valid merge: any possible control values leaving S' must be a valid relaxation of the same control leaving S, and both S and $\overline{S}$ must be relaxed/interchangable. The latter criteria holds from the lack of order-dependent operations in the above formulation. The former criteria holds in that we always take the minimum score for our $C_{max} = \max(\mathbf{f}^{M\prime}_i)$, and we always use the worst-case prerequisite completion time when pulling items from $\mathbf{f}^O$. Model 2 is merged similarly with V_L receiving equivalent treatment to V.

We used the merge operation to validate that Bergman's branch-and-bound algorithm [11] worked for JSP. It did not scale well, though. We were unable to make the relaxed DD give a better lower bound than the linear polytope of the disjunctive formulation, so we do not include this in our computational results below. As part of that, we pruned nodes that exceeded the current primal bound, which is known as LocB pruning in this context [21]. See the other thoughts on the algorithm in the appendix.

IV. A SIMPLE LOCAL SEARCH REFINEMENT

In Balas' original paper [6] about solving JSP via disjunctive graph iteration, he included this special proposition:

Proposition of Balas 1969:

> Let C_h be a critical path in G_h [which is a DAG]. Any graph G_k obtained from G_h by complementing one arc $(i, j) \in C_h$ is circuit-free.

Balas' proof: We know that (i, j) must be the longest path from i to j as it is part of C_h. However, it is also the shortest path from i to j or we would have chosen the longer path. Because i to j is the only path, we can reverse it without creating a circuit.

That leads us to this refinement algorithm as a local search method (Alg. 2):

Algorithm 2 Balas Local Refinement 1 (LNS1)

```
Require: $g = (V, \mathcal{A})$ is a DAG of the fixed conjunctive arcs.
Require: $W_e \geq 0$ for $e \in \mathcal{A}$ as the weight of each arc.
Require: $s_{parent}$ is the $s$ value from the caller.
    $p \leftarrow$ LONGEST_PATH$(g)$
    $s \leftarrow \sum_{(u,v)\in p} W_{uv}$
    if $s > s_{parent}$ then
        return $s$                // remove this to search more space
    for $e \in p$ do
        if not $e.fixed$ then
            $g \leftarrow$ REMOVE_ARC$(g, e)$
            $g \leftarrow$ ADD_FIXED_ARC$(g, e.v, e.u, W_{vu})$
            $t \leftarrow$ RECURSE$(g, W, s)$
            if $t < s$ then
                $s = t$               // can also track swaps here
            $g \leftarrow$ REMOVE_ARC$(g, e.v, e.u)$
            $g \leftarrow$ ADD_FIXED_ARC$(g, e)$
    return $s$                    // return swaps also if desired
```

V. COMPUTATIONAL EXPERIMENTS

A. Use as a Heuristic

In this section we show how a restricted Model 2 compares to other common heuristics including MOR, MWR, and the shifting bottleneck [22]. We include multiple widths for the restricted DD, but this parameter does not substantially improve the bound it computes; see [15].

Observation: The restricted DD always produces at least one feasible solution. Given any partial solution that is feasible, the remaining operations can always be added in a feasible order. Note that this does not hold if you filter the nodes in the DD with anything additional to the maximum width filter. For example, if you filter nodes whose C_{max} exceeds some threshold (in addition to reducing row width), you may filter all possible paths toward the conclusion of the DD.

In Table VIII, we run each heuristic over twenty random 10×10 JSP instances, making use of Gurobi [13], CPLEX [14], and Job Shop Library [23]. $\mathcal{W}$ refers to the maximum width of the DD. Overage is how far the final bound was above the optimum. We use default solver settings with the exception of Gurobi's AggFill=10 and GomoryPasses=1, which was recommended by their tuner for these problems. The random instances are integer and similar to those published with [22]).

TABLE VIII: Avg. heuristic overages for 20 random

Heuristic	Time	Overage	After LNS1
Gurobi, MIPGap=0.25	0.27s	11.1%	9.18%
Shifting Bottleneck	4.0s	15.8%	15.3%
Restricted DD, $\mathcal{W}$=200	0.25s	18.5%	14.6%
Restricted DD, $\mathcal{W}$=400	0.50s	16.8%	12.5%
Most Work Remaining (MWR)		26.9%	16.1%
Most Ops. Remaining (MOR)		20.1%	14.1%
Shortest Proc. Time (SPT)		90.0%	40.1%

TABLE IX: Avg. heuristic overages for 18 from JSPLIB

Heuristic	Time	Overage	After LNS1
Gurobi, MIPGap=0.25	1.6s	7.77%	5.85%
Gurobi, MIPGap=0.40	0.12s	25.8%	12.5%
Shifting Bottleneck		22.8%	21.9%
Restricted DD, $\mathcal{W}$=200		14.4%	11.9%
Restricted DD, $\mathcal{W}$=400		11.6%	9.70%
Most Work Remaining (MWR)		31.4%	19.8%
Most Ops. Remaining (MOR)		29.9%	23.6%
Shortest Proc. Time (SPT)		80.8%	39.7%

Interestingly, (Table IX) the DD approach worked better on the real-world problems found in JSPLIB [24] – it's eighteen 10×10 problems. Similarly, the small 0.25 MIPGAP for Gurobi was much more difficult to achieve on the JSPLIB problems.

The comparison is a little bit unfair, in that the Restricted DDs generate many feasible solutions for the LNS1 whereas the top four only produce a single solution to be refined. However, it shows that the local refinement eliminates the need for the shifting bottleneck heuristic.

We recognize that there are more sophisticated versions of the shifting bottleneck algorithm, e.g. [25]. There are also a variety of other local search mechanisms designed for JSP that are far more sophisticated and far-reaching, typically built on taboo search, e.g. [26], [27]. We did not consider simulated annealing nor any evolutionary algorithm as part of this research either, though papers on those approaches for JSP abound.

Regarding Gurobi NoRel: We ran Gurobi's NoRel heuristic for 4 seconds on the same 20 problems. It failed to find any solutions on 80% of the problems, but on the other four, it found solutions within 3% of optimal. Note that Gurobi can solve a 10x10 disjunctive program in 2 to 8 seconds on our test machine, so running a 4-second heuristic for it would not make sense generally. The NoRel runtime has to be specified as an input.

Regarding runtimes: Note that the Shifting Bottleneck (SB) without readjustment of machines in M_0 plus the LNS at the end still achieves 18%. This takes about two seconds to run whereas the other takes 4 seconds per 10x10. SB can be modified to solve subproblems in parallel, which was not a part of our implementation. We rely on Gurobi to solve the $1|r_j|L_{max}$ subproblems in the SB. This takes up 90% of the runtime for it. Generally, though, when using SB one would use Carlier's approximation [28] for the L_{max} instead of solving it via a MIP solver (and probably still run them in parallel). See other ideas here: [5], [29]. With our DD written in Go, the 10x10 on a max-width of 400 takes half a second to run and half that time for the 200 width (using no parallelism). Most of that time is in comparison to previous nodes on the same row. Better hashing would improve that time. The LNS1 adds some additional time to that as it runs on each resulting node. This is not included in the time measurement listed above. The dispatching rule approximations obviously use a trivial amount of time.

Most of the items arriving at the bottom row of a DD tend to be similar, which comes from the sort-and-truncate approach. It needs some other selection criteria toward the top of the tree so that it keeps more diversity early on, which should give it better chances of enabling a good/unique neighborhood. The DD is quite sensitive to changes in the running C_{max} computation. For example, you can use the C_{max}-so-far or you can add to the trailer for the remaining items to be done on each machine or you can add the work remaining on the job. Those selections all change results quite drastically. The run recorded above does not add a trailer, as going without seemed slightly better on average.

B. The value of a starting point

Here we demonstrate the value of using a heuristic to select a starting solution when solving the JSP to its optimum. In Table X we run each solver over the same 20 instances used above but give it a single starting point. The starting points are derived using MOR followed by the refinement of the LNS1 algorithm described above.

TABLE X: Solve time with single warmstart

Solver on 10x10	Time/problem	With warmstart
Gurobi, big-M	2.1s	1.9s
Gurobi, indicator	4.0s	3.6s
CPLEX MP, big-M	3.4s	3.8s
CPLEX MP, indicator	130s	110s
CPLEX CP	1.17s	1.16s
Solver on 12x12	**Time/problem**	**With warmstart**
Gurobi, big-M	36s	43s
CPLEX MP, big-M	66s	60s
CPLEX CP	5.3s	5.5s

We conclude from this table that you should be using a constraint solver for exact solutions on this, and that the big-M path is more optimized than the indicator constraint feature, and that warm-starting it is unnecessary. Runs were made with Gurobi 11.0.2, CPLEX 22.1.1 on a i7-8750H processor.

Gurobi supports a heuristic parameter for controlling the percentage of time spent in heuristics. The default is 5%. We explored other values from 0% through 50% but could find no other value to improve the average time. Increasing the parameter by 5% generally added 5% to the runtime.

C. The value of LNS via callback

We demonstrate (Table XI) the value of calling LNS1 in the MIP node callback (CB). We take the ordering given by the start time variables (S) and run the local search on that. We can actually just submit the solution given by the ordering instead of running a local optimizer on it – herein called "Nearest".

TABLE XI: LNS1 in MIP-node callback

Solver	Default	LNS1	Wins	Nearest	Wins
Gurobi	2.1s	2.6s	5.6	2.1s	6.7

Note that we subtract the time of the callback itself, in that it is assumed that we can come up with more efficient implementations of it or curtail the calls to it as it becomes unlikely to assist. This optimization of how often the heuristic is called is a separate issue. Writing our LNS algorithm in a heuristic form that is fast enough to justify its use is nontrivial.

CPLEX, as documented, supports heuristic callbacks. However, in attempting to use them with version 22.1.1 on Linux, accessed via the Python docplex wrapper, we were unable to determine whether or not the solver was utilizing the provided heuristic solutions. No errors were given, but the incumbent scores were not updating as expected, so we did not include the numbers for it. We also have interesting numbers for FICO Xpress, but their license prohibits publishing.

D. Using our state model for A*

Unfortunately, Model 2 alone does not seem to be sufficient to allow solving a 10x10 via A*. We note that there are other efforts to make A* utilize relaxation features of the DD approach such as [30]. The nearness of the running C_{max} to reality is of critical importance in A*-search. It is possible to improve the trailer estimate by solving 2 of ten machines: see [31]. This is fairly quick, especially as the problems progress and most tasks have release times available. However, empirically, it's not enough accuracy to make the A* approach feasible.

VI. CONCLUSION

Conclusions from our experimentation:

1) Relaxed decision diagrams are useful as a simple JSP heuristic. They are not difficult to write/use and run fast. They produce better results than other common (simple) heuristics on real-world problems.
2) Passing a start point to the solver is not useful at 14% away from optima. Perhaps it would be worth it if you were using some more sophisticated heuristic that could

generate starting points within just a few percentage points from optima.

3) Running Balas' critical path refining, the LNS1, does generally improve a given feasible solution. It is fast to run and generally worth it.
4) The big-M handling in Gurobi and CPLEX is superior to their indicator constraint handling at present. It may be that our $\overline{M}$ value was small enough to tip that balance.
5) For problems where feasible solutions are rare, it may be helpful to find a nearby feasible solution in the callback if it can be done quickly. This computation is very fast for the JSP. It did help significantly on some of the test problems. Especially consider it if you use FICO Xpress.

Ideas for future work:

1) The selection of keeper nodes in the restricted DD is of critical importance. Using a basic rule like keeping the smallest 200 is a general failure – most of the nodes with the optimum are weeded out early on. That's the curse of these scheduling problems – the conflicts on the attractive solutions don't show up until late in the game. If we had some kind of machine learning approach that could identify bad nodes early on, we would have a higher chance of retaining the good nodes (or vice-versa). Node selection ideas from modern solvers may also apply [32].
2) Another idea for retaining nodes is to try to keep a diverse set using some kind of diversity measure that would increase the likelihood of keeping the optimum path.
3) Relaxed decision diagrams produce many infeasible solutions. Order them and you can expand these infeasible paths until you arrive at the first and best feasible solution, a best first approach similar to A*. We attempted this. However, there are so many infeasible solutions to weed through that this is generally not a viable approach for problems at scale. If we had some equivalent to cuts-for-LP, perhaps we could cut out subtrees in a way that allowed us to arrive at the best solution much faster.

APPENDIX

A. Notes on existing algorithms for exact solutions via DD

[11] presents two general algorithms for finding an exact solution to any program representable by a DD. The first mechanism is what they term "compiling DDs by separation", condensed form of the algorithm given in [33].

Algorithm summary: Begin with a relaxed decision diagram (DD) and identify the optimal path through it. If this path violates any constraints, separate the relaxed nodes on that path into two or more replacements. Adjust the inputs feeding into the violating node so that some go to each replacement. Similarly, replicate the outputs of the violating node to each replacement. Continue this process until the optimal path is feasible.

The process requires a split or "separation" operation, which essentially undoes the merge operation, though the necessary bookkeeping for this may be expensive. If no split operation is available, a possible solution is to backtrack to the parents of the merged nodes and regenerate their children. Additionally, we assume that the arcs store and maintain the variable value (also known as the control) and the cost of traversing them. This assumption differs from our previous experiments, where we kept only the fringe nodes with the running cost in the state. Furthermore, decision diagram (DD) creation generally employs node reduction (combining identical nodes), and this reduction must be maintained after adding additional nodes to the graph. If all identical nodes are on the same layer, the check is reasonable; otherwise, it becomes too expensive.

The second algorithm represents a branch-and-bound (BnB) approach, where you branch on a cutset of exact nodes, making a new subtree pair, both relaxed and restricted, for the decedents of each node in the cutset. Cutset refers to a set of exact nodes such that any path through the tree goes through one and only one of these nodes (before hitting any relaxed nodes).

Algorithm summary: While there are nodes in the queue, remove node u. Update the primal bound, which is the cost to node u plus a heuristic from it to the end, potentially determined by a restricted tree. Construct a relaxed decision diagram (DD) with u as its root. If the best relaxation is worse than the primal bound, exclude the entire u subtree. Otherwise, add the exact cutset of u to your queue and repeat the process. Some general notes on this algorithm:

1) The processing of these subtrees is parallelizable (as noted in the reference).
2) It does not require a split operation, although it does require a working merge operation for building the relaxed trees.
3) It doesn't require a restricted tree if you have some other heuristic mechanism that completes partial solutions, as that may also provide a reasonable primal bound, especially if it's refined by a fast local search as the final step of the heuristic.
4) It doesn't make any use of the dual bound for subtree exclusion. This is its fundamental weakness.
5) Empirically, it's highly unlikely that to be able to exclude the whole relaxed tree based on its best node being worse than the current overall primal bound. Hence, you can simply return the cutset as soon as it is discovered. This eliminates the need for a merge operation.

The algorithm makes use of two things from the relaxed DD: its best path cost and its exact cutset. [11] gives three algorithms for selecting the cutset: take the first layer, take the last layer before any nodes are merged, or take the "frontier", meaning all the exact nodes that have at least one relaxed child. From that, we make these observations:

1) If we merge many nodes into one, that node has a high likelihood of being very relaxed. Thus we will keep it, as it has a good score, which will in turn lead to the best path through the relaxed tree being a poor estimate

of reality. Hence, again we will keep that tree's cutset, as our primal bound won't be able to exclude it.

2) If we take some layer before we merge any nodes, our cutset will be very shallow. Shallow nodes have lower likelihood of being excluded by constraints, assuming most constraints incorporate more than the first few variables. Moreover, it is utilizing less of our DD.
3) If we choose merge a lot in hopes of not over-relaxing any one path, we will force our cutset to be more shallow, thus getting less advantage from our DD expansion.

REFERENCES

[1] P. Martin and D. B. Shmoys, "A new approach to computing optimal schedules for the job-shop scheduling problem," in *Integer Programming and Combinatorial Optimization*, W. H. Cunningham, S. T. McCormick, and M. Queyranne, Eds. Springer Berlin Heidelberg, 1996, vol. 1084, pp. 389–403. ISBN 978-3-540-61310-7 978-3-540-68453-4. [Online]. Available: http://link.springer.com/10.1007/3-540-61310-2_29

[2] Y. P. Gupta, M. C. Gupta, and C. R. Bector, "A review of scheduling rules in flexible manufacturing systems," *International Journal of Computer Integrated Manufacturing*, vol. 2, no. 6, pp. 356–377, Nov. 1989. doi: 10.1080/09511928908944424. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09511928908944424

[3] S. S. Panwalkar and W. Iskander, "A Survey of Scheduling Rules," *Operations Research*, vol. 25, no. 1, pp. 45–61, Feb. 1977. doi: 10.1287/opre.25.1.45

[4] A. Allahverdi, "The third comprehensive survey on scheduling problems with setup times/costs," *European Journal of Operational Research*, vol. 246, no. 2, pp. 345–378, Oct. 2015. doi: 10.1016/j.ejor.2015.04.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377221715002763

[5] N. Grigoreva, "Worst-case analysis of an approximation algorithm for single machine scheduling problem," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, ser. FedCSIS 2021. IEEE, Sep. 2021. doi: 10.15439/2021f66. ISSN 2300-5963

[6] E. Balas, "Machine Sequencing Via Disjunctive Graphs: An Implicit Enumeration Algorithm," *Operations Research*, vol. 17, no. 6, pp. 941–957, Dec. 1969. doi: 10.1287/opre.17.6.941

[7] W.-Y. Ku and J. C. Beck, "Mixed integer programming models for job shop scheduling: A computational analysis," *Computers & Operations Research*, vol. 73, pp. 165–173, 2016. doi: 10.1016/j.cor.2016.04.006

[8] G. Da Col and E. C. Teppan, "Industrial-size job shop scheduling with constraint programming," *Operations Research Perspectives*, vol. 9, p. 100249, 2022. doi: 10.1016/j.orp.2022.100249

[9] B. Naderi, R. Ruiz, and V. Roshanaei, "Mixed-integer programming vs. constraint programming for shop scheduling problems: New results and outlook," *INFORMS Journal on Computing*, vol. 35, no. 4, pp. 817–843, 2023. doi: 10.1287/ijoc.2023.1287

[10] C.-H. Pan, "A study of integer programming formulations for scheduling problems," *International Journal of Systems Science*, vol. 28, no. 1, pp. 33–41, Jan. 1997. doi: 10.1080/00207729708929360. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00207729708929360

[11] D. Bergman, A. A. Cire, W.-J. Van Hoeve, and J. Hooker, *Decision Diagrams for Optimization*, ser. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, 2016. ISBN 978-3-319-42847-5 978-3-319-42849-9. [Online]. Available: http://link.springer.com/10.1007/978-3-319-42849-9

[12] Michael L. Pinedo, *Scheduling: Theory, Algorithms, and Systems*. Springer International Publishing, 2022. ISBN 978-3-031-05920-9 978-3-031-05921-6. [Online]. Available: https://link.springer.com/10.1007/978-3-031-05921-6

[13] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2023. [Online]. Available: https://www.gurobi.com

[14] IBM, *IBM ILOG CPLEX Optimization Studio*, 2023. [Online]. Available: https://www.ibm.com/products/ilog-cplex-optimization-studio

[15] J. N. Hooker, "Job Sequencing Bounds from Decision Diagrams," in *Principles and Practice of Constraint Programming*, J. C. Beck, Ed. Springer International Publishing, 2017, vol. 10416, pp. 565–578. ISBN 978-3-319-66157-5 978-3-319-66158-2. [Online]. Available: http://link.springer.com/10.1007/978-3-319-66158-2_36

[16] ——, "Improved Job Sequencing Bounds from Decision Diagrams," in *Principles and Practice of Constraint Programming*, T. Schiex and S. De Givry, Eds. Springer International Publishing, 2019, vol. 11802, pp. 268–283. ISBN 978-3-030-30047-0 978-3-030-30048-7. [Online]. Available: http://link.springer.com/10.1007/978-3-030-30048-7_16

[17] D. Bergman, A. A. Cire, and W.-J. Van Hoeve, "Lagrangian bounds from decision diagrams," *Constraints*, vol. 20, no. 3, pp. 346–361, Jul. 2015. doi: 10.1007/s10601-015-9193-y. [Online]. Available: http://link.springer.com/10.1007/s10601-015-9193-y

[18] P. Van Den Bogaerdt and M. De Weerdt, "Multi-machine scheduling lower bounds using decision diagrams," *Operations Research Letters*, vol. 46, no. 6, pp. 616–621, Nov. 2018. doi: 10.1016/j.orl.2018.11.003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016763771830227X

[19] ——, "Lower Bounds for Uniform Machine Scheduling Using Decision Diagrams," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, L.-M. Rousseau and K. Stergiou, Eds. Springer International Publishing, 2019, vol. 11494, pp. 565–580. ISBN 978-3-030-19211-2 978-3-030-19212-9. [Online]. Available: http://link.springer.com/10.1007/978-3-030-19212-9_38

[20] M. P. Castro, A. A. Cire, and J. C. Beck, "Decision Diagrams for Discrete Optimization: A Survey of Recent Advances," *INFORMS Journal on Computing*, vol. 34, no. 4, pp. 2271–2295, Jul. 2022. doi: 10.1287/ijoc.2022.1170

[21] X. Gillard, V. Coppé, P. Schaus, and A. A. Cire, "Improving the filtering of branch-and-bound mdd solver," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, P. J. Stuckey, Ed. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-78230-6_15. ISBN 978-3-030-78230-6 pp. 231–247.

[22] J. Adams, E. Balas, and D. Zawack, "The Shifting Bottleneck Procedure for Job Shop Scheduling," *Management Science*, vol. 34, no. 3, pp. 391–401, Mar. 1988. doi: 10.1287/mnsc.34.3.391

[23] Pablo Ariño, *Job Shop Library*. [Online]. Available: https://github.com/Pabloo22/job_shop_lib

[24] T. Yasumasa, *JSPLIB-Benchmark instances for the job-shop scheduling problem*. [Online]. Available: https://github.com/tamy0612/JSPLIB

[25] E. Balas, N. Simonetti, and A. Vazacopoulos, "Job shop scheduling with setup times, deadlines and precedence constraints," *Journal of Scheduling*, vol. 11, no. 4, pp. 253–262, Aug. 2008. doi: 10.1007/s10951-008-0067-7. [Online]. Available: http://link.springer.com/10.1007/s10951-008-0067-7

[26] É. D. Taillard, "Parallel Taboo Search Techniques for the Job Shop Scheduling Problem," *ORSA Journal on Computing*, vol. 6, no. 2, pp. 108–117, May 1994. doi: 10.1287/ijoc.6.2.108

[27] E. Nowicki and C. Smutnicki, "An Advanced Tabu Search Algorithm for the Job Shop Problem," *Journal of Scheduling*, vol. 8, no. 2, pp. 145–159, Apr. 2005. doi: 10.1007/s10951-005-6364-5. [Online]. Available: http://link.springer.com/10.1007/s10951-005-6364-5

[28] J. Carlier, "The one-machine sequencing problem," *European Journal of Operational Research*, vol. 11, no. 1, pp. 42–47, 1982.

[29] M. Sinai and T. Tamir, "Minimizing tardiness in a scheduling environment with jobs' hierarchy," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, ser. FedCSIS 2021. IEEE, Sep. 2021. doi: 10.15439/2021f36. ISSN 2300-5963

[30] M. Horn, J. Maschler, G. R. Raidl, and E. Rönnberg, "A*-based construction of decision diagrams for a prize-collecting scheduling problem," *Computers & Operations Research*, vol. 126, p. 105125, 2021. doi: 10.1016/j.cor.2020.105125

[31] B. Jurisch, "Lower bounds for the job-shop scheduling problem on multi-purpose machines," *Discrete Applied Mathematics*, vol. 58, no. 2, pp. 145–156, 1995. doi: 10.1016/0166-218X(93)E0124-H

[32] T. Achterberg and T. Berthold, "Hybrid branching," *Lecture Notes in Computer Science*, p. 309–311, 2009. doi: 10.1007/978-3-642-01929-6_23

[33] T. Hadzic, J. N. Hooker, B. O'Sullivan, and P. Tiedemann, "Approximate Compilation of Constraints into Multivalued Decision Diagrams," in *Principles and Practice of Constraint Programming*, P. J. Stuckey, Ed. Springer Berlin Heidelberg, 2008, vol. 5202, pp. 448–462. ISBN 978-3-540-85957-4 978-3-540-85958-1. [Online]. Available: http://link.springer.com/10.1007/978-3-540-85958-1_30

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 77–82
DOI: 10.15439/2024F5990
ISSN 2300-5963 ACSIS, Vol. 41

Combining Local and Global Weather Data to Improve Forecast Accuracy for Agriculture

František Koutenský*, Jiří Pihrt*, Miroslav Čepek*, Vojtěch Rybář*, Petr Šimánek*,
Michal Kepka†, Karel Jedlička‡, and Karel Charvát§
*Faculty of Information Technology, Czech Technical University in Prague
Thákurova 9, Praha 6, Czech Republic
Email: {koutefra, pihrtjir, miroslav.cepek, vojtech.rybar, petr.simanek}@fit.cvut.cz
†Lesprojekt s.r.o.
Martinov 197, Záryby, Czech Republic
Email: kepka@lesprojekt.cz
‡Wirelessinfo
Cholinská 1048/19, Litovel, Czech Republic
Email: jedlicka@wirelessinfo.cz
§HELP SERVICE – REMOTE SENSING s.r.o.
Husova 2117, Benešov, Czech Republic
Email: charvat@hsrs.cz

Abstract—**Accurate local weather forecasting is vital for farmers to optimize crop yields and manage resources effectively, but existing forecasts often lack the precision required locally. This study explores the potential of combining data from local weather stations with global forecasts and reanalysis data to improve the accuracy of local weather predictions. We propose integrating the HadISD data set, which contains data from 27 stations in the Czech Republic, with the Global Forecast System predictions and ERA5-Land reanalysis data. Our goal is to improve 24-hour weather forecasts using Multilayer Perceptrons, CatBoost, and Long Short-Term Memory neural networks. The findings demonstrate that combining local weather station data with global forecasts improves the accuracy of weather predictions in specific locations. This advancement holds promise in optimizing agricultural practices and mitigating weather-related risks in the region.**

I. INTRODUCTION

ACCURATE weather forecasting is crucial for farmers to make informed decisions that optimize crop yields and manage resources efficiently. However, available weather forecasts often lack the precision required for agricultural planning, leading farmers to invest in their own weather stations. This study explores the potential of combining data from local weather stations with global forecasts to improve local weather predictions.

As local weather station data could suffer from various inconsistencies, we propose testing this idea by integrating the local weather station data (HadISD data set [1], [2]), with the Global Forecast System (GFS) predictions. In addition, we incorporate ERA5-Land reanalysis data to introduce information on weather conditions in surrounding areas. Our goal is to improve the accuracy of 24-hour weather forecasts by evaluating three machine learning techniques: Multilayer Perceptons (MLP) [3], gradient-boosting regression trees method CatBoost [4], and Long Short-Term Memory (LSTM) [5] neural network.

In this study we focus on the Czech Republic, utilizing data from 27 stations recorded in the HadISD data set within the country and close neighborhood. We supplement this data with GFS forecasts, which provide weather predictions on a 0.25-degree grid resolution, corresponding to an approximately 27.8 km × 27.8 km area in Central Europe. The GFS model predicts various meteorological parameters at different atmospheric levels, offering a comprehensive data set for creating our machine-learning models.

To further improve our predictions, we employ the ERA5-Land data set, renowned for its high-accuracy reanalysis data. Recognizing the latency in the availability of ERA5-Land's data, we trained a U-Net [6] model to map the GFS forecast data to the ERA5-Land's high-resolution grid. This approach enables us to generate ERA5-Land-like predictions in near real-time, potentially enhancing the accuracy of our weather forecasts.

Our methodology involves training and comparing the performance of CatBoost, MLP, and LSTM machine learning techniques against two baseline models, the raw GFS predictions, and the last measured values from the stations. The training data set is constructed using weather data from 2022, pairing each station's observations with corresponding GFS grid data. The models are then validated using data from 2023.

In this paper, we present a detailed analysis of our approach, including data pre-processing, model architectures, training processes, and evaluation metrics. We discuss the performance improvements achieved by integrating ERA5-Land predictions.

Our findings demonstrate that combining local weather station data with global forecasts and incorporating ERA5-

©2024, PTI

Land reanalysis data can substantially improve the accuracy of weather predictions in specific locations. This advancement holds significant promise for optimizing agricultural practices, mitigating weather-related risks, and ultimately enhancing food security in the region.

II. RELATED RESEARCH

Weather forecasting is a well-explored area of research with numerous methodologies and models developed to improve prediction accuracy. The integration of local weather station data with global models is promising in improving forecast precision, particularly in agricultural contexts.

a) Global Forecast Models and Their Limitations: Global forecast models such as the Global Forecast System (GFS) from the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) provide detailed predictions on a global scale. These models use numerical algorithms and vast amounts of atmospheric data to produce forecasts at various temporal and spatial resolutions. There is a large body of literature devoted to improving the global model using in-situ data, (e.g. [7], [8]) However, the coarse resolution of these models often limits their utility for local weather predictions, which are crucial for agricultural decision-making.

b) Machine Learning in Weather Forecasting: Machine learning has become increasingly prominent in weather forecasting due to its ability to handle large data sets and capture complex, nonlinear relationships within the data. Various ML techniques, including neural networks, support vector machines, and ensemble methods, have been applied to enhance forecast accuracy. For example, Multilayer Perceptron (MLP) models have been used to predict temperature [9] and precipitation with notable success. More recently, gradient-boosting algorithms like CatBoost have demonstrated superior performance in regression tasks related to weather prediction [10]. Various concepts have been used for weather forecasting based on weather station data: a 2D-convolutional LSTM in [11], Temporal Convolutional Network (TCN) in [12], and Copulas in [13], to name a few. Large convolutional neural networks (CNNs) were used for global machine learning weather forecasting [14].

c) Hybrid Approaches Combining Global and Local Data: Several studies have explored hybrid approaches that combine global forecast data with local observations to improve prediction accuracy. These methods often involve the use of statistical downscaling or machine learning models to integrate diverse data sources. For instance, a study by [15] demonstrated the effectiveness of combining a global climate model with local weather station observations using an ML model. Similarly, the other works [16]) highlighted the benefits of integrating ERA5 reanalysis data with local meteorological data to refine precipitation forecasts.

d) Application of U-Net for Spatial Predictions: The U-Net [6] architecture, initially developed for biomedical image segmentation, has been adapted for various geospatial applications, including weather forecasting. U-Net's ability to capture spatial hierarchies and produce high-resolution output maps makes it suitable for transforming coarse global forecast data into fine-scale local predictions. Recent studies have successfully employed U-Net to downscale climate model outputs, demonstrating significant improvements in prediction accuracy and spatial resolution [17].

e) Focus on Agricultural Applications: The intersection of weather forecasting and agriculture has been a focal point for research aimed at enhancing food security and optimizing resource management. Accurate local weather predictions can help farmers make timely decisions regarding planting, irrigation, and harvesting, thereby improving crop yields and reducing losses. Multiple research papers ([18], [19]) have emphasized the potential of combining local weather station data with advanced modeling techniques to support precision agriculture.

III. DATA

In this work, we utilize three data sets, each serving a different purpose:

- **HadISD** [2], [1]: This data set is used to extract hourly records from 27 weather stations situated randomly across the area of the Czech Republic. These records contain the variables we aim to predict (temperature, dew point, wind speed), as well as additional variables like cloud coverage, precipitation depth across multiple periods (1h, 2h, ..., 24h), and sea level pressure.
- **GFS** [20]: Unlike HadISD's station-specific data, the GFS (Global Forecast System) data set delivers broader area weather predictions with various frequencies and forecast lead times, encompassing a comprehensive range of atmospheric variables at various altitudes. This feature-rich data set serves both as a baseline for our predictions and as a source for enhancements, utilizing every available feature across all altitude levels.
- **ERA5-Land** [21]: The ERA5-Land data set is a reanalysis tool, meaning it does not provide real-time data but rather offers a retrospective view of land variables over several decades. As a reanalysis data set, ERA5-Land integrates model data with historical observations using the laws of physics to create a globally consistent and comprehensive data set. This characteristic makes it ideal for understanding past climate conditions but limits its use for immediate weather events. We utilize the ERA5-Land data set to train an additional model that can generate features from the GFS data set. The details of this technique will be discussed further in the text.

Having HadISD and GFS data sets, for our experiments, we mapped a GFS rectangle to each HadISD station and merged the data sets accordingly. The final combined data set, consisting of GFS and HadISD data sets, contains 157 features, where 25 of them come from HadISD and 132 of them are from GFS.

We identified the most significant features for our weather prediction task by leveraging the CatBoost model (described later) and computing its SHAP values [22]. These values were

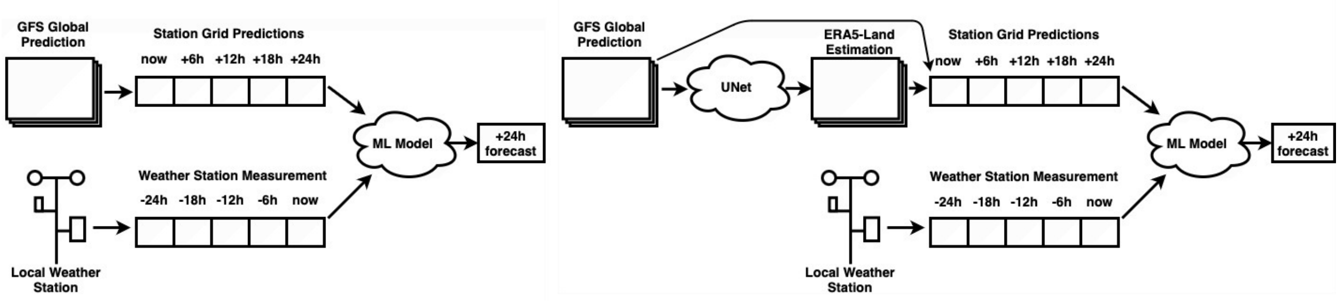


Fig. 1. Illustration of data flows in our experiments. The left schema illustrates a setup where we forecast local weather with the use of weather station data and GFS forecasts. The right schema shows how we improve the accuracy of the forecast with the superresolution model estimating the ERA5-Land reanalysis based on GFS data.

averaged across all target features (temperature, dew point, wind speed) and all stations. Below are the identified features, along with their origins and a brief explanation of each:

- **Helicity** (GFS, height above ground layer): Measures the potential for rotation in the atmosphere, which can be important for predicting severe weather events.
- **Temperature** (GFS, surface): A fundamental parameter, that influences various atmospheric processes and weather conditions.
- **Precipitable water** (GFS, atmosphere single layer): Representing the total atmospheric water vapor, is essential for forecasting precipitation and humidity levels.
- **Dew point** (HadISD): Dew point temperatures from station data indicate the atmospheric moisture content, aiding in humidity and fog predictions.
- **Minimum temperature** (GFS, height above ground): Minimum temperature above ground level might help in identifying cold spells and frost conditions.
- **Temperature** (HadISD): Observed temperatures from station data, a direct measurement of local weather conditions, is crucial for accurate forecasting.

IV. EXPERIMENT DESIGN

The model takes the last 24 hours of weather data from the station and the weather forecast for the next 24 hours for the corresponding grid. Based on this, the model creates a prediction of selected weather parameters 24 hours from now. The left side of Figure 1 illustrates the idea.

The training data set is constructed from weather data from selected weather stations combined with its GFS rectangle for the year 2022. The next step is to split the whole history of station data and GFS predictions into time windows containing inputs and corresponding target values from the HadISD station. For the validation of the model, we used data from 2023.

As the experimental results below show, the models can improve the accuracy of the local prediction. Since the GFS data set has low spatial resolution and is known to have limited accuracy in predictions, we decided to explore possibilities to incorporate another data set, with better accuracy and higher spatial resolution. The ERA5-Land data set [21], [23], which is suitable for our case, is a reanalysis of past weather conditions and it is assessed to be a good approximation of the actual weather. However, the ERA5-Land predictions are available with considerable delay and thus it is not possible to use it directly.

To tackle this issue, a model to estimate the ERA5-Land values was developed. We use a subset of the GFS data set as coarse grid input and a U-Net architecture to create a fine-grained grid estimating ERA5-Land. In the training process of the U-Net superresolution model, GFS predictions and corresponding ERA5-Land data for the years 2015 to 2021 are utilized. The data were split time-wise to prevent information leaks. The earlier 80% are used for training and the remaining 20% are used for validation. The incorporation of the ERA5-Land data set into our AI forecast is illustrated on the right side of Figure 1.

V. METHODOLOGY

Our experiments can be divided into two stages. The aim of the first stage was to train a model that could transform GFS predictions to be closer to ERA5-Land. The variables we focused on included temperature, dew point, and the u and v components of wind.

We approached this problem as an image-to-image translation task. For this purpose, we employed the U-Net [6] convolutional neural network. The name is inspired by its U-shaped architecture shown in Figure 2. It can be described as a symmetrical encoder-decoder architecture, consisting of a contracting path to capture context and a symmetric expanding path that enables precise localization, further enhanced by the usage of skip/shortcut connections.

First, GFS data are interpolated to match the grid of ERA5-Land. Both are then cropped to 128x64 pixels, which is enough to cover the Czech Republic, and the dimensions are divisible by 16, which is required by U-Net. The variables are concatenated in the channel dimension and normalized. U-Net is trained with 4 input channels and 4 output channels, as well as other hyperparameters listed in Table I. Example prediction is shown in Fig. 3. The output of a trained U-Net is then denormalized, and time series for each station is generated by interpolating the variables at specific coordinates of the station.

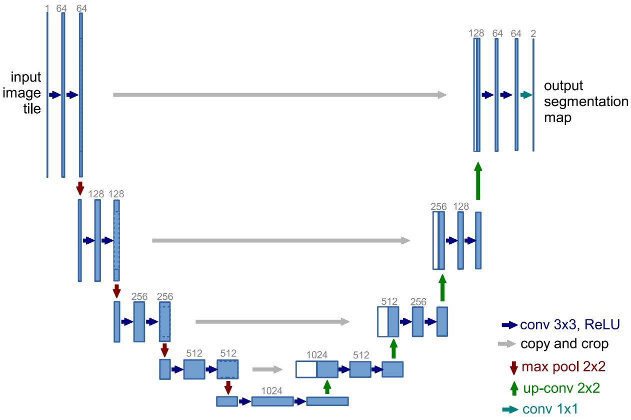


Fig. 2. U-Net architecture [6] used to estimate ERA5-Land like weather forecasts from GFS predictions. The exact hyperparameters of the U-Net architecture are outlined in Table I.

TABLE I
U-NET TRAINING HYPERPARAMETERS

Hidden dimensions	[32, 64, 128, 256, 512]
Batch size	16
Optimizer	Adam
Learning rate	1×10^{-3}
Loss function	MAE

The goal of the first stage was to create more reliable features by utilizing the GFS and ERA5-Land data sets for subsequent weather forecasts at specific stations. This process effectively generates 4 new features, which are then incorporated into the data set used for station forecasts.

The second stage involved comparing models on our enhanced station data set, which was augmented by the U-Net-generated ERA5-Land predictions. Since we are dealing with time series data, we utilized the LSTM (Long Short-Term Memory) network [5], known for its ability to capture long-term dependencies and patterns in sequential data. The LSTM model is based on a sequence-to-sequence (seq2seq) architecture. The model consists of an encoder that encodes the input time series sequence into a fixed-length vector representation and a decoder that generates the predicted output sequence based on the encoded representation. The encoder takes an input sequence of a given length and produces hidden states and cell states that capture the temporal dependencies in the data. The decoder takes the last hidden state of the encoder as its initial hidden state and generates the output sequence recursively. The input sequence consists of four previous time steps of a 10-time series, selected based on SHAP [24] analysis of a CatBoost model. The LSTM seq2seq model takes this input and regressively generates predictions for the next four time steps, effectively making a 24-hour forecast for the desired quantities.

Alongside LSTM, we also examined MLP (Multi-Layer Perceptron) [3] and CatBoost [4] models. MLP, a class of feedforward artificial neural networks, consists of multiple layers of nodes, with each node connected to every node in the subsequent layer. It is particularly adept at capturing complex relationships in the data through its dense connections and non-linear activation functions. MLPs are typically composed of an input layer, one or more hidden layers, and an output layer. Each node (or neuron) in a layer applies a weighted sum of the inputs followed by a non-linear activation function, which allows the model to learn and represent complex functions. MLPs are effective in scenarios where the relationship between inputs and outputs is highly non-linear and intricate, making them suitable for various predictive modeling tasks in weather forecasting.

CatBoost, a gradient-boosting algorithm, is well-suited for this task because it handles categorical data well and mitigates the problem of overfitting. It builds an ensemble of decision trees where each new tree is trained to correct the errors of the previous ones, leading to improved accuracy. CatBoost is particularly effective in scenarios where the relevance of historical data might vary, providing robust predictions even when the importance of past data fluctuates. One of its key advantages is its ability to handle categorical features natively, without requiring extensive preprocessing or encoding, which simplifies the model training process and enhances performance. Additionally, CatBoost incorporates ordered boosting and other advanced techniques to reduce overfitting and improve the generalization of the model.

For each of these three models, we developed two versions: (a) one trained on GFS and station data, and (b) another trained on GFS, ERA5-Land predictions, and station data. This protocol was designed to assess the impact of ERA5-Land predictions, which we hypothesized might play a crucial role in enhancing the accuracy and reliability of the weather forecasts. The Figure 1 illustrates both versions.

VI. EXPERIMENTAL RESULTS

Table II summarises our results. We present results for three models in two versions, as mentioned above. In addition, two additional baseline techniques were added for comparison: using the GFS forecast directly and assuming the weather in 24 hours will be the same as the current conditions. The forecast results are presented for three weather parameters: Temperature, Dew Point, and Wind Speed, all measured at a height of 2 meters above ground. The tables show the mean absolute error (MAE) of all examined methods for all target parameters. The metric is calculated over all 27 selected weather stations for predictions in the year 2023.

Table II shows that the direct GFS forecast has the second worst accuracy. It is on average more than $2°C$ off the actual measured temperature. Using actual current weather achieves slightly worse accuracy with $2.5°C$ average absolute error. Similar results are seen for the Dew Point and Wind Speed.

The results for our models show that we can greatly improve over the GFS forecasts as well as the last value prediction. The models built on top of the GFS predictions and weather recorded by a weather station achieve better accuracy. The best model, CatBoost, achieves mean absolute errors of $1.07°C$ in Temperature, $1.04°C$ in Dew Point, and $1.01\ m/s$ in

TABLE II
ACCURACY OF PRESENTED METHODS AND COMPARISON TO BASELINE MODELS. THE VALUES SHOW THE MEAN ABSOLUTE ERROR BETWEEN THE MODEL FORECAST AND ACTUALLY MEASURED VALUE. ALL FORECASTS ARE 24 HOURS INTO THE FUTURE.

Method	**Temperature** 2m above ground MAE [$°C$]	**Dew Point** 2m above ground MAE [$°C$]	**Wind Speed** 2m above ground MAE [m/s]
Last value (persistence)	2.53	2.54	1.83
Direct GFS Prediction	2.18	1.38	2.04
LSTM with GFS	1.20	1.14	1.12
MLP with GFS	1.16	1.13	1.09
CatBoost with GFS	1.07	1.04	1.01
LSTM with GFS and estimated ERA5-Land	1.15	1.23	1.12
MLP with GFS and estimated ERA5-Land	1.14	1.15	1.05
CatBoost with GFS and estimated ERA5-Land	**1.06**	**1.02**	**1.01**

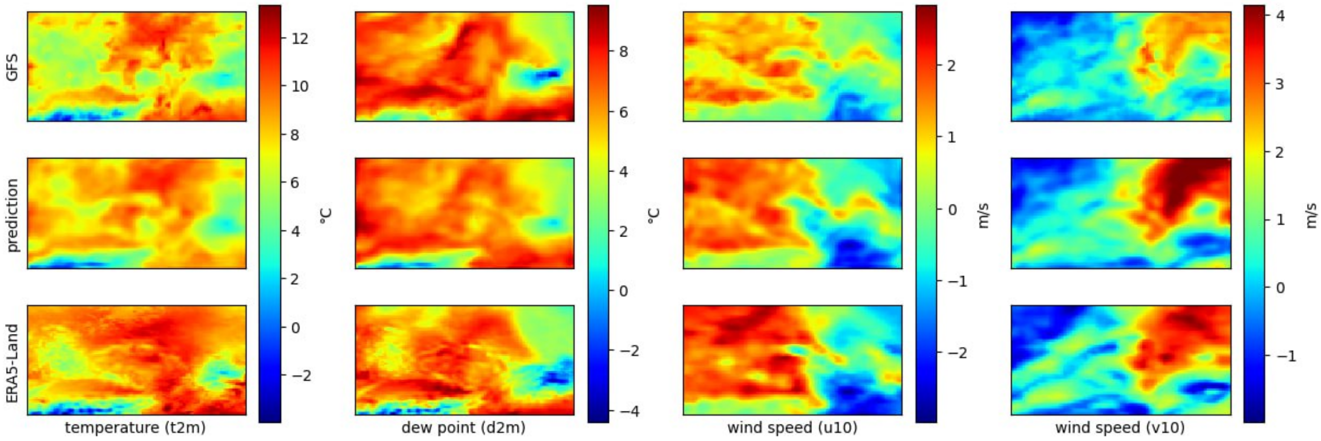


Fig. 3. Example of ERA5-Land data estimated by U-Net superresolution model from GFS predictions. The top row shows the GFS forecast, the middle row is the output of our U-Net model and the bottom row is ground truth ERA5-Land data.

TABLE III
MAE OF GFS FORECASTS AND U-NET MODEL ESTIMATING ERA5-LAND

Variable	GFS to ERA5-Land	Estimated ERA5-Land to ERA5-Land
Temperature [$°C$]	3.03	1.10
Dew Point [$°C$]	2.77	1.09
Wind Speed E-W [m/s]	1.47	0.46
Wind Speed N-S [m/s]	1.51	0.47

Wind Speed. Which represent around 50% improvement for Temperature, Dew Point, and Wind Speed respectively over GFS predictions.

The models incorporating the estimated ERA5-Land data set achieve slightly better results. Similar to no ERA5-Land data set, the CatBoost model achieves the lowest error. The mean absolute error for the Temperature is $1.06°C$ for the 24-hour forecast. The Dew Point forecasts show a mean absolute error of $1.02°C$ and the Wind Speed forecasts show a 1.05 m/s error.

Comparing the three selected ML-based models, the best accuracy is achieved by the CatBoost model. The difference in accuracy between MLP and LSTM is not so significant.

Figure 3 illustrates the results of the U-Net architecture when estimating the ERA5-Land weather reanalysis from GFS forecasts. We illustrate the results using the temperature, the dew point, and wind speed. The wind speed is shown as south-north and east-west components of the speed vector. The top row shows the original GFS forecasts, the bottom row shows the ERA5-Land targets and the middle row represents the U-Net estimate. Table III shows the mean absolute error between our estimated ERA5-Land values and the ground truth values. For comparison, the table also shows the mean absolute error between GFS and ground truth ERA5-Land values. The numbers show about 60% to 70% improvement in the estimation of the ERA5-Land value. This improvement supports the results presented earlier when the introduction of estimated ERA5-Land data improved the accuracy of the forecast.

VII. CONCLUSION

The experimental results confirm the viability and effectiveness of the proposed methodology in generating highly accurate localized forecasts. The AI-driven 24-hour predictions, which integrate GFS data with local measurements, demonstrate markedly superior accuracy compared to GFS alone. This enhanced precision empowers farmers to refine their planning processes, potentially leading to improved crop yields, more efficient fertilizer application, and strengthened food security.

Our findings indicate that the inclusion of estimated ERA5-Land data does not contribute significantly to model accuracy improvements.

Moving forward, research efforts will concentrate on minimizing the volume of historical data required from weather stations. These refined techniques will be implemented across approximately 200 weather stations in Czech and Slovakian vineyards[1].

ACKNOWLEDGMENT

This project is funded with state support from the Technology Agency of the Czech Republic and the Ministry of Industry and Trade of the Czech Republic under the TREND Program.

REFERENCES

[1] R. J. H. Dunn, K. M. Willett, D. E. Parker, and L. Mitchell, “Expanding hadisd: quality-controlled, sub-daily station data from 1931,” *Geoscientific Instrumentation, Methods and Data Systems*, vol. 5, no. 2, pp. 473–491, 2016. [Online]. Available: https://gi.copernicus.org/articles/5/473/2016/

[2] D. R. J. H., “Hadisd version 3: monthly updates,” Hadley Centre Technical Note, Tech. Rep., 2019.

[3] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[4] A. V. Dorogush, V. Ershov, and A. Gulin, “Catboost: gradient boosting with categorical features support,” 2018.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation.”

[7] D. Salisbury, K. Mogensen, and G. Balsamo, *Use of in situ observations to verify the diurnal cycle of sea surface temperature in ECMWF coupled model forecasts*. European Centre for Medium Range Weather Forecasts, 2018.

[8] G. Balsamo, A. Agusti-Parareda, C. Albergel, G. Arduini, A. Beljaars, J. Bidlot, E. Blyth, N. Bousserez, S. Boussetta, A. Brown *et al.*, “Satellite and in situ observations for advancing global earth surface modelling: A review,” *Remote Sensing*, vol. 10, no. 12, p. 2038, 2018.

[9] M. He, J. Qin, N. Lu, and L. Yao, “Assessment of era5 near-surface air temperatures over global oceans by combining modis sea surface temperature products and in-situ observations,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[10] A. Callens, D. Morichon, S. Abadie, M. Delpey, and B. Liquet, “Using random forest and gradient boosting trees to improve wave forecast at a specific location,” *Applied Ocean Research*, vol. 104, p. 102339, 2020.

[11] D. Kreuzer, M. Munz, and S. Schlüter, “Short-term temperature forecasts using a convolutional neural network — an application to different weather stations in germany,” *Machine Learning with Applications*, vol. 2, p. 100007, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827020300074

[12] P. Hewage, A. Behera, M. Trovati *et al.*, “Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station,” *Soft Computing*, vol. 24, no. 24, pp. 16 453–16 482, 2020. [Online]. Available: https://doi.org/10.1007/s00500-020-04954-0

[13] S. C. Maina, D. Mwigereri, J. Weyn, L. Mackey, and M. Ochieng, “Evaluation of dependency structure for multivariate weather predictors using copulas,” *ACM J. Comput. Sustain. Soc.*, vol. 1, no. 2, dec 2023. [Online]. Available: https://doi.org/10.1145/3616384

[14] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, “Accurate medium-range global weather forecasting with 3d neural networks,” *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[15] G. J. Anderson and D. D. Lucas, “Machine learning predictions of a multiresolution climate model ensemble,” *Geophysical Research Letters*, vol. 45, no. 9, pp. 4273–4280, 2018. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL077049

[16] M. Cucchi, G. P. Weedon, A. Amici, N. Bellouin, S. Lange, H. Müller Schmied, H. Hersbach, and C. Buontempo, “Wfde5: bias-adjusted era5 reanalysis data for impact studies,” *Earth System Science Data*, vol. 12, no. 3, pp. 2097–2120, 2020. [Online]. Available: https://essd.copernicus.org/articles/12/2097/2020/

[17] Y. Sha, D. J. Gagne II, G. West, and R. Stull, “Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part ii: Daily precipitation,” *Journal of Applied Meteorology and Climatology*, vol. 59, no. 12, pp. 2075–2092, 2020.

[18] A. Bramantoro, W. S. Suhaili, and N. Z. Siau, “Precision agriculture through weather forecasting,” in *2022 International Conference on Digital Transformation and Intelligence (ICDI)*. IEEE, 2022, pp. 203–208.

[19] C. El Hachimi, S. Belaqziz, S. Khabba, and A. Chehbouni, “Towards precision agriculture in morocco: A machine learning approach for recommending crops and forecasting weather,” in *2021 international conference on digital age & technological advances for sustainable development (ICDATA)*. IEEE, 2021, pp. 88–95.

[20] National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, “Ncep gfs 0.25 degree global forecast grids historical archive,” Boulder CO, 2015.

[21] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach *et al.*, “Era5-land: A state-of-the-art global reanalysis dataset for land applications,” *Earth system science data*, vol. 13, no. 9, pp. 4349–4383, 2021.

[22] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

[23] M.-S. J., “Era5-land hourly data from 1950 to present,” Copernicus Climate Change Service (C3S) Climate Data Store (CDS), Tech. Rep., 2019.

[24] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” *CoRR*, vol. abs/1705.07874, 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

---

[1] https://ekovin.cz/

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 83–94
DOI: 10.15439/2024F5992
ISSN 2300-5963 ACSIS, Vol. 41

Exploring Linguistic and Cultural Differences in Online Job Advertisement Analysis for NLP Applications

Lea Grüner
German Federal Institute
for Vocational Education and Training
Email: lea.gruener@bibb.de

Kai Krüger
German Federal Institute
for Vocational Education and Training
Email: kai.krueger@bibb.de
ORCID: 0009-0000-0596-3621

***Abstract*—Online job advertisements (OJAs) have become a significant data source for analyzing labor market dynamics, offering insights into shifts within occupations, industry sectors, skills, and tasks. This paper investigates the cross-lingual and cultural differences in OJAs and their impact on the transferability of Natural Language Processing (NLP) methods and research scope. By analyzing OJAs from Austria, France, Germany, Italy, Spain, the UK, and the US, we point out substantial variations in document length, diversity metrics, syntactic structures, and content features such as salary information. These differences underscore the challenges in applying NLP methods universally across languages and cultures. Our findings emphasize the need for tailored approaches in NLP research and offer a starting point for developing standardized pipelines for analyzing text genres across different languages.**

I. INTRODUCTION

ONLINE Job advertisements (OJAs) have garnered significant attention from researchers across various fields, including social sciences, economics, and computational linguistics. Studies utilizing OJAs have explored labor market trends, skill demands, and occupational shifts [1], [2], [3], [4], [5], [6]. Additionally, OJAs offer insights into demographic targeting and potential discrimination, making them critical for research in Human Resources (HR) and gender studies [7], [8], [9]. The digital migration of job ads has enhanced their accessibility, prompting the development of Natural Language Processing (NLP) methods to process and analyze these texts. However, given the diverse linguistic and cultural contexts of OJAs, it is essential to investigate how these differences affect the transferability of NLP methods and research findings.

The main contributions of our exploratory study are:

1) **Cross-Lingual Variation in Job Advertisements**: We identify and quantify[1] substantial differences in language, structure, and content in job advertisements across seven countries, emphasizing the need for localized approaches in NLP research.
2) **Impact on NLP Method Transferability for Downstream Tasks**: The study highlights how cross-lingual and cultural variations in the text genre job ads can hinder the direct transfer of NLP methods, providing tangible evidence and metrics to support this claim. We discuss how this impacts other text genres as well.
3) **Comprehensive Data Collection and Analysis Pipeline**: We present a robust methodology for scraping OJAs from multiple countries using consistent data sources, ensuring high comparability and reliability of the results.

By exploring the form and content variations in OJAs from different countries and languages, this paper aims to identify potential pitfalls in cross-lingual and -cultural research, providing a foundation for more effective and nuanced NLP applications in the context of global applied NLP.

II. JOB ADVERTISEMENTS AS RESEARCH SUBJECTS

In recent years, the interest in job advertisements (job ads) has increased among a diverse group of actors, including researchers from fields such as social sciences and economics, as well as government agencies and private corporations, who use job ads as a data source to gain insights into labor market dynamics. These insights include shifts in occupations [1], [2], industry sectors, skills [3], [4] and tasks [5], [6]. Beyond these aspects, job ads can also provide perspectives on how various demographic groups are targeted or potentially discriminated against, offering critical data for HR or gender studies [7], [8], [9] research. Furthermore, job ads have been used to study how employers attempt to attract candidates, for example by analyzing benefits they offer [10] or the way they present their corporate identity [11], [12]. Research on job ads has been performed in many different countries and on job ads of many different languages, including, among others, all EU states and the UK [13], [14], Brazil [15], Canada [16], China [7], Japan [17], Mexico [15], Taiwan [18] and the United States [19].

While some of the mentioned studies were of qualitative nature, the rise of big data in recent years and the digital migration of job advertisements have significantly enhanced their accessibility as a data source, which motivated the development of NLP methods to structure and analyze these texts. This study aims to investigate how linguistic and cultural

---

[1] Our code and supplementary plots can be accessed here: https://github.com/TM4VETR/linguistic-cultural-differences-OJAs/

©2024, PTI

differences in job ads impact the transferability of NLP methods and research findings across different languages and cultures.

III. JOB ADVERTISEMENTS AS TEXT GENRE

From a linguistic perspective, job ads can be considered a text genre [20], characterized by their communicative purpose [21], the recurrence of situations they address, the discourse community that produces them, and their primary audience [22]. Thus, the actors working with these texts have certain expectations about the texts' content and linguistic characteristics based on their genre knowledge [23]. However, some studies on other text genres have shown that these assumptions might not always apply to all languages and cultures [24], [25], [26], [27], [28], [29], [30], even though in other studies such evidence has not been found [31]. Studying cross-lingual and -cultural differences within a text genre has also gained great attention by researchers from translation studies [20], [32], [33], [34]. In NLP, however, research on cross-lingual text genre has instead focused on getting models to transfer and adapt genre-specific language information from resource rich languages to resource low languages. We argue that practitioners in applied NLP working with specific text genres can benefit from analyzing cross-lingual and cultural differences within their text genre. Our core argument is that researchers may falsely assume that findings derived from one language or cultural context apply universally within the same genre. This can unfold in two ways:

1) **NLP Methods:** Researchers assume their NLP methods for processing job ads are universally applicable. This could include the usefulness/ translation of word lists, structural patterns for rule based systems or the potential, limitations, strengths and weaknesses of Machine Learning with encoder-based models like BERT [35] or autoregressive models such as GPT [36].
2) **Research Scope:** Given the diverse purposes for which job ads are used, it is unclear whether all information expected to be included in job ads actually is included across countries.

IV. RESEARCH QUESTION

Derived from the observations described in the previous two chapters, our central research question is: **Can differences in the form or content of OJAs potentially hinder the transferability of NLP methods or research scope?** To answer this question, we perform quantitative analyses on OJAs from different countries and languages. Our analyses focus on different aspects of OJAs and were designed partly to explore the linguistic differences between the data and partly with regard to specific features derived from potential pitfalls. We chose this mixed approach, because we believe both types of analyses might be beneficial to researchers. Analysing cross-lingual or -cultural differences with regard to a specific pitfall can help uncover and overcome it. However, it is impossible for researchers to identify all such pitfalls exhaustively beforehand, which is why using general descriptive analyses of the language data might help uncover additional instances. We describe our exact methods in Section VI and indicate whenever we use a certain analysis with respect to a specific pitfall.

V. DATA

Since no publicly available, comparable job ad dataset across multiple languages exists, we decided to collect our own data by scraping[2]. We faced the challenge that country-specific differences could be influenced by factors like occupation or industry sector, acting as confounding variables. Classifying ads into the respective taxonomies would require complex models and normalization, beyond this project's scope. Our exploratory research indicated that different websites target different audiences, affecting the types of employers and employees. To mitigate this, we scraped data from the same website operating across multiple countries. We scraped data from CareerJet[3] from Austria, France, Germany, Italy, Spain, the United Kingdom, and the United States.

We chose these countries specifically for several, mostly practical reasons. Firstly, we aimed to include different countries sharing the same main language to see whether differences were merely linguistic or also cultural. Subsequently, we have chosen two German and two English speaking countries. Then, for our analyses we make use of existing NLP libraries and models. Especially the use of multilingual models when performing tokenization or NLI analyses (see Section VI) required us to limit ourselves to languages that these models have been pre-trained on. Additionally, since we use Zero-Shot models for our analyses, we wanted to focus on languages that our team members had access to, to be able to manually examine the performance of these models for some basic sanity checks. Lastly, the countries mentioned are all rather large and strong in research, which ensured that there would be enough data, and with regard to our research question it is likely that methods for OJA processing will be developed on data from these countries. The choice of countries, however, is somewhat vulnerable as we discuss in Section IX.

Our target was to gather a dataset of 10,000 OJAs from each country. We argue that this number is sufficiently large to provide meaningful insights into the research questions while remaining computationally manageable considering some of the applied analyses are quite resource intensive, for example the vendi-score calculation (Section VI). We scraped slightly more data in case some of the scraping results were corrupted and randomly sampled to get the desired amount. Data scraping for all countries was completed within a week, with one or two days dedicated to each country. Therefore, our data is also comparable with regard to origin time.

In addition to the job ad data scraped from CareerJet, we ran our linguistic analyses on the Wikilingua dataset [37] for comparison where applicable. This dataset contains a compilation

---

[2] We are working on publishing our data in compliance with data privacy regulations.

[3] https://www.careerjet.com/, accessed April 2024.

of WikiHow[4] articles and summaries in 18 different languages. Using a reference dataset is useful to distinguish between language and genre related features and differences to some extent. We chose the Wikilingua dataset because it is available for all examined languages (English, German, French, Italian, Spanish), easily accessible, and the texts are of suitable length. To run the analyses, 10.000 Wikilingua articles per language are sampled. Noteworthily, the contents of these articles may partly overlap between languages, since the Wikilingua data contains translations of English articles.

VI. METHOD

We executed an analysis pipeline with each of the seven data splits from Austria, France, Germany, Italy, Spain, the UK, and the USA to compare job ad data across countries and languages. The examined features include structural, linguistic and job-ad related aspects, precisely document length, lexical and structural diversity, Part-Of-Speech (POS) tags, paragraph and list count and length, language detection, and presence of salary information as a content feature. All aspects except the latter three were analyzed using the Wikilingua [37] reference dataset as well. In the following, the specific measures used to determine the aforementioned features are described.

Document length: The mean, median, and standard deviation of document length in tokens were calculated for each country split. To get the length in tokens, we used two different tokenization methods. The first one is simple white-space splitting, with additionally separating symbols and punctuation and counting them as single tokens. Since pre-trained language models such as BERT [35] use Wordpiece tokenization, we used the BertTokenizer for multilingual BERT [35] for splitting documents as well. This results in higher token counts because tokens are split into sub-tokens. Document length in tokens can be highly relevant for processing in pre-trained language models since BERT-like models usually truncate texts after 512 tokens.

Type-token ratio: The type-token ratio (TTR) is a simple measure for lexical diversity of a text (corpus). It is calculated by dividing the number of types (unique tokens) by the number of tokens (all tokens in a text). Since the TTR is sensitive to text length, we used the standardized TTR (STTR) that computes the TTR for each window w of n tokens and averages over all windows W.

$$STTR = \frac{\sum_{w \in W} \frac{count(types)}{count(tokens)}}{count(w)}. \tag{1}$$

The STTR is a value between zero and one. Values close to one indicate high lexical diversity, lower values point to less diversity. In our experiments, the STTR was calculated with a window size of n=1000. For each country corpus, it was calculated using both on white-space tokenized texts, and lemmas obtained from language models from the SpaCy NLP pipeline [38]. Especially for morphologically rich languages like German, lemmas are more insightful for measuring lexical diversity.

Vendi-score: The vendi-score (VS) [39] is a diversity metric for Machine Learning that can be used for a broad range of matrix-based data types. It is defined as

$$VS_k(x_1, ..., x_n) = exp(-\sum_{i=1}^{n} \lambda_i log \lambda_i) \tag{2}$$

where $x_1, ..., x_n$ is a collection of n samples, k is a pairwise similarity function with the kernel matrix K, and $\lambda_i, ..., \lambda_n$ denote the eigenvalues of K/n. Higher values indicate more diversity within a collection of samples.

We calculated the n-gram VS with n ranging from one to three to measure lexical-structural diversity within a data split. Additionally, the embedding VS with contextual embeddings obtained from multilingual BERT [35] was used to assess semantic diversity. The similarity function k is the cosine similarity for both n-gram and embedding VS.

Generally, diversity on different linguistic levels can offer insights into how complex the downstream task is. This could, for example, be used as an indicator of how much evaluation data is required. Examining diversity is particularly important for approaches that are evaluated on the basis of very few manual examples, e.g. prompting [40], other zero- or few-shot methods [41] or synthetic job ads [42].

POS-tags and POS n-grams: To get insights into the linguistic structure of the texts, we used POS-tags obtained using the SpaCy NLP pipeline [38]. For each country, the frequency of POS-tags was determined. If the frequency distribution of tags varies a lot between countries and datasets, this may point to language- or genre-specific differences in job ad data. It is also interesting to examine whether there are significant discrepancies between the two same language country pairs (Germany-Austria and UK-US). In addition, we count n-grams of POS-tags with n=2 and n=3. Again, this could provide evidence for language- or genre-specific contrasts. For instance, the comparison of POS n-grams allows conclusions on syntactic differences, which can be important for analysis systems based on POS-patterns.

Considering the differences in linguistic structures can be relevant specifically for the transferability of syntax based methods. For example, [6] use verb object pairs to extract tasks from OJAs. While this is reasonable for English job ads, an exploratory analysis of German job ads showed that instead of using such verb-object pairs these tasks were often expressed as compounds (e.g. *Kundenberatung*).

Paragraph/ List count and length: Scraping data from CareerJet allowed us to maintain structural HTML information from the website such as linebreaks and listings (ordered and unordered lists). It is worth mentioning, however, that this approach possibly misses out on some listings, since some may not be appropriately formatted in HTML. We created paragraphs based on linebreaks and lists (a list is counted as one paragraph). The paragraphs and lists were counted per document, and the average count per document was calculated for each country split. In addition, we computed the average paragraph length in whitespace-tokens and the average list

---

[4] https://www.wikihow.com/Main-Page, accessed April 2024.

length in list items.

This analysis might offer insight into structural differences between job ads from different countries. This is relevant, because the structure of job ads is frequently used in the OJA analysis pipeline. For example, researchers performed text zoning to identify different segments within the job ads such that specific aspects like skills were only extracted from relevant segments [43], [44], [45]. By counting the amount of list items we hoped to gain insight into how many skills, tasks and benefits can on average be found in OJAs of a given country. We assumed that lists almost exclusively list one of these three entities, i.e. not mix these entities or contain information about other topics. One downside of our current approach is, however, that it lacks the ability to classify these entities.

Language Detection: We used an XLM-RoBERTa-base pretrained language detection model [46] to identify job ads not written the respective country's main language. The model is trained to distinguish 21 languages, including the five examined (English, French, German, Italian, Spanish). Determining the percentage of foreign language texts and the languages present in a dataset is crucial for applying NLP methods, since many approaches are built for a single particular language. Choosing a multilingual method, such as a suitable pretrained multilingual model, might be necessary in some cases. Social science researchers in particular are usually interested in investigating the entire country, regardless of the language. Moreover, detecting the languages used in job ads could also provide interesting insights into the target group to be addressed by the employer. For instance, if many job ads from a non-english speaking country are written in English, this could point to a particular interest in international candidates.

Salary information: To determine the presence of salary information in a job ad we used two indicators. First, we extracted information from a text field of the website by storing the corresponding HTML-tag information in our database. This information usually consists of precise numbers such as the hourly, monthly or yearly salary. Second, a zero-shot classification approach based on a multilingual mDeBERTa model trained for Natural Language Inference (NLI) [47] is used. Each paragraph of a job ad was used as a premise tested against the hypothesis *'The line contains information on salary.'*. If the model yielded an entailment probability ≥ 0.9, the paragraph was marked and the whole job ad was labeled with *true*. For each country split, the relative number of job ads containing salary information was computed.

We consider our investigation of salary information as a means to test how research scope varies between cultures, as outlined in Section III. Previous studies have leveraged salary data to analyze phenomena such as the impacts of introducing minimum wage [48]. Therefore, researchers in other countries, where the minimum wage has been implemented, might be interested in using this study as a reference.

VII. Results

In this section we present our results. We focus on the findings most relevant to our discussion, supplementary results and plots can be accessed in our repository.

Document length: Figure 1 plots mean and standard deviation of document length based on the two methods described in Section VI. Overall, there are more tokens using the BERT tokenization, which is expected given the subword tokenization. The majority of country-wise proportions are comparable for both tokenization methods, although some discrepancies can be observed. For example, German, Austrian and French texts are proportionally longer with the BERT tokenization.

It is striking that the UK and the US not only have the longest OJAs, but are also the only countries where the job ads are on average longer than the Wikilingua texts. US ads are longest with a mean length of 740 BERT tokens while Italian ads are shortest with a mean length of 369 tokens, followed by Spain. German, Austrian and French OJAs display similar mean lengths around 550 tokens. For all countries except the UK and the US, texts from the Wikilingua reference dataset are notably longer on average. French Wikilingua texts are longest with a mean length of 728 tokens. In general, the Wikilingua data shows less discrepancies between languages than the OJA data does between countries. However, one possible explanation is the content-wise overlap in Wikilingua data.

Diversity Metrics: STTR, n-gram VS, and embedding VS are all metrics indicating different aspects of corpus diversity. Therefore, we use Table I to give an overview of text diversity in our datasets. We can see that consistently OJA data is more diverse in terms of STTR than the respective Wikilingua counterpart. For both VS metrics we observe the opposite. The Wikilingua data is more diverse here. For STTR, the difference is highest for Austria and lowest for Italy. For n-gram VS it is highest for Italy and lowest for Germany, whereas for embedding based VS it is highest for the UK and lowest for Germany.

The within metric differences between countries are generally low to moderate with a few exceptions. Overall, we see that French and Spanish data is less diverse across most metrics compared to other country splits, whereas German, Austrian and Italian OJA data is rather diverse.

TABLE I Comparison of different text diversity metrics across countries, plus diversity metrics of Wikilingua data for reference.

Country	**STTR↑**		**VS_ngram↑**		**VS_embedding↑**	
	OJA	**Wiki**	**OJA**	**Wiki**	**OJA**	**Wiki**
AUT	.45	.36	275	331	1.42	1.56
DE	.43	.36	284	331	1.4	1.56
ES	.38	.34	242	379	1.39	1.61
FR	.37	.32	164	311	1.34	1.50
IT	.41	.38	280	542	1.4	1.66
UK	.40	.34	240	397	1.19	1.46
US	.41	.34	234	397	1.21	1.46

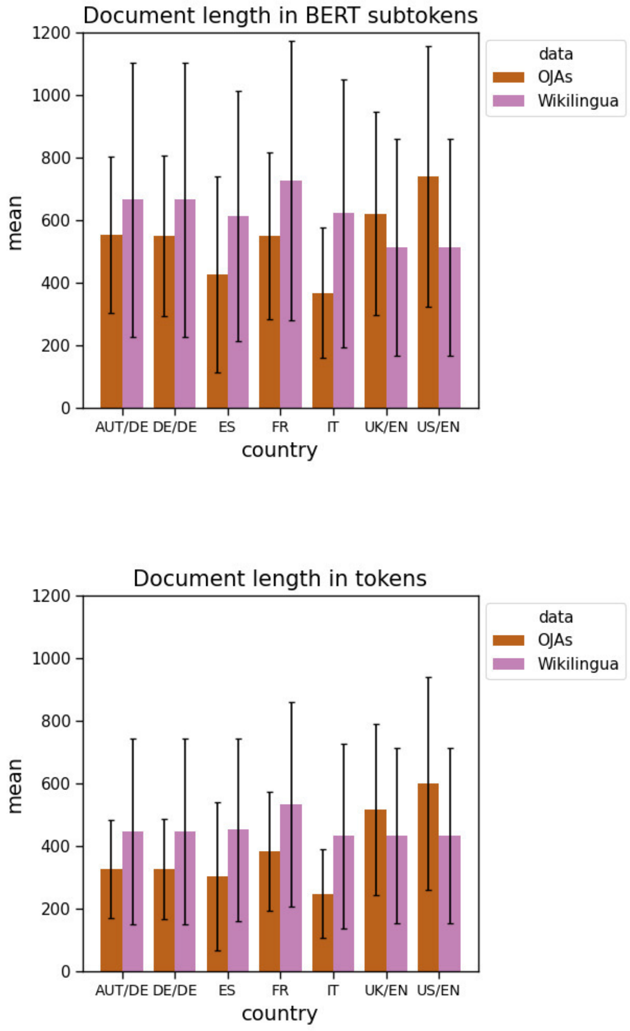


Fig. 1: **Document length.** Upper plot shows mean token count based on BERT subtokens, lower plot based on simple whitespace tokenization. Comparison between countries (x-axis) and datasets indicated by colors.

POS analysis: We analyzed the distribution of different POS types using radar plots as demonstrated in Figure 2 to quickly access the differences between countries with regard to selected POS types. We chose the POS tags ADJ, ADV, AUX, ADP, NOUN, VERB, PRON and PROPN, because we argue that these, mostly containing content words, are most relevant to NLP method design. The most relevant findings from the comparison of OJAs from different countries are:

- In general, for all country splits, there is a major overlap in the distribution of the POS tags examined.
- Nouns have the highest frequencies across all countries.
- Some differences appear to depend on the language family. The Romance languages Spanish, French and Italian have more adpositions (ADP) than the Germanic languages English and German.
- Although the discrepancy is not very profound, data from the English-speaking countries (UK and US) shows an increased occurrence of verbs compared to all other languages. OJAs from German-speaking countries Austria and Germany have the lowest proportion of verbs which may be due to the popularity of substantivations in German.

The corresponding radar plots comparing the two datasets per country can be found in the supplementary material accessible in our repository. The most relevant observations are summarized in the following:

- All countries have more nouns in OJAs than in the Wikilingua data, except from Spain where it is about equal.
- All countries have less verbs in OJAs than in the Wikilingua data.
- All countries have more proper nouns in OJAs than in the Wikilingua data, although the discrepancy is greater for Spain than it is for other countries, like Germany.
- OJAs also tend to have more adjectives and adpositions, whereas Wikilingua data has more auxiliaries and adverbs, although the differences are not very large for the most part. Wikilingua data also has substantially more pronouns.

With regard to bi- or trigrams, the analysis becomes even more complex given the large number of patterns. This makes it difficult to choose individual patterns for comparison. Therefore, we employed Principal Component Analysis (PCA) [49] to identify key syntactic structures that significantly contribute to variations among datasets. PCA effectively reduces dimensionality, transforming the data into principal components that capture the major patterns of variation. This method allows us to highlight the most influential n-grams across different countries. Figures 3a and 3b plot the PCA results across countries. In Table II, we list the five influential bigrams and trigrams identified from the PCA of OJA data. These n-grams have the highest absolute loadings on the first two principal components, indicating their significant contribution to the patterns of variation captured by these components.

Countries sharing the same official language are very close in their principal components. Likewise, there is a clear vertical separation between countries with Romance languages Spain, France and Italy and countries with Germanic languages Austria, Germany, UK and US. This clearly points to similar sentence patterns of the related languages.

When adding the Wikilingua data to the plot, the overlap of each language data was not very pronounced, but rather there was a clear separation between OJA and Wikilingua data. Upon investigating the influential bi- and trigrams per component of the PCA with and without the additional Wikilingua data, we found that the former were much more related to function words rather than content words. This indicates that certain structural patterns in OJA or Wikilingua text genres govern language specific differences to some extent. On the other hand, the examination of influential bi- and trigrams in OJA data for the components in Figures 3a and 3b reveals that they almost exclusively contain at least one noun. As we have described above, nouns are very frequent in OJAs based on our unigram analysis. However, the exact patterns in which nouns appear seem to vary between languages (see Table II).

TABLE II Influential bigrams and trigrams for the first two principal components.

Component	Influential N-grams
Component 1	Bigrams: ('ADP', 'NOUN'), ('ADJ', 'NOUN'), ('NOUN', 'ADP'), ('NOUN', 'ADJ'), ('DET', 'NOUN') Trigrams: ('NOUN', 'ADP', 'NOUN'), ('ADP', 'NOUN', 'ADP'), ('DET', 'NOUN', 'ADP'), ('ADP', 'DET', 'NOUN'), ('ADP', 'NOUN', 'ADJ')
Component 2	Bigrams: ('DET', 'NOUN'), ('NOUN', 'NOUN'), ('PROPN', 'PROPN'), ('NOUN', 'SPACE'), ('SPACE', 'NOUN') Trigrams: ('NOUN', 'SPACE', 'NOUN'), ('NOUN', 'NOUN', 'PUNCT'), ('NOUN', 'DET', 'NOUN'), ('PROPN', 'PROPN', 'PROPN'), ('ADJ', 'NOUN', 'SPACE')

Paragraph & List information: Table III provides the mean and median amount of paragraphs and lists per country as well as their length. France has the most paragraphs while Austria and Italy have the fewest. US and UK data has the longest paragraphs, and Germany and Italy have the shortest. Likely, these trends are to some extent related to the total lengths of ads per country. However, while having similarly long ads, Germany has substantially more paragraphs than Austria while consequently the paragraphs in Austrian ads are longer.

With regard to lists the most prominent observation is that in Spain, France and Italy more than half of the ads did not have a single list (median = 0). On the other hand, Austria, Germany, the UK and the US had around 2 lists per document on average. However, the average amount of items per list was rather similar across all countries with around 5 items per list as median and average.

Language Detection Overall, the amount of job ads labeled with a language other than the countries' main language was small (Figure 4). Spain had a substantially larger amount compared to the other countries, albeit still minor with around 2% of the texts being labeled with a language other than Spanish. The foreign language texts detected for Spain were exclusively labeled with English and Portuguese. In the other non English-speaking countries, the majority of the foreign texts were detected as English. In the US, the predominant foreign language was Spanish, and Arabic in the UK. Noteworthily, the language detection also revealed a few job ads only contained a very short text, such as a city name and a postal code. The language detection model did not properly work for such instances, resulting in misclassification of the respective samples. Arguably, these texts cannot actually be counted as real job ads. However, since only a very small number of samples was affected, we decided to not further address this issue.

Salary Information: Figure 5 shows the comparison of salary information obtained from a structured website field vs. from the texts. It is apparent that all countries have salary information in the texts that is not included in the structured information of the website. However, the ratio differs significantly. While Germany displays a discrepancy of almost 40 percentage points, in the UK only about 10% of the ads contain additional salary information in text. Generally, Italy has the lowest share of salary information in job ads, whereas UK, US and Austria have a relatively large proportion across both identification strategies.

VIII. DISCUSSION

In this section, we closely examine the results from Section VII, interpreting the outcomes of each method individually (Section VIII-A). We provide explanations for observed phenomena, sometimes overlapping with Section IX, where we discuss limitations of specific methods. However, Section IX focuses more on macro-level limitations rather than individual methods. In Section VIII-B, we adopt a broader perspective, relating our findings to our research question and reflecting on the usefulness of our experiments in gaining insights.

A. Interpreting the Results

Document length: One important finding of looking at the document length is that indeed OJAs frequently exceed the token limit of 512 tokens typically used by transformer-based language models, although for some countries like Italy and Spain, the job ads are exceptionally short and on average below this limit. The large variations of document lengths observed in OJAs (ads in the US are more than twice as long on average) is much more profound than for the Wikilingua data. This indicates that job ads as a text genre can have culture-specific variations that cannot solely be attributed to characteristics of the language.

Several explanation factors are plausible. For example, some cultures might prefer brevity and directness in communication whereas in others there may be a preference for more detailed and comprehensive communication. Another factor could be different price structures for job ad platforms. We know that some job ad platforms have pricing based on text length, and this can differ dependent on the country where the job is

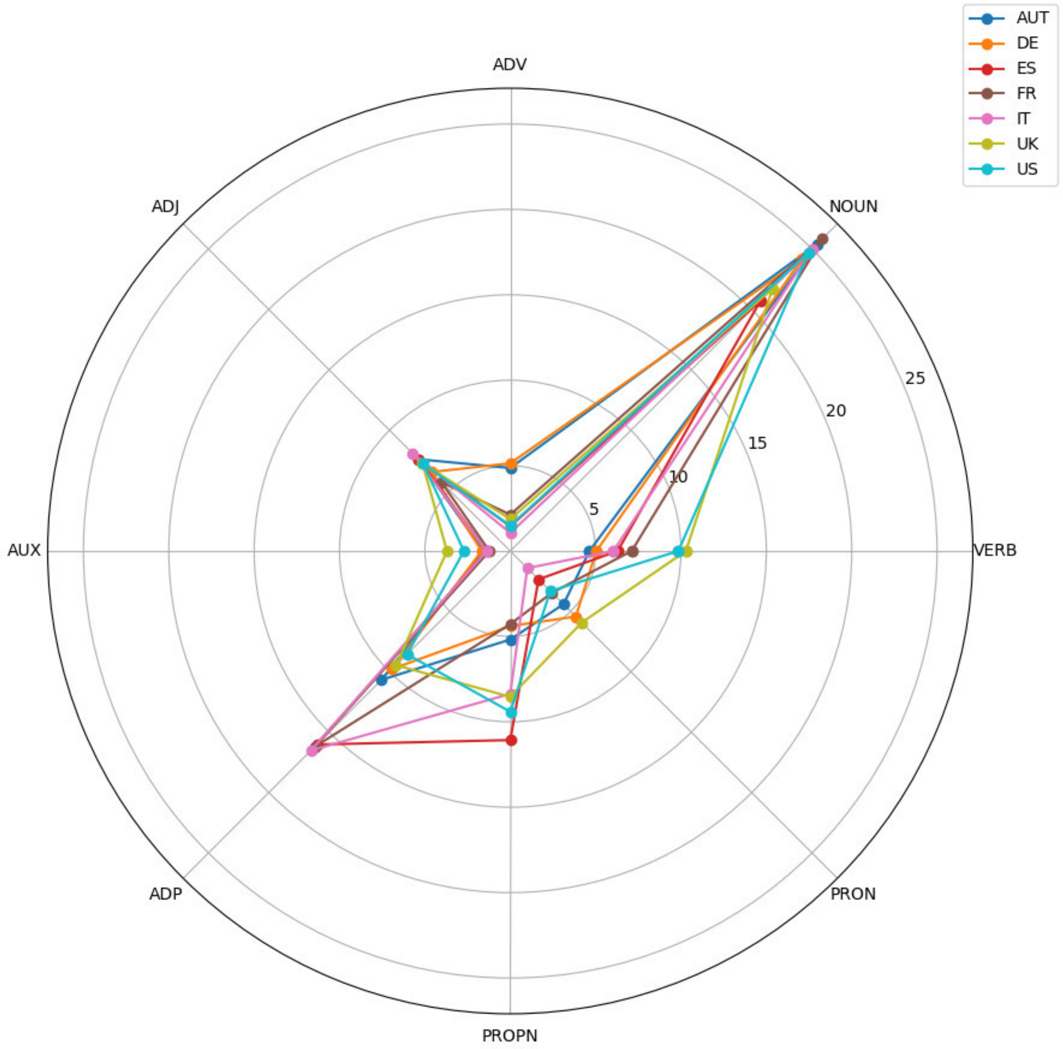
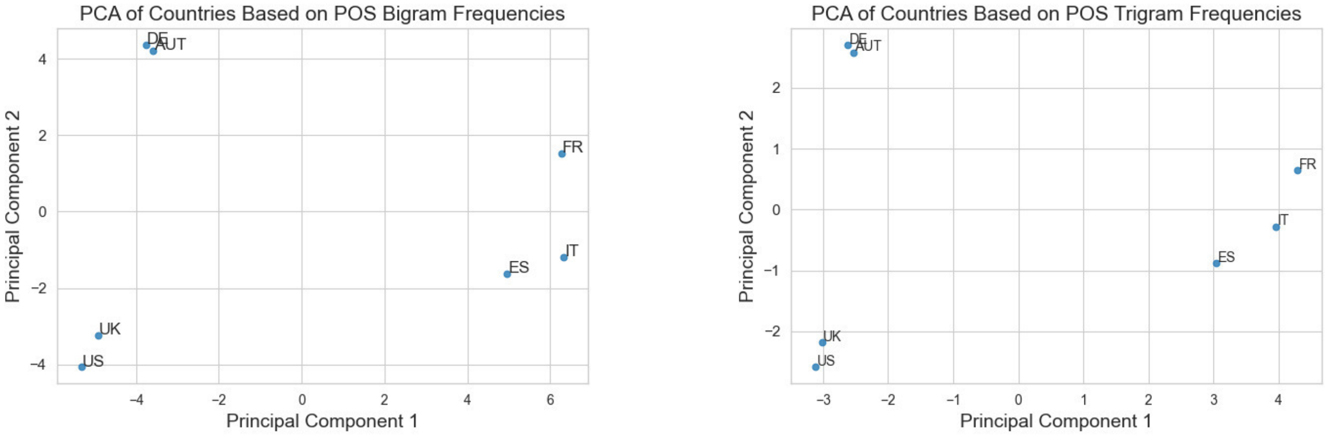


Fig. 2: **POS frequency radar plot.** Radar plot of selected POS types. The plot shows the share of the POS types in the seven OJA country splits.Values in the centre indicate lower frequency of the corresponding POS tag in a country split, values at the edge indicate higher frequency.

TABLE III Comparison of rounded mean and median for the amount of paragraphs per document, length of each paragraph, amount of lists per document, and amount of items per list across countries.

Country	Mean Par./Doc.	Median Par./Doc.	Mean Tokens/Par.	Median Tokens/Par.	Mean Lst./Doc.	Median Lst./Doc.	Mean Items/Lst.	Median Items/Lst.
AUT	13.6	12	23.5	7	2.3	3	5.3	5
DE	20.1	17	16	6	1.8	2	5.2	5
ES	16	13	19	12	0.9	0	5.1	5
FR	21.2	20	18.1	12	0.7	0	5.5	5
IT	14.3	13	17.2	9	0.7	0	4.2	4
UK	18.8	17	27.2	14	2.0	2	6	5
US	20.7	17	28.6	12	2.4	2	6.2	5



(a) Principal Component Analysis of Bigrams (OJA data).

(b) Principal Component Analysis of Trigrams (OJA data).

Fig. 3: **Principal Component Analysis (PCA).** Results for bigrams and trigrams in OJA data. The first two components derived from bigram and trigram frequencies across countries are shown.

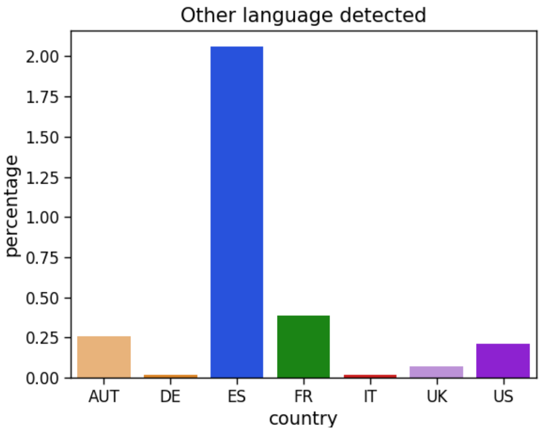


Fig. 4: **Other language detected.** Comparison of the relative share of ads where a language that is not the countries main language was detected.

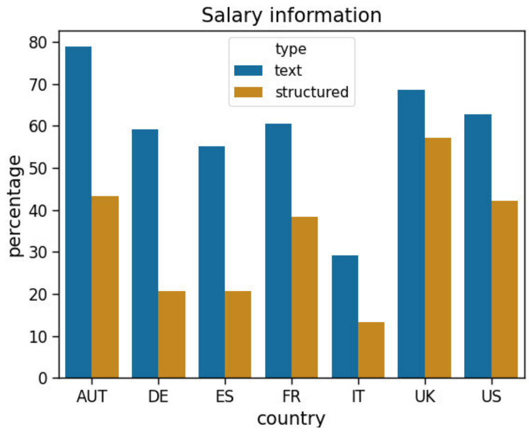


Fig. 5: **Salary information.** Comparison of the relative share of ads with salary information across countries and the two methods of detecting salary.

offered. We do not know as to whether this is the case for our CareerJet data. Even if not, it still might have indirect impact, given that employers may write their ads for several platforms. Another factor could be how formalized qualifications are in a country. For example, Germany has a very formalized vocational education system in place. Therefore, if a company mentions that a potential employee is expected to bring a finished apprenticeship training as, for example, an electrician with a certain specialization, a lot of their expected skills are presumed. In a country with a less formalized vocational education system, the company might feel the need to elaborate more on what exactly they are looking for. Other factors might include legal aspects, labor market dynamics (e.g. what type of occupations are in demand?) or awareness for SEO-optimization.

Diversity Metrics: Our results in Table I show that text diversity differs substantially across countries, but job ads are generally more diverse on a purely lexical level than on other (semantic, syntactic) levels. This indicates that job ads as a text genre have specific linguistic properties. At the same time, the differences between countries, for example between the n-gram VS of France and Germany are quite strong. Also, no country is consistently highest or lowest in all diversity rankings, showing that OJAs from different countries have different properties with regards to various linguistic levels. This leads us to conclude that researchers should carefully reflect the lexical, semantic and syntactical structures of their OJA data when designing research projects. More diverse data might require additional evaluation to ensure robust conclusions.

Syntactic analysis: The analysis of selected POS as well as the PCA analysis of bi- and trigrams show that OJAs can be seen as a text type with unique syntactic characteristics across countries, such as the increased amount of nouns we found in most countries compared to the Wikilingua data. However, our analysis also reveal that these characteristics do not hold for all countries, e.g. US job ads did not have substantially more nouns than English Wikilingua data. This points to the need of reflecting NLP methods when trying to transfer methods or models from one language to another. Certain patterns, that are typical for entities in one country, might be expressed differently in another country. At the same time, researchers cannot rely solely on knowledge they have about the characteristics of the different languages used in these countries, because differences may be specific to the text genre OJA.

Paragraph & List information: The variability in job ad structure indicated by paragraph amount and length impacts NLP tasks such as text zoning for information extraction. Training NLP models predominantly on data from one country may reduce their effectiveness when applied to structurally different ads from other countries. Moreover, the notable variation in list usage across countries like Spain, France, and Italy, where lists often are absent, complicates the application of list-based analytical methods developed in countries like Germany or the UK. The high similarity in the amount of list items indicates that the amount of skills, tasks or benefits listed in OJAs does not differ substantially across countries.

However, our detection of lists and paragraphs is exclusively based on HTML-analysis. It is possible that a text may simply contain line-breaks combined with list-indicating symbols (like hyphen) as a list. We do, however, know that the website we scraped the data from offers a bullet-point-button in the field where employers put the main body of the job ad.

Language Detection: Our results indicate that ads in other languages than the countries' official language do not occur equally frequent across languages. Based on our results, in Spain this would be an influential factor in the OJA pipeline, possibly requiring an additional data cleaning step or a multilingual approach. At the same time, when having a much larger dataset, this would also provide additional research opportunities, because the occurrence of ads in different languages could be related to other factors. For example, OJAs with different languages might differ regionally or with regard to job requirements.

However, we found several uncertainties with our method. First, the model was limited to the languages it knows. We found that some texts from the Spanish dataset labeled with Spanish or Portuguese are actually written in Catalan, which the model was not trained on. Also, some ads appear to contain very short and noisy texts. If a data point consists of mostly noise, such as symbols or contact addresses, etc., the model tended to predict a completely unrelated language. These cases could easily be identified as false classifications in a visual inspection. However, these instances are misleading in quantitative analysis. Finally, ads might include two (or more) languages [22] to cater to a local as well as international audience. We found that sometimes the ad was repeated in another language. A more refined approach could detect such instances.

Salary Information: With regard to salary information there were major differences between the countries for both structured and text-based information. This shows that transferring research on the basis of correlating salary information to other characteristics from one country to another is not straightforward and requires careful consideration. Also, our method only gives sparse information about the type of salary information detected in texts. Firstly, we do not know if the structured information (e.g. hourly wage) is always repeated in the free text. It might be plausible for the employer to omit repeating it, because they know that employees get that information trough the text field and the visual aid of the websites' structure. We do, however, know that information is repeated at least in some cases, because for some countries the sum of text and structured information is above 100%.

Furthermore, we suspect that there are two major types of information that our NLI model detects except from precise salaries based on exploratory qualitative analysis. The first group is about collective wage information. Our expertise is mostly in the German labor market and here our intuition is that employers rarely mention concrete numbers like an hourly wage, but will mention, if the position is paid based on a collective wage. Especially jobs in the large public sector always include this type of information. The second group is employers advertising their payment using phrases like "attractive" or "above-average" salary. Differentiating between these (and possible more) groups of payment information might give further insight into how the differences between countries can be explained. For example, we suspect the large increase of textual compared to structured salary information in German job ads to be mostly caused by the great amount of collective wage mentions in Germany. Knowing what type of salary information is present in a country and to what extent might further help researchers develop or discard research ideas using salary information obtained from job ads.

B. *Reflection on Experiments*

Our central research question was whether or not we could find differences in form or content in OJAs that could hinder the transfer of NLP methods or research scope in a major way. Based on our exploratory analysis, we can confirm that there are substantial differences between OJAs from different countries. This applies to both linguistic and content features. Also, the linguistic differences we detected did only partially correspond to the respective cross-lingual differences observed in the Wikilingua reference data. This indicates that there are differences between countries that are specific to the text genre OJA. Interestingly, OJA data from countries that share the same official language had rather similar linguistic features. For example, English texts are the shortest in Wikilingua. However, US OJA texts are longest of all countries, indicating that long texts are not a property to English, but only to US OJAs, indicating cultural factors. Yet, we observe the same (although not quite as extreme) for data from the UK. This raises the question whether this behavior of the language pairs Germany/Austria and UK/US can be explained by cultural similarities of these countries that share a language or other factors we did not uncover.

With regard to the different methods used, we find that experiments motivated by specific questions, for example text length ("Can we use a 512 token truncation?") or salary information ("Can we correlate salary information to other properties of job ads?") made it easier to draw more concrete conclusions and derive actionable recommendations. For less concretely motivated methods such as text diversity, where we argued that it might be an indicator for how complex various NLP tasks might be (for example, "How many examples do we need to manually evaluate in Zero Shot scenario?"), we cannot simply derive answers based on our results. At the same time, the differences for some of these metrics proved to be quite large. This has implications on two major levels.

Firstly, NLP practitioners developing OJA analysis pipelines should be aware that OJAs can have substantial differences that may go beyond the linguistic differences expected for different languages. If applicable, our results can be used directly to draw conclusions on methods to be applied. Otherwise, our analysis pipeline can be used to analyze other data and then relate to our findings to check whether there may be problems in applying existing NLP OJA methods. Ideally, researchers should add further experiments for the concrete problem they are facing. This prerequisites that they find ways to quantify text properties in a manner that is useful to their problem, which may be challenging.

Secondly, with the rise of NLP methods in various contexts like legal texts, medical texts, social media, literature and so on, we will find an increasing amount of text genres that NLP methods are being developed for. Consequently, NLP practitioners from text genres will look into building upon published work from other languages or countries. As we have shown, some pitfalls exist in this process. Therefore, we advocate researchers working with other text genres to perform analyses similar to ours, using easily accessible tools to gain quantitative insight into their data and how it behaves compared to texts from the same genre in other languages or countries. Ideally, future work would reflect further upon the specific experimental methods and refine a more standardized, yet flexible pipeline that researchers can revert to when they intend to perform an analysis like ours. In this sense, our paper provides a starting point for researchers looking to develop a similar pipeline.

IX. LIMITATIONS

Despite our promising results, there are limitations to our study, which we want to discuss in this section. As explained in Section V we chose to scrape data from one website that operates in multiple countries to minimize biases introduced by confounding variables based on target audiences from different job portals. At the same time this choice also means that we equated the properties of the text genre OJA from a given country with OJA data from only a single source. So, while our approach helps in comparability between countries, the countries themselves are not thoroughly represented. Ideally, future work reproduces our experiments with a dataset from mixed source websites that at the same time ensures comparability by choosing similar distributions of relevant variables like occupation or industry sector. Perhaps, this would require an even larger dataset, which, however, makes the experiments more computationally expensive.

Furthermore, there are two factors that are partly of ethical nature. In our study we mentioned that we chose two country pairs that share the same official language in order to better differentiate whether results were based simply on language factors or on cultural factors. That way, we equated a different country to a different culture. This is clearly a very simplified view of the intricacies of cultures and states. We are aware, for example, that within one country, major cultural differences may exist. Also, linguistic discrepancies between a country's regions can be strong, one example being the Catalan ads we found in our Spanish dataset. This may also lead to limited replicability of our experiments for countries where multilingualism is even wider spread. Somewhat related to this issue is the choice of countries, which was mostly based on practicability. We are aware of issues in the NLP community with regards to underrepresented languages [50]. The languages we investigated here all belong to the better represented ones. This is particular important due to the fact that especially underrepresented languages logically have a greater need to adopt methods developed originally for other languages and would therefore profit most from our research. However, due to the reasons explained in Section V including these languages was beyond the scope of this paper. Future research should therefore focus on including underrepresented languages to our analysis.

X. CONCLUSION

Our analysis of OJAs across various countries and languages reveals substantial linguistic and content-related differences, emphasizing the complexity of transferring NLP methods and research findings across different contexts. The variations in document length, diversity metrics, syntactic structures, and salary information highlight the need for tailored approaches in NLP research. While our study offers valuable insights, it also points to the necessity of further research, particularly involving underrepresented languages and larger, more representative datasets. By acknowledging these differences and adapting NLP methods accordingly, researchers can improve the accuracy and relevance of their analyses in the context of global labor markets. Our findings serve as a foundation for developing standardized yet flexible pipelines for analyzing text genres across different languages and cultures.

REFERENCES

[1] A. Lima, B. Bakhshi *et al.*, "Classifying occupations using web-based job advertisements: an application to stem and creative occupations," *Economic Statistics Centre of Excellence Discussion Paper*, vol. 8, 2018.

[2] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, and A. Seveso, "Neo: A system for identifying new emerging occupation from job ads," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 16 035–16 037.

[3] E. Senger, M. Zhang, R. van der Goot, and B. Plank, "Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings," 2024.

[4] M. Buchmann, H. Buchs, F. Busch, S. Clematide, A.-S. Gnehm, and J. Müller, "Swiss job market monitor: A rich source of demand-side micro data of the labour market," *European Sociological Review*, vol. 38, no. 6, pp. 1001–1014, 2022.

[5] E. Atalay, P. Phongthiengtham, S. Sotelo, and D. Tannenbaum, "The evolution of work in the united states," *American Economic Journal: Applied Economics*, vol. 12, no. 2, pp. 1–34, 2020.

[6] E. Atalay, S. Sotelo, and D. Tannenbaum, "The geography of job tasks," *Journal of Labor Economics*, 2023.

[7] P. Kuhn and K. Shen, "Gender discrimination in job ads: Evidence from china," *The Quarterly Journal of Economics*, vol. 128, no. 1, pp. 287–336, 2013.

[8] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.

[9] S. Chaturvedi, K. Mahajan, and Z. Siddique, "Words matter: Gender, jobs and applicant behavior," *Jobs and Applicant Behavior (February 18, 2024)*, 2024.

[10] M. Ganesan, S. P. Antony, and E. P. George, "Dimensions of job advertisement as signals for achieving job seeker's application intention," *Journal of Management Development*, vol. 37, no. 5, pp. 425–438, 2018.

[11] B. Bullinger, "Companies on the runway: Fashion companies' multimodal presentation of their organizational identity in job advertisements," in *Multimodality, meaning, and institutions*. Emerald Publishing Limited, 2017, vol. 54, pp. 145–177.

[12] J. Binnewitt and T. Schnepf, "Join us to turn the wor (l) d greener!—investigating online apprenticeship advertisements' reference to environmental sustainability," *Zum Konzept der Nachhaltigkeit in Arbeit, Beruf und Bildung—Stand in Forschung und Praxis*, 2022.

[13] Cedefop, *Online job vacancies and skills analysis – A Cedefop pan-European approach*. Publications Office, 2019.

[14] P. Descy, V. Kvetan, A. Wirthmann, and F. Reis, "Towards a shared infrastructure for online job advertisement data," *Statistical Journal of the IAOS*, vol. 35, no. 4, pp. 669–675, 2019.

[15] A. T. S. Calazans, R. A. Paldes, E. T. S. Masson, I. S. Brito, K. F. Rezende, E. Braosi, and N. Pereira, "Software requirements analyst profile: A descriptive study of brazil and mexico," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE, 2017, pp. 204–212.

[16] L. Bowker, "What does it take to work in the translation profession in canada in the 21st century? exploring a database of job advertisements," *Meta*, vol. 49, no. 4, pp. 960–972, 2004.

[17] D. Rear, "Converging work skills? job advertisements and generic skills in japanese and anglo–saxon contexts," *Asian Business & Management*, vol. 12, pp. 173–196, 2013.

[18] C.-H. Chung and L.-J. Chen, "Text mining for human resources competencies: Taiwan example," *European Journal of Training and Development*, vol. 45, no. 6/7, pp. 588–602, 2021.

[19] M. A. Kennan, F. Cole, P. Willard, C. Wilson, and L. Marion, "Changing workplace demands: What job ads tell us," in *Aslib Proceedings*, vol. 58, no. 3. Emerald Group Publishing Limited, 2006, pp. 179–196.

[20] A. Trosborg, "Text typology: Register, genre and text type," *Benjamins Translation Library*, vol. 26, pp. 3–24, 1997.

[21] C. L. Engstrom, J. T. Petre, and E. A. Petre, "Rhetorical analysis of fast-growth businesses' job advertisements: Implications for job search," *Business and professional communication quarterly*, vol. 80, no. 3, pp. 336–364, 2017.

[22] F. van Meurs, B. Planken, H. Korzilius, and M. Gerritsen, "Reasons for using english or the local language in the genre of job advertisements: Insights from interviews with dutch job ad designers," *IEEE Transactions on Professional Communication*, vol. 58, no. 1, pp. 86–105, 2015.

[23] C. Berkenkotter and T. N. Huckin, *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge, 2016.

[24] Y. Sun, "Genre formation in contexts: a cross-lingual comparison of english ma thesis introductions." *Linguistics & the Human Sciences*, vol. 10, no. 3, 2014.

[25] B. Melander, "Culture or genre? issues in the interpretation of cross-cultural differences in scientific papers," *Genre studies in English for academic purposes*, vol. 9, pp. 211–226, 1998.

[26] Y. Zhu, "A situated genre approach for business communication education in cross-cultural contexts," in *The Routledge handbook of language and professional communication*. Routledge, 2014, pp. 26–39.

[27] C. C. Nickerson, "The usefulness of genre theory in the investigation of organizational communication across cultures," *Document Design*, vol. 1, no. 3, pp. 203–215, 1999.

[28] D. Kuhi and M. Mojood, "Metadiscourse in newspaper genre: A cross-linguistic study of english and persian editorials," *Procedia-Social and Behavioral Sciences*, vol. 98, pp. 1046–1055, 2014.

[29] H. Marefat and S. Mohammadzadeh, "Genre analysis of literature research article abstracts: A cross-linguistic, cross-cultural study," *Applied research on English language*, vol. 2, no. 2, pp. 37–50, 2013.

[30] L. Filipović, "The role of language in legal contexts: A forensic cross-linguistic viewpoint," *Law and Language: Current Legal Issues*, vol. 15, no. 19, pp. 328–343, 2013.

[31] H. Ansary and E. Babaii, "A cross-cultural analysis of english newspaper editorials: A systemic-functional view of text for contrastive rhetoric research," *RELC Journal*, vol. 40, no. 2, pp. 211–249, 2009.

[32] I. G. Izquierdo and V. M. i Resurrecció, "Translating into textual genres," *Linguistica Antverpiensia, new series–themes in translation studies*, vol. 1, 2002.

[33] V. Montalt, P. Ezpeleta-Piorno, and I. García-Izquierdo, "The acquisition of translation competence through textual genre," 2008.

[34] Ł. Biel, "Genre analysis and translation," in *The Routledge handbook of translation studies and linguistics*. Routledge, 2017, pp. 151–164.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[37] F. Ladhak, E. Durmus, C. Cardie, and K. McKeown, "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4034–4048.

[38] SpaCy, https://spacy.io/models, Accessed April 2024.

[39] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *Transactions on Machine Learning Research*, 2023.

[40] 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.

[41] I. Beltagy, A. Cohan, R. Logan IV, S. Min, and S. Singh, "Zero- and few-shot NLP with pretrained language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, L. Benotti, N. Okazaki, Y. Scherrer, and M. Zampieri, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 32–37.

[42] A. Magron, A. Dai, M. Zhang, S. Montariol, and A. Bosselut, "Jobskape: A framework for generating synthetic job postings to enhance skill matching," *arXiv preprint arXiv:2402.03242*, 2024.

[43] A.-S. Gnehm, "Text zoning for job advertisements with bidirectional lstms," 2018.

[44] A.-S. Gnehm and S. Clematide, "Text zoning and classification for job advertisements in german, french and english," in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 2020, pp. 83–93.

[45] J. Hermes and M. Schandock, "Stellenanzeigenanalyse in der qualifikationsentwicklungsforschung," *Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten. Bundesinstitut für Berufsbildung, Bonn*, 2016.

[46] L. Papariello, "xlm-roberta-base-language-detection (revision 9865598)," 2024.

[47] M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised

machine learning with deep transfer learning and bert-nli," *Political Analysis*, vol. 32, no. 1, pp. 84–100, 2024.

[48] E. Andrieu and M. Kuczera, "Minimum wage and skills: Evidence from job vacancy data," The Productivity Institute, Tech. Rep., 2023.

[49] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002.

[50] J. Nee, G. M. Smith, A. Sheares, and I. Rustagi, "Linguistic justice as a framework for designing, developing, and managing natural language processing tools," *Big Data & Society*, vol. 9, no. 1, p. 20539517221090930, 2022.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 95–104 DOI: 10.15439/2024F1140 ISSN 2300-5963 ACSIS, Vol. 41

Predictive System of the Water Deficit Analysis for the Black Sea Lowland (an example of the Kherson Region)

Lyudmyla Kuzmych
0000-0003-0727-0508
[a] Kherson State Agfrarian and Economic University,
Stritenska Str., 23,
73006 Kherson, Ukraine;
[b] Institute of Water Problems and Land Reclamation,
Vasylkivska Str. 37,
03022 Kyiv, Ukraine
Email:
kuzmychlyudmyla@gmail.com

Mykola Voloshyn
0000-0003-0467-1963
Kherson State Agfrarian and Economic University,
Stritenska Str., 23,
73006 Kherson, Ukraine;
Email:
voloshin_nik_1977@ukr.net

Yuriy Kyrylov
0000-0001-5061-2644
Kherson State Agfrarian and Economic University,
Stritenska Str., 23,
73006 Kherson, Ukraine;
Email:
kirilov_ye@ukr.net

***Abstract*— Developing a predictive system for water deficit analysis in the Black Sea Lowland, especially in climate change, involves integrating various data sources, modeling techniques, and technological tools to forecast water availability and demand.**

The analysis of the change in moisture deficit in the Kherson region is provided for the period from 1955 to 2022. The description of with temperature gradients across the Kherson region is provided. The distribution of precipitation throughout the years in terms of quantity and intensity is provided.

As part of the Black Sea Lowland, the Kherson region is critically important for southern Ukraine's agriculture and water security. Given the region's reliance on irrigation and the challenges posed by climate change, developing a predictive system for water deficit analysis is essential. Such a system can help stakeholders make informed decisions to ensure sustainable water management and mitigate the adverse effects of water scarcity.

***Index Terms*— predictive system, water deficit, climate change, the Kherson region, lack of moisture, weather stations, precipitation**

I. INTRODUCTION

AS PART of the Black Sea Lowland, the Kherson region is critically important for southern Ukraine's agriculture and water security. Given the region's reliance on irrigation and the challenges posed by climate change, developing a predictive system for water deficit analysis is essential [1-5].

The Kherson region is a major agricultural hub, producing significant quantities of grains, vegetables, and fruits, which rely heavily on irrigation. Agriculture is a key economic driver, contributing significantly to the local and national economies. Ensuring water availability for agriculture is crucial for food security and economic stability [6].

The Dnieper River and associated canals (e.g., the North Crimean Canal) are primary water sources for irrigation and municipal use. Groundwater resources are also vital, though their overuse can lead to depletion and degradation, impacting long-term water availability [7-11].

Rising temperatures lead to higher evaporation rates, reducing water availability. This can exacerbate water shortages, especially during peak agricultural periods [12–17].

Changes in precipitation patterns, with more frequent droughts and irregular rainfall, affect water supply consistency. Unpredictable weather patterns make it difficult to plan for water usage effectively [17-19].

Due to increasing water demand for irrigation due to intensive farming practices efficient water management becomes crucial to sustain agricultural productivity [3]..

Growing population and industrial activities further strain water resources. Ensuring an adequate supply of water for all sectors is essential to support regional development [19].

Therefore a need for a predictive system of the water deficit analysis, which can provide early warnings of potential water shortages, allowing for proactive management and planning. This can help in preparing for droughts and other extreme events. Also, it helps in devising strategies to mitigate risks associated with water scarcity, such as crop failure and economic losses.

II. METHODS AND TECHNIQUES

It is proposed the structured approach to designing the predictive system for water deficit analysis consists of the following steps :

1. Data Collection and Integration:
    1.1. Climate data - collect historical data on temperature, precipitation, humidity, and wind patterns.
    1.2. Hydrological data:
        - River flows: data on river discharge rates, particularly for major rivers like the Dnipro River;
        - Reservoir levels: historical and current water levels in reservoirs and major water bodies;
        - Groundwater levels: data from groundwater monitoring wells.
    1.3. Water Usage Data:
        - Agricultural Demand: Information on crop types, irrigation practices, and water usage patterns;
        - Domestic and Industrial Use: Data on municipal and industrial water consumption.
2. Modeling Framework:
    2.1. Climate models - apply regional climate models to downscale global climate projections to the Black Sea Lowland region for more accurate local predictions.
    2.2. Hydrological models:

- Rainfall-runoff models - use models like SWAT (Soil and Water Assessment Tool) to simulate the conversion of rainfall to runoff and river flow;
- Water balance models - implement models that account for inputs (precipitation, river inflows) and outputs (evaporation, water extraction) to estimate water availability.

2.3. Demand models:

- Agricultural demand - develop models to estimate future irrigation needs based on crop types, planting schedules, and climate conditions;
- Municipal and Industrial demand - use demographic and economic projections to forecast changes in water demand for domestic and industrial uses.

3. System Design:

3.1. Data integration platform:

- GIS integration - use a Geographic Information System (GIS) to integrate spatial data on climate, hydrology, and land use;
- Database Management - implement robust databases to store and manage large volumes of data from various sources.

3.2. Predictive Analytics / Scenario Analysis - develop tools to simulate different scenarios (e.g., varying levels of water usage, different climate change scenarios) and assess their impact on water deficit.

4. Application and Use Cases:

4.1. Agricultural planning:

- Crop planning - assist farmers in planning crop types and irrigation schedules based on predicted water availability;
- Water allocation - help water managers allocate water resources efficiently during dry periods.

4.2. Urban and industrial water management:

- Demand management - support municipalities and industries in managing water demand and implementing conservation measures;
- Infrastructure planning - guide investment in water infrastructure, such as reservoirs and pipelines, to mitigate future water deficits.

To characterize the climatic features of the object, data from the weather stations Askania-Nova, Velyka Oleksandrivka, Kherson, Nova-Kakhovka, Nizhny Sirogozy, Khorly (Fig. 1.) and the relevant sources were used [20-25] .

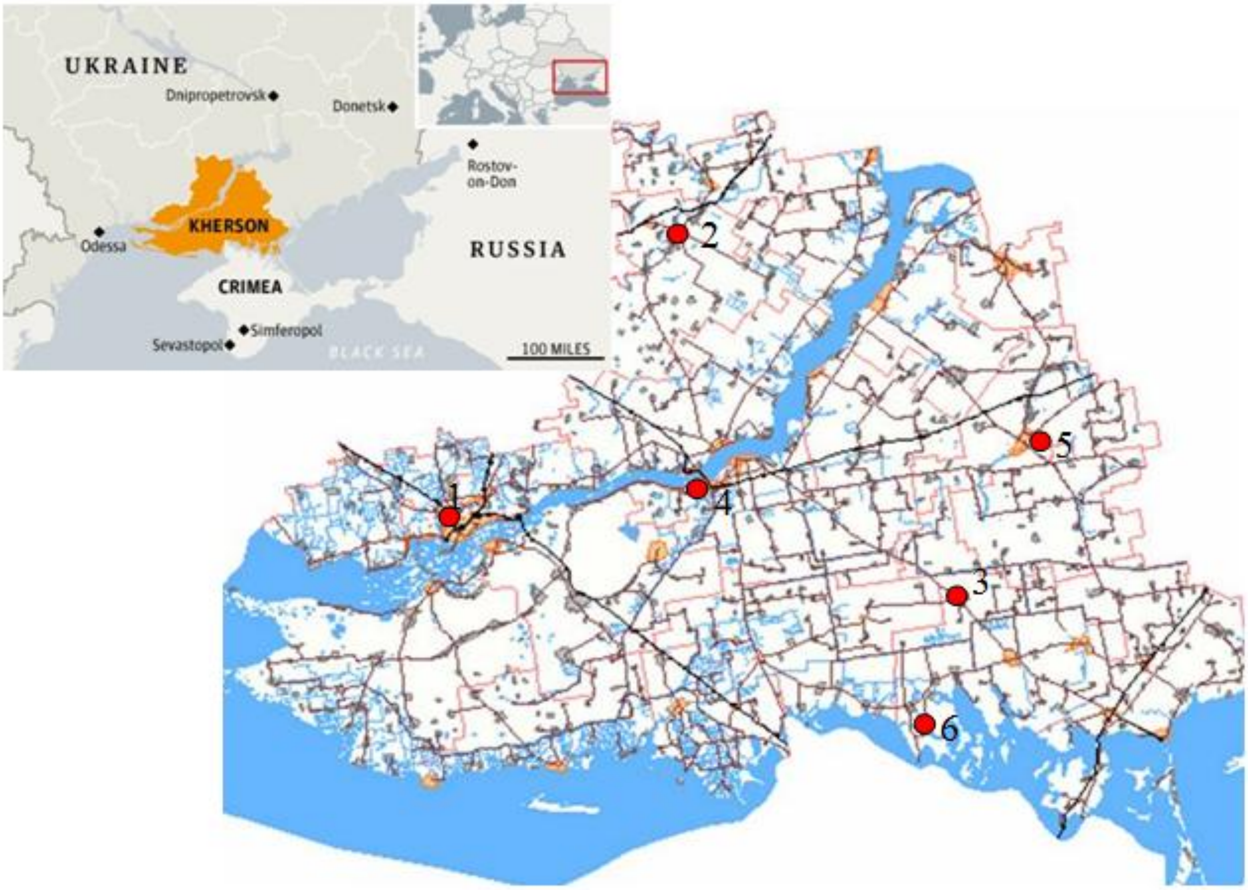


Fig. 1. Scheme of the map of the Kherson region and the location of weather stations: 1 – Kherson; 2 – Velyka Oleksandrivka; 3 – Askania-Nova; 4 – Nova Kakhovka; 5 – Nizhny Sirogozy; 6 - Khorly.

III. RESULTS AND DISCUSSION

According to the Köppen climate classification [20-22], the classification of climates Alisova et al. [23-25] this territory belongs to the Black Sea sub-region of the Atlantico-continental steppe region. The climate is typically continental with a high heat resource and insufficient humidity.

The change of seasons occurs gradually without sharp temperature fluctuations, the average annual air temperature ranged from 8.1°C (1976) to 11.4°C (1966), with an average annual value of 9.7°C. The coldest month is January, with an average monthly temperature ranging from -11.2°C (1972) to 1.9°C (1966) and an average annual value of -3.5°C. The winter period does not exceed 100 days. Winter is short, moderately cold, mild, with frequent thaws. The snow cover usually appears in November-December, characterized by instability, and melts in February-March.

The snow depth does not exceed 5-10 cm. Snow is not the main source of spring moisture accumulation in the soil in this area. The instability of the temperature regime is due to frequent soil thawing in winter, affecting soil moisture in the aeration zone during the winter period.

The spring increase in average daily temperature in March leads to the complete thawing of the frozen layer. The increase in evaporation of moisture in spring, along with rising air temperatures, causes a sharp increase in moisture deficit. The warmest month is July, with an average monthly air temperature ranging from 20.5°C (1969) to 24.3°C (1972) and an average annual value of 23.2°C.

By the degree of humidity, the northern and central parts of the territory belong to the zone of insufficient humidity with a humidity coefficient greater than 0.5 (for Askania-Nova - 0.68), while the southern part (Prisivashshya) belongs to the coastal arid zone with a humidity coefficient of approximately 0.4. The annual precipitation varied from 238.5 mm (1984) to 640.8 mm (1966). According to the book "Climate of Ukraine," on average in the northern hemisphere, the surface air temperature increased by only 0.5°C from 1961-1990, and globally by 0.4°C. The change in annual temperature over a 100-year period in the Steppe region is 0.2-0.3°C towards warming. Winter warming is 1.2°C, in spring - 0.8°C, with minor changes in summer and autumn.

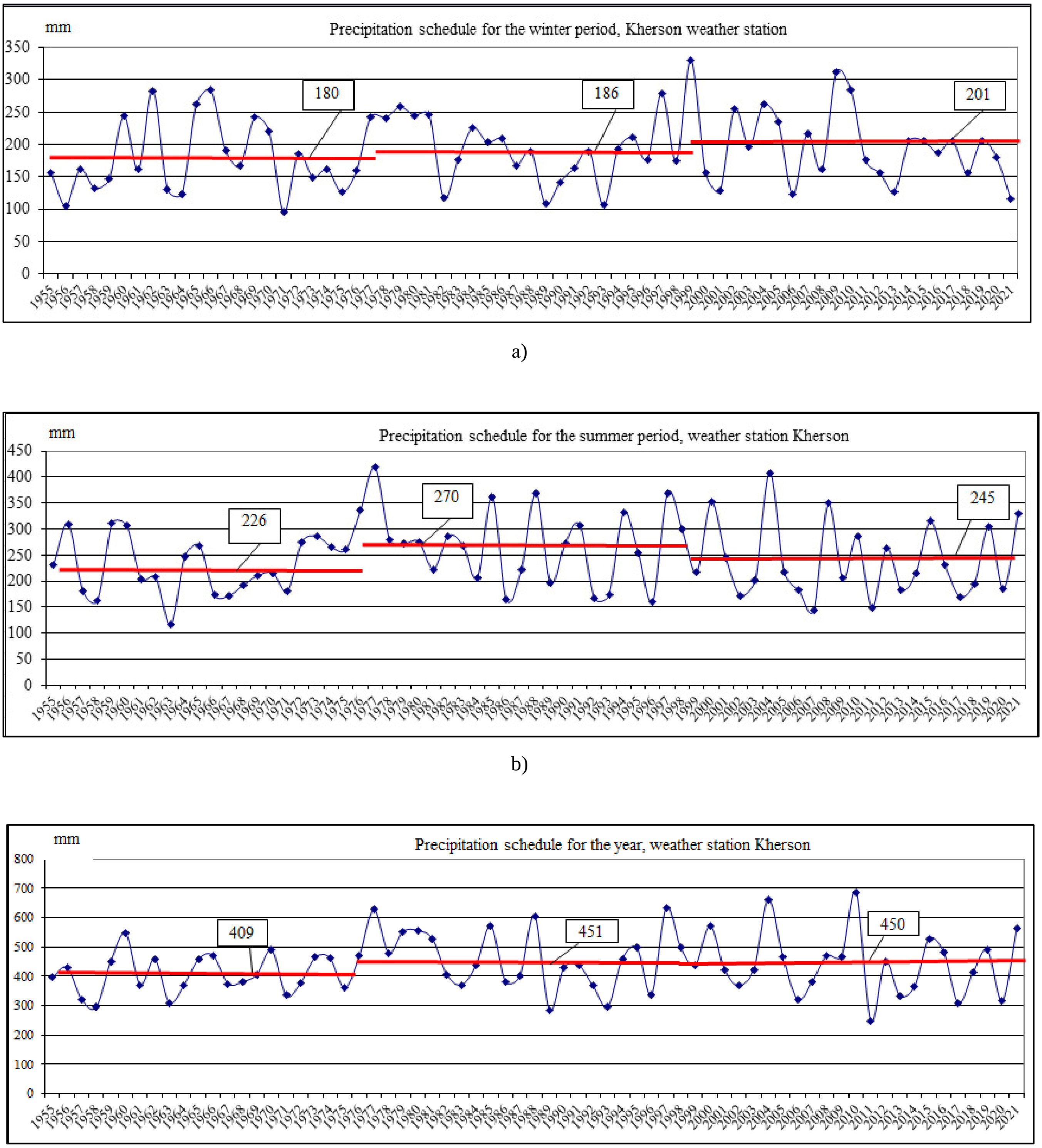
From 1900 to 2020, the annual amount of precipitation in Ukraine varied unevenly. In some regions, there was an increase in precipitation by 7-10% (over 40 mm) from the climatological norm, while in the rest of the territory, it remained within the norm. During the period of maximum global warming, starting from 1975, a decrease in the amplitude of precipitation fluctuations from year to year was observed almost throughout Ukraine. This means that the moisture regime stabilized and is within the climatological norm. It is known that seasonal unevenness in precipitation, an increase in average annual precipitation in recent decades, and the amplitude of precipitation in certain years are natural factors contributing to flooding.

The distribution of precipitation throughout the year is uneven both in quantity and intensity. The highest amount of precipitation falls during dry months when evaporation is high. Summer rainfall (35-40% of the annual total) occurs in the form of short downpours and heavy rains, often accompanied by thunderstorms. The main spatial distribution pattern of precipitation in Ukraine, determined by general circulation factors, is their decrease from the northwest to the southeast.

In the Kherson region, the average annual precipitation decreases from 450 to 300 mm and less from northwest to southeast, reaching 230 mm on the coast of the seas. The distribution of precipitation throughout the year is uneven both in quantity and intensity. The highest amount of precipitation falls during dry months when evaporation is high. Summer rainfall (35-40% of the annual total) occurs in the form of short downpours and heavy rains, often accompanied by thunderstorms. The most significant increase in precipitation is observed in the observation zone of the Kherson and Velyka Oleksandrivka weather stations, with slightly less growth in the observation zone of Nova Kakhovka and Nyzhni Sirohozy weather stations, and very slight increase in the observation zone of Askania-Nova and Khorly weather stations.

An analysis of the long-term dynamics of precipitation at the Kherson and Velyka Oleksandrivka weather stations was conducted for the periods 1955-1975, 1976-1995, and 1996-2022. Comparison of the long-term seasonal unevenness of precipitation at the Kherson weather station indicates a gradual increase in the average precipitation over twenty years, especially in the last forty-five years (41 mm), indicating a gradual increase in overall natural loading. In the winter period (most critical for replenishing groundwater with atmospheric precipitation), the overall increase in average precipitation was 21 mm, while in the summer period, precipitation increased by 19 mm (Fig. 2).

.



a)

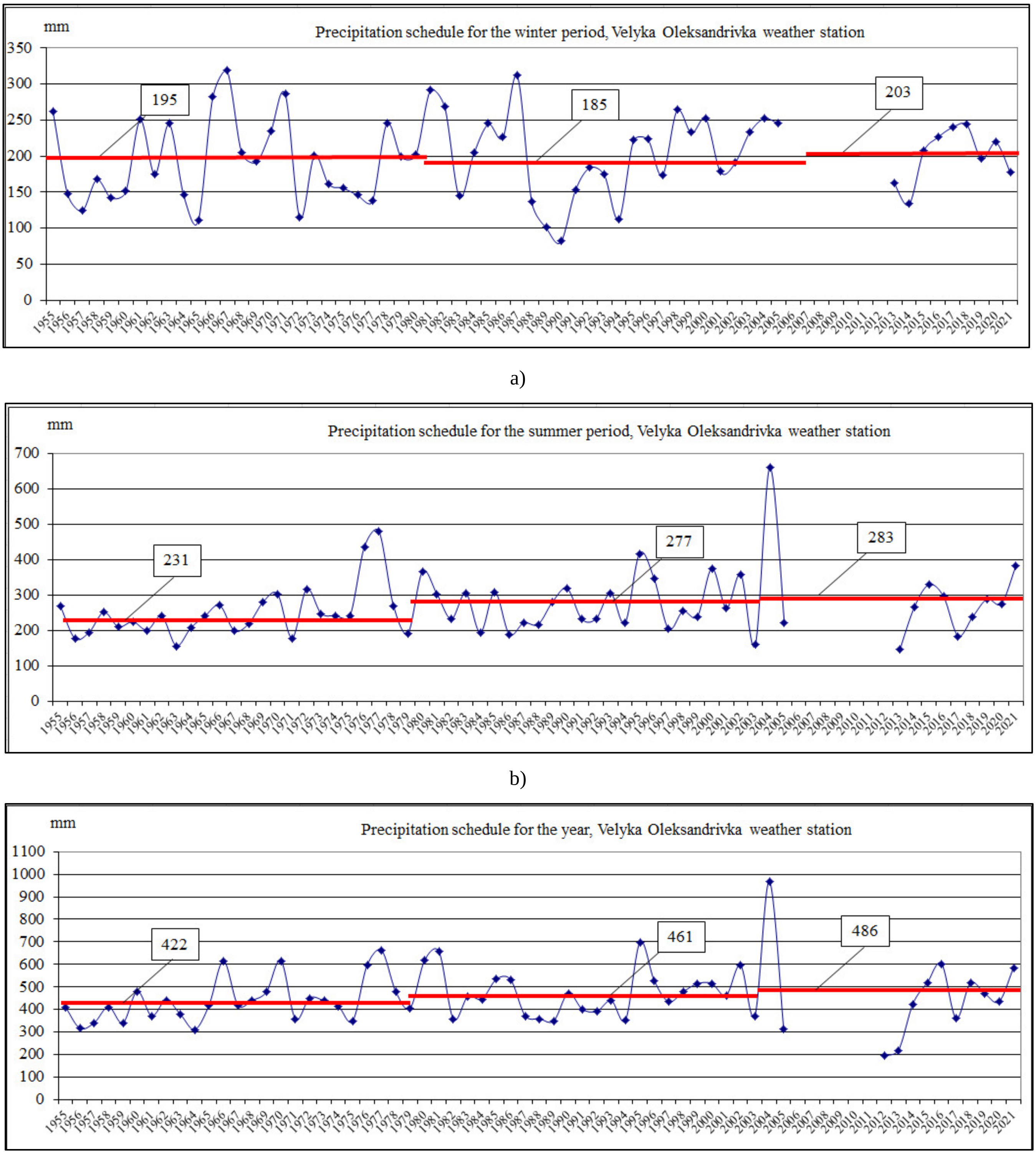
b)

c)

Fig. 2. Precipitation schedule according to Kherson weather station data:
a) winter period; b) summer period; c) for a year.

A comparison of the long-term seasonal unevenness of precipitation at the Velyka Oleksandrivka weather station was carried out in a similar way (Fig. 3). The analysis shows a gradual increase in the average rainfall over sixty-five years (44 mm), which indicates a significant increase in the total natural load. Moreover, in the winter period (the most threatening, in terms of replenishment of groundwater by atmospheric precipitation), the total increase in average precipitation was 8 mm, in the summer period precipitation increased by 48 mm.

.



a)

b)

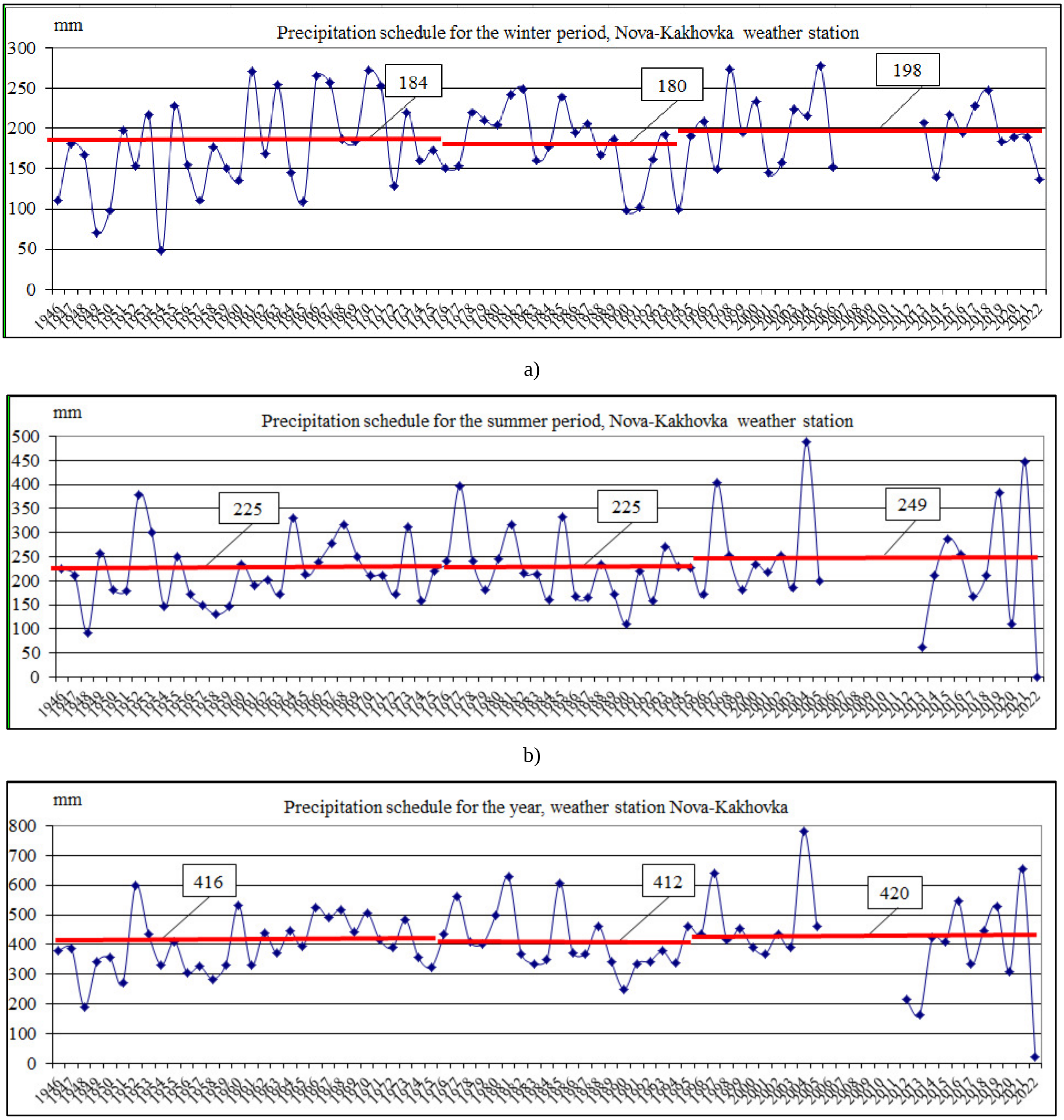
c)

Fig. 3. Precipitation schedule according to data from the Velyka Oleksandrivka weather station:
a) winter period; b) summer period; c) for a year.

Taking into account the conducted analysis and the location of the weather stations, there is reason to say that the significant cause of flooding in the area covered by the weather stations Kherson (western part of Kherson region) and Velyka Oleksandrivka (north-western part of Kherson region) is precisely the natural factor - an increase in atmospheric precipitation. Particular attention should be paid to the increase in the amplitude of precipitation in the summer period in 2005, which exceeds the average value by 360 mm, which was practically not observed in previous years.

Analysis of the multi-year dynamics of precipitation at the Nova Kakhovka and Nizhni Syrogoza weather stations. A comparison of the long-term seasonal irregularity of precipitation at the Nova Kakhovka weather station (Fig. 4) shows an increase in the average annual precipitation from 416 mm in the period 1946-1975 to 420 mm in 1996-2022. Moreover, in the winter period (the most threatening, in terms of the replenishment of groundwater by atmospheric precipitation), the total increase in average precipitation was 14 mm, and in the summer - 24 mm.

.



a)

b)

c)

Fig. 4. Precipitation schedule according to the data of the Nova Kakhovka weather station: a) winter period; b) summer period; c) for a year.

A comparison of the long-term seasonal unevenness of precipitation according to the data of the Nizhny Sirogoza weather station (Fig. 5) shows a gradual increase in the average amount of precipitation over seventy-five years (98 mm), which indicates a slight increase in the total natural load. Moreover, in the winter period (the most threatening, in terms of replenishment of groundwater by atmospheric precipitation), the total increase in average precipitation was 24 mm, in the summer period the increase was 54 mm.

.

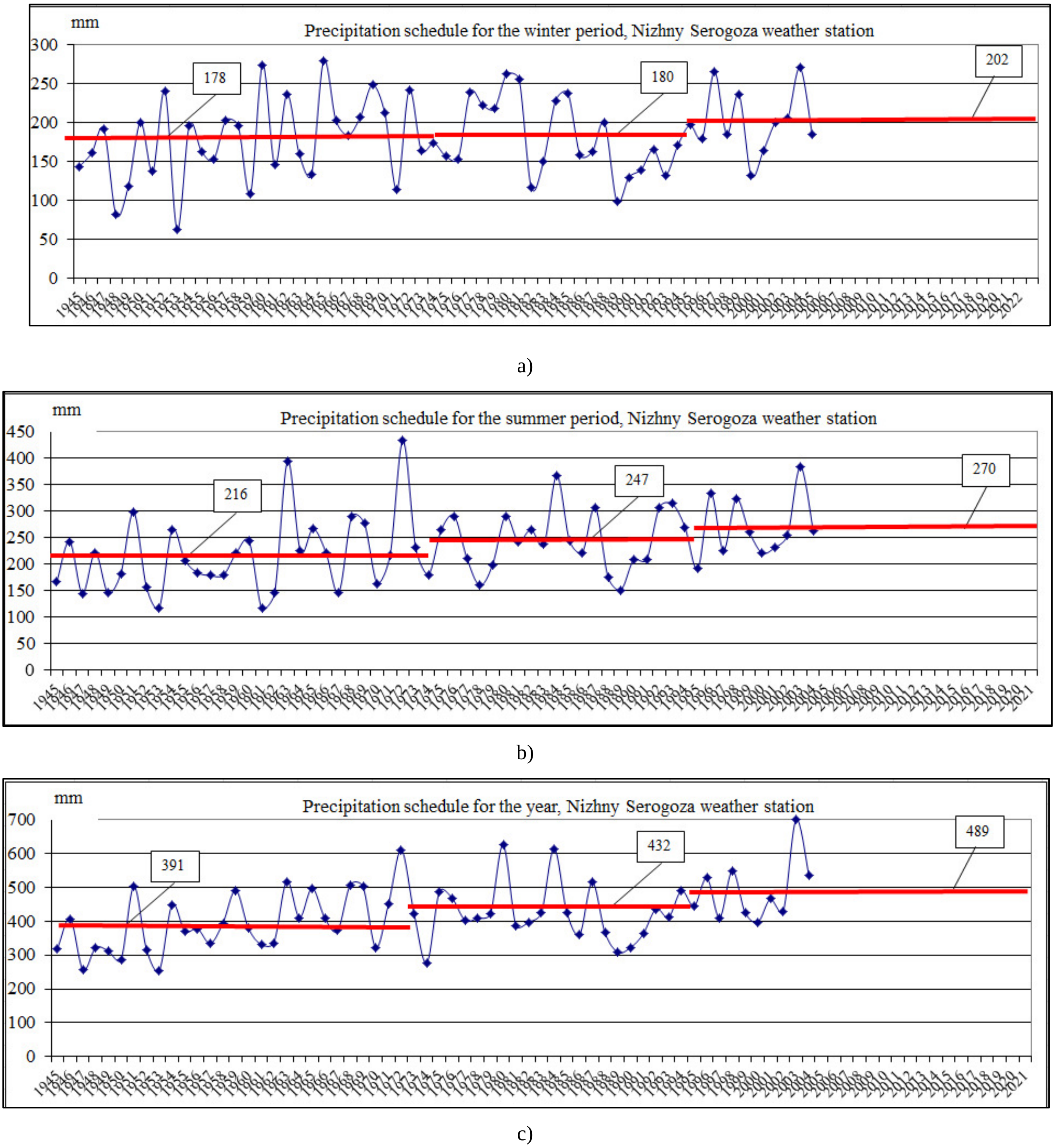


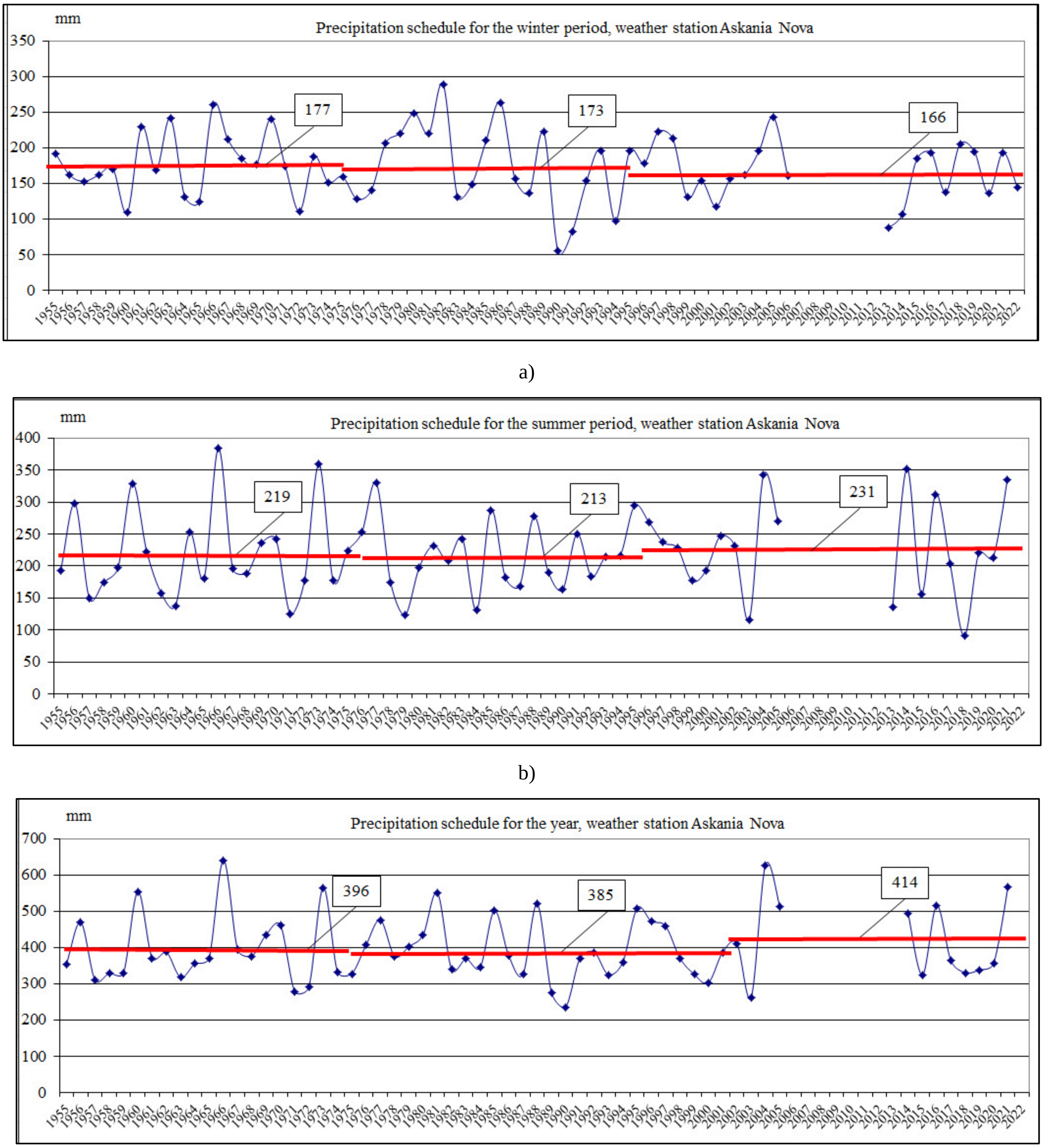
Fig. 5. Precipitation schedule according to the data of the Nizhny Sirogoza weather station: a) winter period; b) summer period; c) for a year.

Taking into account the conducted analysis and the location of weather stations, there is reason to say that one of the reasons (natural factors) of flooding in the area covered by the Nova Kakhovka and Nizhni Syrogozy weather stations in the recent period (1996 - 2022) is also an increase in atmospheric precipitation. At the same time, the amplitude of precipitation in the summer period in 2005 is increasing, which exceeds the average value by 240 mm, which was practically not observed in previous years.

Analysis of long-term precipitation dynamics of the Askania-Nova and Khorly weather stations. Weather stations Askania-Nova and Nizhny Sirogozy are located in the Kherson region, on the left bank of the Dnipro River.

The analysis of long-term precipitation dynamics was carried out for the following periods: 1955-1975, 1976-1995, 1996-2022.

A comparison of the long-term seasonal unevenness of precipitation according to the data of the Askania-Nova weather station (Fig. 6) shows a slight increase in the average amount of precipitation over twenty years (18 mm), which indicates an increase in the overall atmospheric load. Moreover, in the winter period (the most threatening, in terms of replenishment of groundwater by atmospheric precipitation), there was a slight (11 mm) decrease in average precipitation, in the summer period the increase was 12 mm.



a)

b)

c)

Fig. 6. . Precipitation schedule according to the data of the Askania-Nova weather station:
a) winter period; b) summer period; c) for a year.

Analyzing the comparison of precipitation according to the data of the Khorla weather station (Fig. 7), we can see that in the period 1976-1995, there was a decrease in the average amount of precipitation for many years by 35 mm. In the last period (1996 - 2022), the reverse process is observed (an increase in the average annual rainfall by 65 mm).

Moreover, in the winter period (the most threatening period, in terms of replenishment of groundwater by atmospheric precipitation), the total increase in average precipitation compared to the period 1976-1995 was 34 mm, and in the summer - 15 mm.

.

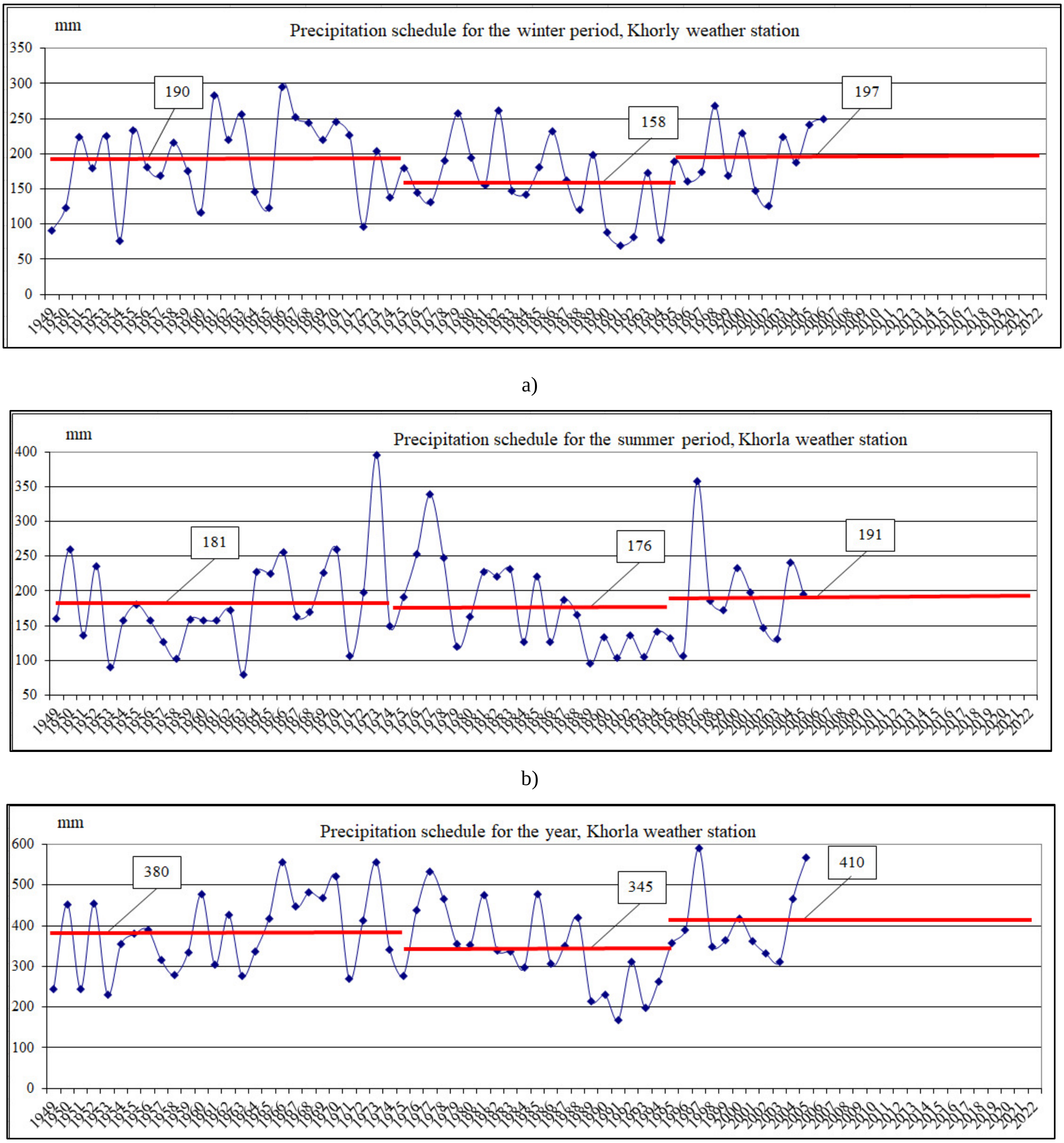


Fig. 7. Precipitation schedule according to data from the Khorly weather station:
a) winter period; b) summer period; c) for a year.

Taking into account the conducted analysis and the location of the Askania-Nova and Khorly weather stations, there are reasons to say that a small increase in the average amount of precipitation is not able to significantly affect the acceleration of flooding processes in this area.

CONCLUSIONS

Long-term analysis of precipitation in the Kherson region indicates an increase in natural load on the right bank of the Dnieper River (weather stations Kherson, Velyka Oleksandrivka) and in the north of the Kherson region (Nova Kakhovka, Nyzhni Sirohozy), which is one of the important natural factors increasing the risk of flooding. At the same time, on the left bank in certain areas (weather stations Askania-Nova, Khorly), there has not been a significant increase in the average amount of precipitation, which suggests the possibility of long-term anthropogenic influence on the flooding process.

Analysis of precipitation showed in some cases an increase in their amplitude in recent years (Kherson weather station 1998, 2004, 2010; Velyka Oleksandrivka weather station 2004, Askania-Nova weather station 2004). This leads to years with an increased risk of flooding and inundation of corresponding areas, as observed in the Kherson region in 1997-1998 and 2004-2005.

Analysis of seasonal unevenness of precipitation revealed an increase in the average amount of precipitation in the winter period over twenty years (Kherson weather station by 33 mm, Velyka Oleksandrivka weather station by 29 mm), which is the most threatening in terms of groundwater replenishment through infiltration.

REFERENCES

[1] Kuzmych, L., Voloshin, M., Kyrylov, Y., Dudnik, A., & Grinenko, O. (2023). Development of Neural Network Control and Software for Dispatching Water Distribution for Irrigation. CEUR Workshop Proceedings, 3624, pp. 352–367 chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ceur-ws.org/Vol-3624/Paper_29.pdf

[2] Turcheniuk, V., Rokochinskiy, A., Kuzmych, L., Volk, P., Prykhodko, N., (2023). Formation of a Favorable Filtration Regime of Soils in Saline Areas of the Danube Delta Rice Irrigation Systems. Archives of Hydro-Engineering and Environmental Mechanics. Volume 70, Issue 1, P.115–128. https://doi.org/10.2478/heem-2023-0008

[3] Korobiichuk, I., Kuzmych L, Kvasnikov, V., Nowak, P. The use of remote ground sensing data for assessment of environmental and crop condition of the reclaimed land // Advances in intelligent systems and computing (AISC), volume 550, ICA 2017: Automation 2017, pp 418-424 DOI: 10.1007/978-3-319-54042-9_39.

[4] Alfonso, L., Lobbrecht, A., Price, R. (2010) Optimization of water level monitoring network in polder systems using information theory. Water Resources Research, 46 (12), p. 1–13.

[5] Basharin D, Polonsky A, Stankunavichus G. (2016) Projected precipitation and air temperature over Europe using a performance-based selection method of CMIP5 GCMs. Journal of Water and Climate Change. 2016;7(1), p.103–113.

[6] Kuzmych L., Voropay G., Moleshcha N., Babitska O. (2021): Improving water supply capacity of drainage systems at humid areas in the changing climate. Archives of Hydro-Engineering and Environmental Mechanics. Vol. 68. No. 1: 29–40.

[7] Kuzmych,L., Furmanets,O., Usatyi,S., Kozytskyi,O., Mozol,N., Kuzmych,A.,Polishchuk,V. & Voropai,H.(2022).Water Supply of the Ukrainian Polesie Ecoregion Drained Areas in Modern Anthropogenic Climate Changes. Archives of Hydro-Engineering and Environmental Mechanics,69(1) 79-96. https://doi.org/10.2478/heem-2022-0006

[8] Schultz Bart (2008) Water management and flood protection of the polders in the Netherlands under the impact of climate change and man-induced changes in land use. Journal of water and land development, No. 12, p.71–94.

[9] Shang, S.H. (2014) A general multi-objective programming model for minimum ecological flow or water level of inland water bodies. Journal of Arid Land, 7 (2), p. 166-176.

[10] Su, X., Chiang, P., Pan, S., Chen, G., Tao, Y., Wu, G., Wang, F., Cao, W. (2019) Systematic approach to evaluating environmental and ecological technologies for wastewater treatment. Chemosphere, 218, p. 778-792.

[11] Van Overloop, P.J. (2006) Drainage control in water management of polders in the Netherlands. Irrigation and Drainage Systems, 20 (1), p. 99-109.

[12] Andrić I., Koc M. and Al-Ghamdi S. G. 2019 A review of climate change implications for built environment: Impacts, mitigation measures and associated challenges in developed and developing countries J. Clean. Prod. 211 83-102

[13] Korobiichuk I., Kuzmych L., Kvasnikov V., 2019. The system of the assessment of a residual resource of complex technical structures, MECHATRONICS 2019: Recent Advances Towards Industry 4.0, 350–357. https://doi.org/10.1007/978-3-030-29993-4−43

[14] Ahmed M. R., Hassan Q. K., Abdollahi M. and Gupta A. 2020 Processing of near real time land surface temperature and its application in forecasting forest fire danger conditions Sensors 20 984

[15] Rózsás Á., Kovács N., Gergely Vigh L. and Sýkora M. (2016). Climate change effects on structural reliability in the Carpathian Region Q. J. Hungarian Meteorol. Serv. 120 103-25

[16] Chen K., Blong R. and Jacobson C. (2003). Towards an integrated approach to natural hazards risk assessment using GIS: With reference to bushfires Environ. Manage. 31 546-60

[17] Rokochinskiy, A., Kuzmych, L., Volk, P. (Eds.). (2023). Handbook of Research on Improving the Natural and Ecological Conditions of the Polesie Zone. IGI Global. https://doi.org/10.4018/978-1-6684-8248-3

[18] Rokochinskiy A., Volk P., Kuzmych L., Turcheniuk V., Volk L. and Dudnik A. [2019]. Mathematical Model of Meteorological Software for Systematic Flood Control in the Carpathian Region 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), pp. 143-148. https://doi.org/10.1109/ATIT49449.2019.9030455

[19] Yakymchuk A., Kuzmych L., Skrypchuk P., Kister A., Khumarova N., Yakymchuk Y.. (2022). Monitoring in Ensuring Natural Capital Risk Management: System of Indicators of Socio-Ecological and Economic Security. 16th International Conference Monitoring of Geological Processes and Ecological Condition of the Environment, Nov 2022, Volume 2022, p.1 – 5. https://doi.org/10.3997/2214-4609.2022580047

[20] "Continental Climate". Encyclopedia of the Atmospheric Environment. Manchester Metropolitan University. Archived from the original on 2009-04-27.

[21] Senker, Cath (3 May 2018). Temperate Climates. Raintree Publishers. p. 5. ISBN 9781474738408. Retrieved 23 June 2023.

[22] "Weather & Climate Change: Climates around the world". Education Scotland. Archived from the original on 14 April 2016.

[23] Peel, M. C.; Finlayson B. L. & McMahon, T. A. (2007). "Updated world map of the Köppen–Geiger climate classification" (PDF). Hydrol. Earth Syst. Sci. **11** (5): 1633–1644. Bibcode:2007HESS...11.1633P. doi:10.5194/Hess-11-1633-2007. ISSN 1027-5606.

[24] Kottek, Markus; Grieser, Jürgen; Beck, Christoph; Rudolf, Bruno; Rubel, Franz (2006). "World Map of the Köppen-Geiger climate classification updated" (PDF). Meteorologische Zeitschrift. **15** (3): 259–263. Bibcode:2006MetZe..15..259K. doi:10.1127/0941-2948/2006/0130.

[25] "What Is a Continental Climate?". WorldAtlas. 2019-05-21. Retrieved 2022-04-04.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 105–110 DOI: 10.15439/2024F3810 ISSN 2300-5963 ACSIS, Vol. 41

Toxic Molecule Classification Using Graph Neural Networks and Few Shot Learning.

Bhavya Mehta*, Kush Kothari†, Reshmika Nambiar‡ and Seema Shrawne¶
Department of Computer Engineering and Information Technology, Veermata Jijabai Technology Institute
Mumbai, India.
Email: *bdmehta_b19@ce.vjti.ac.in, †kmkothari_b19@ce.vjti.ac.in, ‡rsnambiar_b19@ce.vjti.ac.in, ¶scshrawne@ce.vjti.ac.in

***Abstract*—Traditional methods like Graph Convolutional Networks (GCNs) face challenges with limited data and class imbalance, leading to suboptimal performance in graph classification tasks during toxicity prediction of molecules as a whole. To address these issues, we harness the power of Graph Isomorphic Networks, Multi Headed Attention and Free Large-scale Adversarial Augmentation separately on Graphs for precisely capturing the structural data of molecules and their toxicological properties. Additionally, we incorporate Few-Shot Learning to improve the model's generalization with limited annotated samples. Extensive experiments on a diverse toxicology dataset demonstrate that our method achieves an impressive state-of-art AUC-ROC value of 0.816, surpassing the baseline GCN model by 11.4%. This highlights the significance of our proposed methodology and Few-Shot Learning in advancing Toxic Molecular Classification, with the potential to enhance drug discovery and environmental risk assessment processes.**

***Index Terms*—Graph Neural Networks, Graph Isomorphic Network, Multi Headed Attention, Graph Data Augmentation, Few Shot Learning, Toxicity Prediction.**

I. INTRODUCTION

Toxicological assessment of molecular compounds plays a pivotal role in drug discovery, environmental risk assessment, and chemical safety evaluation. Accurate prediction of a molecule's toxicity is crucial in ensuring the development of safe and effective drugs while minimizing potential harm to both human health and the environment.

Traditional methods of toxic molecule detection [[1],[2]] possess some inherent limitations. This is because conducting experiments to synthesize a compound and then analyzing its toxicity is time-consuming and often very expensive. It consumes a lot of resources and is not feasible for large-scale testing of molecules.

A number of approaches based on machine learning have also recently been proposed. The methods described above use several molecular characteristics, such as their physical and chemical properties, to predict their toxicity. However, a present problem in the field is lack of sufficient labelled data, due to the difficulties faced in synthesizing and testing new molecules, as explained above. Moreover, often these machine learning techniques only look at certain numerical properties of the molecules and fail to take into consideration the structural aspects of the molecule.

In recent literature, a lot of research is being done in representing molecules as graphs and processing them through Graph Neural Networks (GNNs). While this method does not tackle the low-data scenario we often face in toxicity prediction, newer methods have integrated few-shot learning into GNNs, like the Adaptive Step Model-Agnostic Meta-Learner (AS-MAML). We believe that this intersection of graph-embedding algorithms and few-shot learning is key to creating effective models for molecular toxicity prediction.

The research problem addressed in this paper is to investigate and propose enhancements to the GNN-specific few-shot learning technique in order to achieve favorable results in the toxicity prediction task on the Tox21 data set under the few shot learning scenario.

II. BACKGROUND

Before delving into the specifics of the architecture, it is essential to provide some background information that will be helpful for better understanding.

A. Few Shot Learning (FSL)

As suggested by Vinyals et al. in [3], FSL is the ability of an algorithm to generalize well from limited data points with supervised information available for every class. To achieve this, we employ Model-Agnostic Meta-Learning (MAML) given by Finn et al. which aims to find a good initialization for the model parameters θ, for rapid adaptation to novel classes with only a few labeled examples. This is done by optimizing the model's performance on a set of meta-training experiments, where each task simulates a few-shot learning scenario.

B. Adaptive Step Model Agnostic Meta Learning (AS-MAML)

Introduced by Ma et al. in [5], it is a meta-learning technique that builds upon the MAML[4] algorithm by introducing an Adaptation Controller that employs reinforcement learning techniques to determine the optimal step size and when to stop the adaptation process. A StopController model, incorporating LSTM [6] layers and a sigmoid function, estimates the probability of stopping the adaptation process based on the training loss and embedding quality. This addresses the challenge of finding the optimal learning rate and step size in MAML-based meta-learning approaches.

C. Graph Convolution Network (GCN)

GCNs were introduced as a way to extend convolutional neural networks (CNNs) to handle irregular and non-Euclidean data. Kipf and Welling in [7] mentions that the key challenge

©2024, PTI

in processing graph data is that the number of nodes and their connectivity can vary widely from one graph to another. GCNs address this challenge by learning to exploit the local neighborhood information of each node in the graph to make predictions. The core idea behind GCNs is to perform node feature aggregation through a series of graph convolutions, enabling nodes to gather information from their neighbors and incorporate it into their own representations.

D. Graph Isomorphic Network (GIN)

Graph Isomorphic Networks by Xu et al. in [8] are a class of deep learning models designed for graph classification tasks. Unlike traditional GCNs, GINs do not rely on graph structure during message passing, making them more flexible and suitable for various graph types. The core idea behind GINs is to employ an aggregation function that is permutation-invariant to the node ordering, ensuring that the model produces the same output regardless of how the nodes are arranged. This property allows GINs to capture the global graph information effectively and provide more robust representations for graph classification tasks.

E. Free Large-scale Adversarial Augmentation on Graphs (FLAG)

It is a technique for enhancing graph data to imprive GNNs' performance. FLAG by Kong et al. in [9] suggests augmenting node properties rather than modifying graph topological structures, which is where the majority of existing graph regularizers concentrate their efforts. It improves generalization to out-of-distribution samples by iteratively enhancing node characteristics with gradient-based adversarial perturbations during training. This makes the model invariant to tiny fluctuations in input data. Adversarial data points are created and then inserted into the training data as part of the adversarial training process. The objective of this min-max optimization problem is to minimize the objective function while keeping the perturbation within a predetermined bound.

III. Related Works

Some of the earliest works in toxicity prediction include DeepTox by Mayr et al., who used chemical properties of these compounds fed into a Deep Neural Network to predict their toxicity. By using this method and ample of labelled data Mayr et al., manages to achieve an 0.92 AUC value. Alperstein et al. introduced All SMILES VAE [11], a generative model which uses variational autoencoders (VAEs) for generating SMILES strings using stacked RNNs. The model surpassed state-of-the-art methods and achieved an ROC-AUC score of 0.871 on the dataset. Censnet by [12] learns node and edge features through the use of novel propagation rules while switching the roles of nodes and edges. The method attains about 0.79 AUC score at most on the Tox21 dataset when tested under various splitting scenarios. Zhou et al. proposed Uni-Mol [13], a framework that incorporates the pretraining of transformers in order to use 3D information. It was evaluated on Tox21 as a downstream task and outperformed several methods, achieving an ROC-AUC score of 0.796.

Graph Multiset Transformer (GMT) [14] adopts a novel pooling method wherein multi-head attention is used for learning node interaction based on task relevance. An AUC score of about 0.773 was obtained on Tox21. Meta-MGNN by Guo et al. employs meta-learning to learn molecular representations under few-shot settings. It uses pretrained GNNs and leverages additional tasks to be optimised. When tested on Tox21, an AUC score of 0.769 was obtained under the one-shot setting and about 0.78 under the five-shot setting, outperforming several baseline models. However very few works have obtained significant results in few shot domain with graphs.Chen et al. in [16] achieves an average ROC-AUC score of 0.757 employing the Mean Teacher Semi-Supervised ML Algorithm, which is a 6% increase over GCN models trained using supervised and conventional ML techniques. However for low data scenarios, very few works have been able to get significant results.

IV. Tox21 Dataset

Tox21 is a dataset containing measurements of toxicity of 12 thousand molecules against 12 target proteins. It aims to help analyse the performance of models in predicting the biochemical activity of compounds using their chemical structure. We use the AhR sub-dataset from Tox21 that focuses on chemicals' interactions with this Aryl hydrocarbon Receptor, a ligand-activated transcription factor that is essential for the toxic response to toxins and medications. The dataset is open source and can be downloaded from Tox21 AHR[1].

Each chemical compound in this collection is represented as a graph, with atoms serving as nodes and chemical bonds between atoms serving as edges. Molecules' structural information is preserved in the graph representation, making it ideal for GNN-based approaches that can efficiently handle graph-structured data. By learning from the graph structure and associated node features, GNNs can discern complex relationships and identify key structural characteristics associated with toxic and non-toxic compounds.

V. Baseline Model

The initial configuration we are evaluating serves as the baseline, which is the standard GCN + AS-MAML model utilizing the few-shot learning setup detailed earlier. While we remain consistent with the framework described in the paper, there is one notable difference: we do not employ distinct classes for training and validation. This configuration consists of three successive layers: a GCN convolution layer, followed by a TopK Pooling Layer [17], each with a hidden layer dimension of 128. The Baseline Architecture is shown in Fig.1, has a validation accuracy of 65.02% and an AUC-ROC value of 0.732 on Tox21 AhR data.

---

[1] http://bioinf.jku.at/research/DeepTox/tox21.html

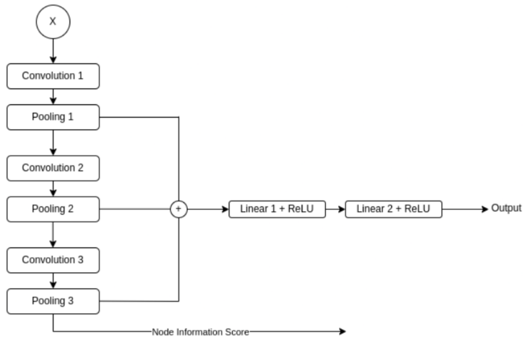


Fig. 1: The image illustrates the baseline sub-architecture.The resulting output vector is subsequently fed into a binary sigmoid classifier. The obtained Node Information Score is utilized by the FSL Reinforcement Learning Agent to optimize gradients and weights, thereby achieving a faster convergence rate.

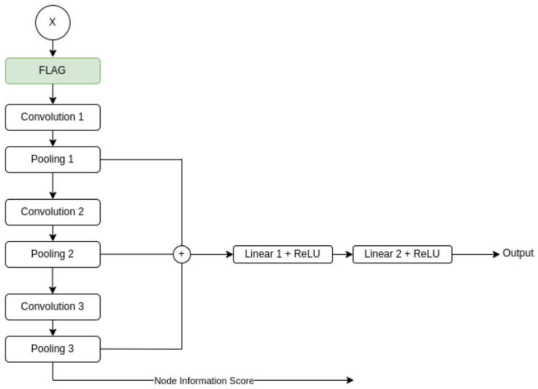


Fig. 2: Proposed sub-architecture of FLAG+GCN based classification model.

VI. PROPOSED ARCHITECTURES

In this research paper, we introduce and empirically evaluate three distinct architectural frameworks, each of which outperforms the baseline model in terms of achieved results. These three novel architectures systematically introduce variations across distinct components of the baseline model's structure the body, the input and the output, enhancing the model's capacity to capture intricate patterns and further enriching its learning capabilities.

A. Augmenting Input data using FLAG

1) Architecture: The first suggested setting adds a preprocessing step of FLAG in order to augment the data being fed into the model as shown in Fig.2. This adds perturbations to node features and provides greater variations in novel tasks available for few-shot learning. The aim of FLAG is to generate additional realistic graph instances that maintain the underlying distribution of the original data, effectively expanding the dataset and boosting model generalization.

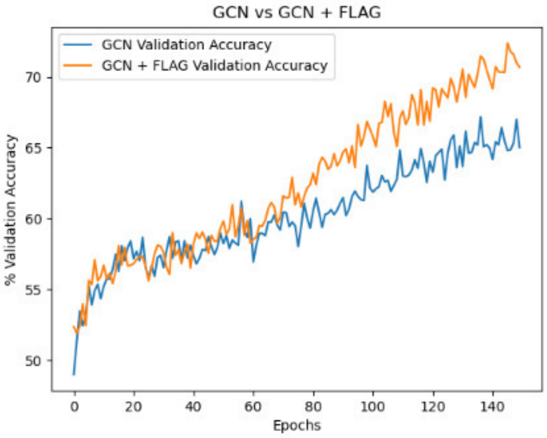


Fig. 3: A comparision of validation accuracy for GCN vs GCN+Flag method.

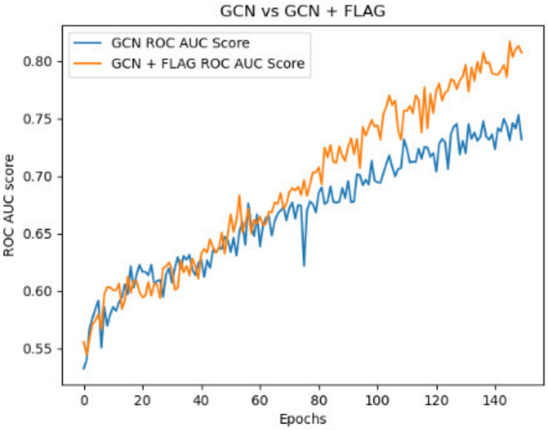


Fig. 4: A plot of ROC score for GCN vs the proposed GCN+FLAG sub-architecture.

2) Experimental Results: As shown in Fig.3 and Fig.4, we observe significant improvement in performance on the use of FLAG. This could be because it preserves the structural integrity of the graphs since only node features are modified. Tox21 is a molecular dataset where random changes in structure may not be realistic. Also, certain constraints are imposed on the perturbations, further improving reliability. FLAG has been found to be effective for discrete features which are commonly encountered in molecular data. In addition to this, FLAG improves generalization, robustness and data diversity, and is computationally efficient with validation accuracy of **70.68%** and validation AUC-ROC score of **0.806,** both of them greater than the baseline GCN model.

B. Replacing GCNs with GINs

1) Architecture: The proposed novel architecture (Refer Fig.5) introduces a modification to the AS-MAML algorithm[5] by replacing the three Graph Convolutional Network (GCN)[7] components with three Graph Isomorphism

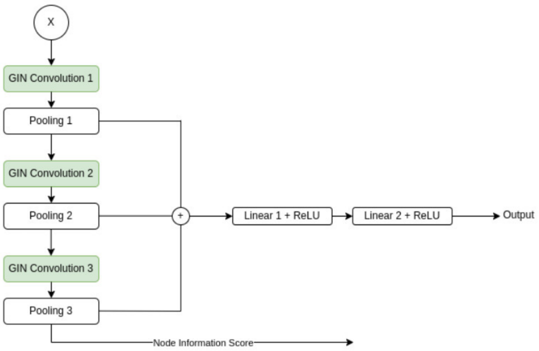


Fig. 5: Proposed sub-architecture of GIN based classification model.

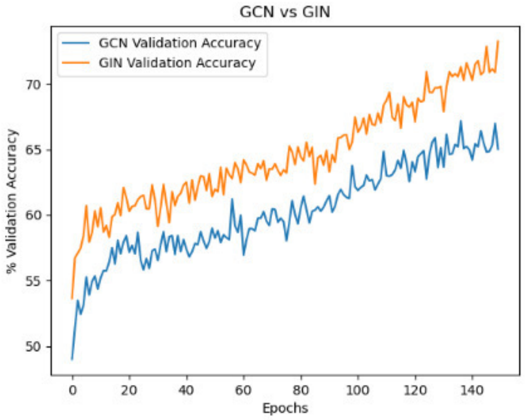


Fig. 6: A plot of validation accuracies of the baseline and the proposed GIN sub-architecture.

Network (GIN) convolution operators. The GIN operator is anticipated to offer enhanced expressibility, resulting in improved hidden layer embeddings compared to the original GCN-based approach. In the forward() operation of the GIN model, a multi-layer perceptron (MLP) is employed, comprising four hidden layers, each containing ten perceptrons. The output layer, consistent with the GCN model, consists of 128 perceptrons.

2) Experimental Results: In the GIN model, we observed significant improvements in validation accuracy and ROC AUC score compared to the baseline GCN model, even at an early stage. The final validation accuracy of **73.23%** as shown in Fig.6 and ROC score of **0.816** as shown in Fig.7 can be attributed to the enhanced expressiveness of the Graph Isomorphism Network Operator utilized in GIN. Notably, this improved accuracy is consistently maintained over the course of 150 epochs, suggesting that while GIN may not necessarily provide an advantage in achieving higher-quality training results, it excels at capturing relevant task information with fewer epochs.

C. Enhanced Aggregation and Extraction using Weighted Multi Headed Attention (MHA).

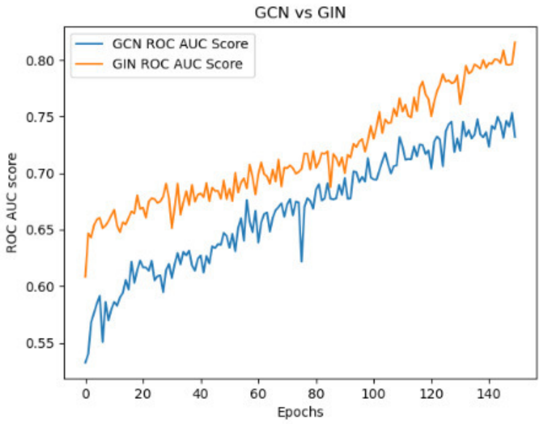


Fig. 7: A plot of ROC score for GCN vs the proposed GIN sub-architecture.

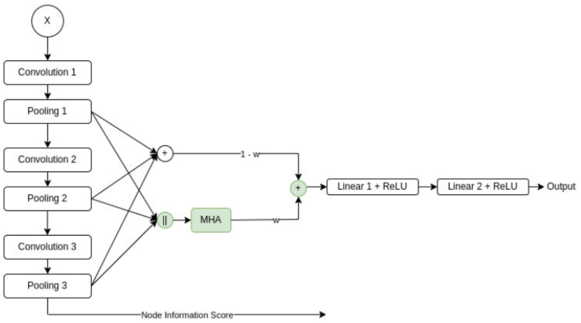


Fig. 8: Proposed sub-architecture of GCN+MHA based classification model.

1) Architecture: The final setting as shown in Fig.8, involves adding a Multi Attention Head (MAH) mechanism as an operator in the last part of the AS-MAML model. The baseline GCN carries out a normal aggregation of the outputs of the Relu layers and passes it to a binary classification network as displayed in the baseline figure above which might be unable to extract all necessary information or give weightage to the important ones. By inculcating a MAH layer, we attempt to change this fact and try to make the best out of the convolutions. MAH takes these three values as input to the Key, Value and Query fields to identify patterns of significance. The weight factor *"w"* adds an extra bias to the inclusion of the attention layer while performing regularization.

2) Experimental Reults: Upon examination of the graphs presented in Fig.9 and Fig.10, it becomes evident that the adapted model consistently outperforms the baseline counterpart during the validation phase. The GCN+MAH model achieves a notable validation accuracy of approximately **69.62%**, showcasing a significant improvement over the baseline's attainment of 65%. Additionally, a discernible discrepancy of 0.055 units is observed in the AUC-ROC values, further substantiating the effectiveness of the modified architecture. This enhanced performance of the GCN+MAH

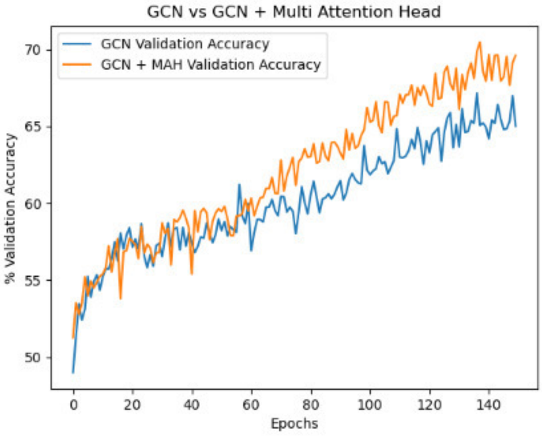


Fig. 9: A plot of validation accuracies of the baseline and the proposed GCN+MAH sub-architecture.

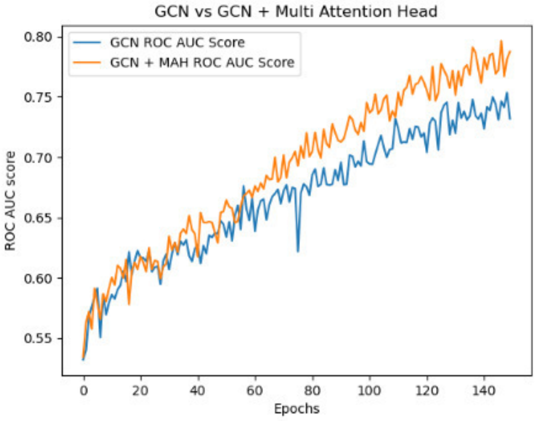


Fig. 10: A plot of ROC Score of the baseline and the proposed GCN+MAH sub-architecture.

model can be attributed to the model's abilty to simultaneously consider multiple attention patterns, enabling it to capture intricate data relationships and patterns more effectively. The experiments have been carried out with the weight factor "w" having a value of 0.4 which is another hyper parameter that we introduce.

VII. Discussion and Evaluations

Within this section, we elaborate on the training regimen and specifications pertinent to Few-Shot Learning (FSL), followed by a comprehensive evaluation of their collective outcomes along with an exhaustive assessment of the best method employed.

A. *Few Shot Learning Specifications*

In Table I, we list some of the common tunable parameters in a few shot learning scenarios and state the settings used for our testing.

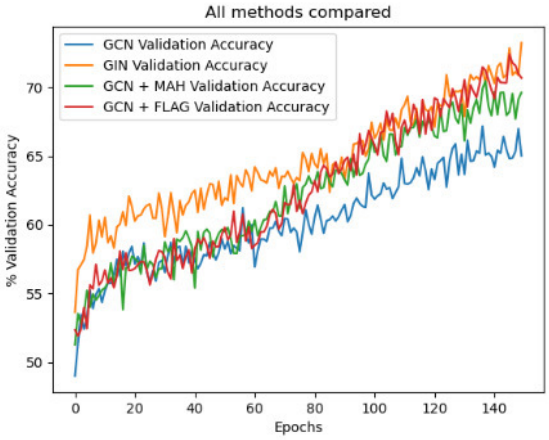


Fig. 11: Plot of validation accuracy for baseline and all three proposed sub-architectures.

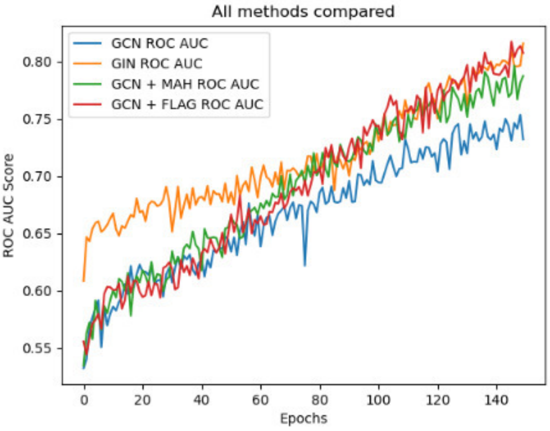


Fig. 12: Plot of ROC score for baseline and all three proposed sub-architectures.

B. *Comprehensive Analysis*

Table II and Fig.11, Fig.12, highlight the enhanced expressibility and feature representation ability of GIN operators, attention modules and data augmentation algorithms.

By adopting these modified architectures, which have the potential to advance the state-of-the-art results. We anticipate achieving improved performance and predictive capabilities in the context of the AS-MAML algorithm for few-shot learning tasks, particularly in toxicological classification of molecular compounds.

GIN's superiority is attributed to its order-agnostic aggregation operation, which ensures robustness and insensitivity to changes in node positions. In contrast, GCN's performance is influenced by the order of nodes in the neighborhood, making it more sensitive to node ordering. Another aspect contributing to GIN's efficacy is its higher expressiveness compared to GCN. While GCN focuses on local information within fixed neighborhoods, it faces limitations in capturing higher-order

TABLE I: Tunable Parameters and Values for FSL setting

Tunable Parameter	Significance	Value in Experiments
Train Shot	The number of labeled examples from the training set used for adapting the model during the few-shot learning process.	10
Validation Shot	The number of labeled examples from the validation set used for fine-tuning or evaluating the model during the few-shot learning process.	10
Train Query Set	The set of unlabeled examples from the training set that are used for prediction or evaluation after model adaptation.	15
Validation Query Set	The set of unlabeled examples from the validation set that are used for prediction or evaluation after fine-tuning or model evaluation.	15
Epochs	Count of how many times the full dataset was run through the model during training.	150
Learning Rate (Outer Loop)	A hyperparameter that determines the step size or rate at which the model's parameters are updated during the training process.	0.001
Learning Rate (Inner Loop)	A hyperparameter that determines the step size or rate at which the model's parameters are updated during the inner loop training process.	0.01

TABLE II: The table presents a detailed comparison between the accuracy and ROC values of the proposed methods.

Model and Algorithm	Validation Accuracy	Δ Accuracy Score	ROC-AUC Score	Δ ROC-AUC Score
GCN (Baseline)	65.02 %	-	0.732	-
GCN + FLAG	70.68 %	**+5.66 %**	0.806	**+0.074**
GIN	73.23 %	**+8.21 %**	0.816	**+0.084**
GCN + MHA	69.62 %	**+4.6 %**	0.787	**+0.055**

graph structures. Conversely, GIN's iterative message passing mechanism enables it to encompass more intricate and global structural patterns, making it more adept at handling complex molecular graphs.

Moreover, GIN's passing of the Weisfeiler-Lehman (WL) test [18], a theoretical measure of a GNN's expressive power, further validates its strength as a graph neural network. The WL test checks whether a GNN can differentiate non-isomorphic graphs with the same initial node labels. GIN's successful performance on this test showcases its stronger representational capacity compared to GCN and other methods.

Nevertheless, it's essential to acknowledge that the efficacy of GNN architectures may vary based on the specific dataset and task, with hyperparameter tuning and data preprocessing also impacting their overall performance.

VIII. CONCLUSION

In conclusion, the GIN method presented in this research paper establishes a new benchmark in drug discovery and toxicity prediction using the Tox21 data, exhibiting remarkable improvements of 8.21% in accuracy and 11.4% in ROC performance compared to existing GCN methods. Moreover, its exceptional performance in low labeled data scenarios, surpassing all other given methods, underscores its robustness and practicality. This novel approach holds immense promise for researchers and practitioners in pharmaceutical and chemical industries, providing valuable insights and advancements in these fields while inspiring further exploration and adoption of graph neural network-based methodologies for addressing real-world challenges

REFERENCES

[1] Y. Jiang, D.-W. Sun, H. Pu, and Q. Wei, "Surface enhanced raman spectroscopy (sers): A novel reliable technique for rapid detection of common harmful chemical residues," *Trends in Food Science & Technology*, vol. 75, pp. 10–22, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924224417307045

[2] M. E. E. Alahi and S. C. Mukhopadhyay, "Detection methodologies for pathogen and toxins: A review," *Sensors*, vol. 17, no. 8, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/8/1885

[3] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *CoRR*, vol. abs/1606.04080, 2016. [Online]. Available: http://arxiv.org/abs/1606.04080

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *CoRR*, vol. abs/1703.03400, 2017. [Online]. Available: http://arxiv.org/abs/1703.03400

[5] N. Ma, J. Bu, J. Yang, Z. Zhang, C. Yao, Z. Yu, S. Zhou, and X. Yan, "Adaptive-step graph meta-learner for few-shot graph classification," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1055–1064. [Online]. Available: https://doi.org/10.1145/3340531.3411951

[6] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *CoRR*, vol. abs/1909.09586, 2019. [Online]. Available: http://arxiv.org/abs/1909.09586

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[8] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *CoRR*, vol. abs/1810.00826, 2018. [Online]. Available: http://arxiv.org/abs/1810.00826

[9] K. Kong, G. Li, M. Ding, Z. Wu, C. Zhu, B. Ghanem, G. Taylor, and T. Goldstein, "Robust optimization as data augmentation for large-scale graphs," 2022.

[10] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," *Frontiers in Environmental Science*, vol. 3, Feb. 2016. [Online]. Available: https://doi.org/10.3389/fenvs.2015.00080

[11] Z. Alperstein, A. Cherkasov, and J. T. Rolfe, "All SMILES VAE," *CoRR*, vol. abs/1905.13343, 2019. [Online]. Available: http://arxiv.org/abs/1905.13343

[12] X. Jiang, P. Ji, and S. Li, "Censnet: Convolution with edge-node switching in graph neural networks," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 2656–2662. [Online]. Available: https://doi.org/10.24963/ijcai.2019/369

[13] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, and G. Ke, "Uni-mol: A universal 3d molecular representation learning framework," May 2022. [Online]. Available: https://doi.org/10.26434/chemrxiv-2022-jjm0j

[14] J. Baek, M. Kang, and S. J. Hwang, "Accurate learning of graph representations with graph multiset pooling," *CoRR*, vol. abs/2102.11533, 2021. [Online]. Available: https://arxiv.org/abs/2102.11533

[15] Z. Guo, C. Zhang, W. Yu, J. Herr, O. Wiest, M. Jiang, and N. V. Chawla, "Few-shot graph learning for molecular property prediction," *CoRR*, vol. abs/2102.07916, 2021. [Online]. Available: https://arxiv.org/abs/2102.07916

[16] J. Chen, Y.-W. Si, C.-W. Un, and S. W. I. Siu, "Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network," *Journal of Cheminformatics*, vol. 13, no. 1, Nov. 2021. [Online]. Available: https://doi.org/10.1186/s13321-021-00570-8

[17] C. Özdemir, "Avg-topk: A new pooling method for convolutional neural networks," *Expert Systems with Applications*, vol. 223, p. 119892, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423003937

[18] N. T. Huang and S. Villar, "A short tutorial on the weisfeiler-lehman test and its variants," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. [Online]. Available: https://doi.org/10.1109%2Ficassp39728.2021.9413523

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 111–116
DOI: 10.15439/2024F4351
ISSN 2300-5963 ACSIS, Vol. 41

Integrating Artificial Intelligence Techniques in Cell Mechanics

Muddasar Naeem
Department of Computer Engineering
Università Giustino Fortunato
Benevento Italy
Email: m.naeem@unifortunato.eu

Mario Fiorino
Politecnico di Torino, Italy
Email: mario.fiorino@polito.it

Pia Addabbo
Department of Computer Engineering
Università Giustino Fortunato
Benevento Italy
Email: p.addabbo@unifortunato.eu

Antonio Coronato
Department of Computer Engineering
Università Giustino Fortunato
Benevento Italy
Email: a.coronato@unifortunato.eu

***Abstract*—The Artificial Intelligence (AI) and Machine Learning (ML) techniques have been revolutionizing many subjects. The AI-empowered methods such as Reinforcement Learning (RL) and Deep Learning (DL) have been employed for various aspects of cell mechanics. This work reviews the state of art of AI and ML technologies that have been used to describe, analyze and predict the mechanics of cells as well as the use of numerical methods for cell mechanics. This review also considers the impact of utilizing physical constraints on the AI and ML models aiming at improved convergences during the training and validation phases. At the end, we will provide a statistical analysis of the reported studies and a discussion on the current challenges and future possibilities**

I. INTRODUCTION

THE DYNAMICS and structure of living cells are managed by the physical properties of the cytoskeleton [1]. But the cytoskeleton is the combination of complex biochemical circuits which controls its spatial organization and dynamics [2]. Untangling this interplay between biochemical and mechanical constraints is the major challenge faced when studying the physical biology of the cell. Traditional modeling techniques are learned mainly on intuition built on classical continuum mechanics, where conservation and symmetries laws control the variables which arise in these models and equations they follow [3]. However, cells are hierarchical, non-equilibrium dynamics, relying instead on distributed enzymatic activity, and non-classical structures [4]. Therefore, such characteristics may complicate system parameterization and coarse-graining with regard to a few collective simply-understood variables [5]. Although we have remarkable progress and research in the area of physical biology, but it is difficult to accurately predict the mechanical response of cells to biochemical perturbations.

Artificial Intelligence (AI) and Machine learning (ML) algorithms have the potential to overcome some of the key challenges faced by conventional modeling techniques by learning models directly from the statistics of data [6]. These modern technologies have useful applications in diverse areas like communication [7], [8], dynamic treatment regimes [9], risk management during nuclear examination [10], rehabilitation [11], health monitoring [12] The use of AI and ML is transforming the way biologists do research, demonstrate their findings, and utilize them to address problems in the field [13], [14] and many more. As science along with other fields have been becoming more interdisciplinary, researchers are taking benefits of AI and ML to solve challenging and emerging problems in cell mechanics and biology. The application of ML especially of Reinforcement Learning (RL) and Deep Learning (DL) in solving an array of research problems like drug discovery, genome analysis, classification of cellular images, and in correlating the genome and image data to electronic medical records. Moreover, AI-powered algorithms have other applications including protein function prediction, analysis of genome-wide association studies to investigate markers of disease, high-throughput microarray data analysis, and enzyme function prediction. In addition, modern AI and ML technologies have assisted the discovery of tools which are solving challenging research problems such as cell profile, DeepVariant, and Atomwise.

The objective of this work is to study cell related issues using AI and ML approaches to characterize the different aspects of the phenomena and highlight the mechanical, thermal, and chemical properties of the systems under scrutiny. The work is structured in different research lines to investigate the topics such as:

1- Analysis of discrete models at the microscale based on elements that exhibit multistability to study how the microscopic details are responsible for emergent phenomena at the meso and macroscale.

2- Study of the complex response of the cells' systems across the scales using AI and Finite Element-based numerical methods.

3- Studies related to experimental validation on adhered cell

©2024, PTI

cultures characterizing elastic properties, surface topography and adhesion phenomena of cells under specific load conditions to confirm the theoretical outcomes about the mechanical behavior.

4- Study of physics-based machine learning techniques that include information obtained from the developed models and the experimental validation and optimize the training phase to improve the output and predictive capabilities of these methods.

After a brief introduction on the scope of this work in section I, we will present an overview of the AI tools in section II. The section III will be the main section where we will present diverse applications of AI for cell mechanics. Then in the discussion section, we will present some of our findings and we will close this study with concluding remarks in section VI.

II. AI POWERED TECHNIQUES

In this section, we will introduce various AI technologies as also indicated in the Figure 1 that can be utilized for different aspects of cell mechanics.

A. Neural Networks

Neural networks (NNs) are computational models inspired by the human brain's ability to process information. They comprise interconnected units known as neurons, organized into layers: input, hidden, and output. These neurons engage in a learning process where they adjust their behavior based on the data they receive. Inputs are represented by numerical values, these are fed into neurons to convey features of the data. Each input is assigned a weight, signifying its impact on the neuron's output. These weights are adjusted during training to optimize the network's performance. A constant value known as bias is added to the weighted sum of inputs, allowing the neuron to learn even when inputs are zero. It helps in avoiding scenarios where the network would be overly influenced by certain features. Activation function is applied as a non-linear transformation to the weighted sum and enables neurons to learn intricate patterns in the data. This non-linearity is crucial for the network to capture complex relationships and avoid being limited to linear mappings.

B. Convolutional Neural Networks (CNNs)

CNNs are advanced forms of NNs tailored for grid-like data, such as images or audio. Filters or convolutions are used in CNNs to extract meaningful features from the input. A conventional CNN consists of some convolutional layers, then some pooling layers, and also fully connected layers.

The convolutional layers use filters to extract features from the input data such as textures or edges. While the use of pooling layers in CNNs is to minimize the dimensionality of the data, avoiding overfitting and simplifying the information.

Finally, the function of fully connected layers in CNNs is the features extraction by convolutional layers for final predictions.

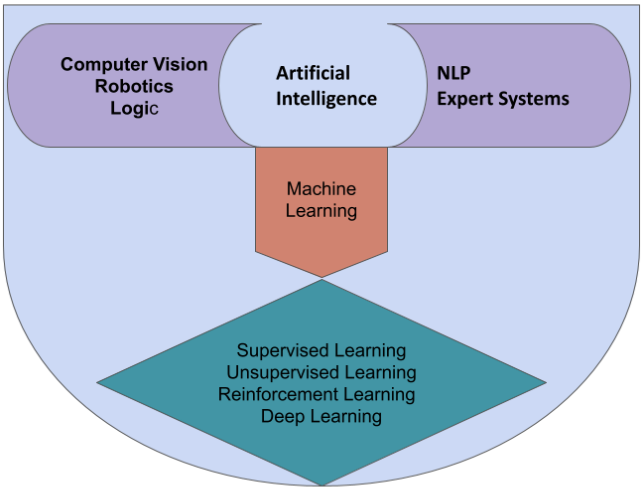


Fig. 1. An overview of AI-powered tools

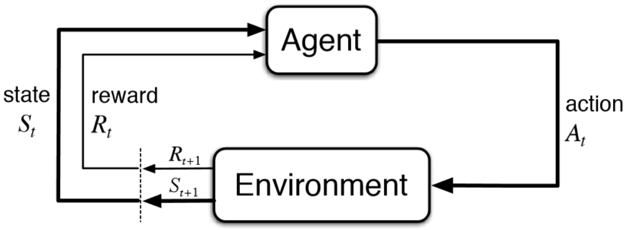


Fig. 2. A demonstration of RL workflow [15]

C. Transfer Learning

Transfer learning is a technique where a pre-trained model is used as a starting point for a new task. The idea is to leverage the knowledge the model gained from a previous task to improve performance on a new task. The pre-trained model, often a deep neural network trained on a large dataset like ImageNet, is modified for the new task by adding or replacing layers. Only the weights of the new layers are trained using the new dataset. Transfer learning is useful when the new dataset is too small to train a model from scratch. The new dataset is different but has some similarities to the original dataset. It has proven effective for tasks like image classification, object detection, and natural language processing, as it reduces the time and resources needed for training while enhancing performance on new tasks.

D. Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning where a computer program learns by taking actions in a given environment as also demonstrated in the Figure 2. It is similar to playing a game where you try different moves to get the highest score [16]. In RL, the program explores actions, learns from the results, and aims to make smart decisions to maximize its overall success.

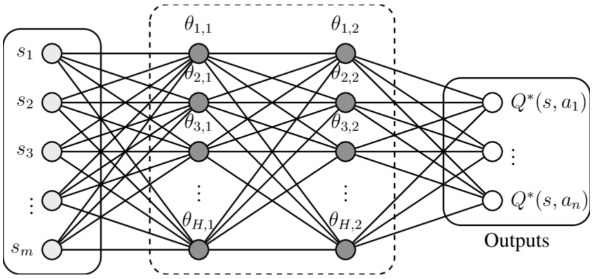


Fig. 3. Structure of the neural network used for the Deep Q-learning Network [18]

In a RL framework an agent has to interact with the given environment. The environment contains a set of states where a state at time t can be represented as S_t as also shown in the Figure 2. The agent has to choose an action at time t, represented as A_t out of set of actions for a state S_t and for each action, the environment returns a reward to the agent which helps an agent to learn and act optimally in an environment. An RL Agent is the program or "player" that makes decisions in the virtual world. Environment is the virtual world where the agent takes actions. Each environment contains different situations or states and the choices the agent can make in those situations are known as actions. The environment returns a reward against each action taken in a state.

The agent starts by randomly trying actions and learning from the outcomes. At first, it doesn't know much about the environment, but with each attempt, it figures out what works and what doesn't [17]. Over time, it gets better at choosing the best actions in different situations. The ultimate goal is to have a strategy (policy) for making the best decisions in any state of the environment. For example, consider the Multi-Arm Bandit (MAB) problem where we have many slot machines. The agent's goal is to learn which machine gives the best reward. It's a simplified RL scenario with only one state, and the agent balances between trying new machines (exploration) and sticking to the best one (exploitation). An advanced form of the RL is the use of NN to accomodate a larger state space environment. One such widely used algorithm is Deep Q-Network (DQN) as also demonstrated in the Figure 3.

III. AI FOR CELL MECHANICS

In this section, we will present some recent and interesting applications of AI in cell mechanics.

Authors in [19] develops interpretable machine learning models of cell mechanics from protein images and neural networks (NN) are used to predict traction forces from a single focal adhesion protein field. NN generalize to unseen biochemical perturbations, cell types and cells. Agnostic as well as physics-constrained methods learn interpretable rules for prediction. A DCell is developed in [20] which is a visible NN embedded in the hierarchical structure of 2,526 subsystems that encompasses an eukaryotic cell. DCell after training on millions of genotypes, can simulates cellular growth with good accuracy. This framework offers a foundation to decode the drug resistance, genetics of disease, and synthetic life.

Similarly, Deep RL (DRL) are exploited in [21] to infer collective cell behaviours and cell–cell interactions in tissue morphogenesis from 3D time-lapse images. Authors use hierarchical DRL to examine cell migrations from the images with an ubiquitous nuclear label. The hierarchical DRL method HDRL reveals a modular, multiphase organization of cell movement to Caenorhabditis elegans embryogenesis. A hybrid RL model is employed in [22] to guide process control efficiently. A probabilistic knowledge graph framework is created characterizing the science and risk-based understanding of quantifying inherent stochasticity and biomanufacturing process mechanisms. The proposed model can assist in learning from heterogeneous process data. Authors also used computational sampling technique to produce posterior samples quantifying model uncertainty while Bayesian RL is utilized for model uncertainty and inherent stochasticity for dynamic decision making.

A DRL method is introduced in [23] to study cell movement in the embryonic development of C.elegans. This agent based modeling mechanism captures the complexity of cell movement patterns in the embryo and addresses the local optimization issues. The model is tested with the rearrangement of the superficial left-right asymmetry and anterior movement of the Cpaaa cell via intercalation. A generic framework based on RL and convolutional NN is introduced in [24] to study navigation rules during cell migration. The proposed system uses a flexible model-free method that directly accepts raw images to the sensory input. The framework manages simulation scenarios involving cell division during embryogenesis.

Inverse RL is applied in [25] to analyze cellular heterogeneity. Authors showed the uses of inverse RL to datasets containing cellular states and actions for inferring how each cell selects actions in heterogeneous states. Authors have also discussed the applications of inverse RL to three cell biology problems. Moreover, the diverse uses of machine learning methods in systems biology are discussed in [26]. Authors also explored the combination of machine learning and mechanistic modeling for systems biology.

Supervised machine learning and generative NN are combined in [27] for patient-derived melanoma xenografts classification as “inefficient” or “efficient” metastatic. Moreover, predictions are validated regarding melanoma cell lines with unknown metastatic efficiency in mouse xenografts, and then the network is used to produce in silico cell images that improve the critical predictive cell features. The proposed work demonstrates how the application of AI can assist in the identification of cellular attributes that are predictive of integrated cell functions and complex phenotypes but are too abstruse for a human expert to identify. A computational framework using machine learning is developed in [28] for individual cell's deformability investigation. This system can reproduce a physical microfluidic. The datasets for the training and testing are produced with high-fidelity fluid-structure interaction simulations.

Similarly, in another work [29], machine learning is applied for extraction of the cellular force distributions from the microscope images. The authors divided the process into three stages. Initially, cells are cultured on a special substrate to measure the cellular traction force and the corresponding substrate wrinkles simultaneously. Then, the extraction of wrinkle positions from the microscope images is performed. At the end, machine learning model is trained with generative adversarial network by using input and traction field images. An analysis pipeline using deep learning is considered in [30] to perform classification of cell morphometric phenotypes from multi-channel fluorescence micrographs. Authors illustrated classification of definite morphological signatures observed when epithelial or fibroblasts cells are available with distinct extracellular matrices using residual NN with squeeze-and-excite blocks.

The intersection between cellular image analysis and deep learning are investigated in [31]. The authors studied about augmented microscopy, object tracking, image segmentation, and image classification. Taxol and cytochalasin D are known chemicals for interaction with the cytoskeleton and affect the cell motility and morphology. Therefore, quantitative measurements of the influence these two chemicals on human neuroblastoma are presented in [32]. The authors have designed customized deep learning and encoder-decoder based cell detection and tracking mechanisms.

The authors of [33] analyzed signal processing techniques such as matched filtering and spectral filtering of the video signal for detection of cellular micromotion. Moreover, they also considered 1D and 3D convolutional NN acting on pixel-wise time-domain data. An interesting and useful software package is proposed in [34] for cell tracking. The software package contains a manual tracking function, a pattern matching based sequential search-type tracking function, and a conventional tracking function composed of link processing and recognition processing. A deep learning based recognition function is also part of the package which is useful for several targets including stain-free images.

A foundation work has been done in [35] to promote genetic research and cellular biology. The authors constructed a novel model called scGPT for single-cell biology. The model is based on a generative pretrained transformer using a repository of more than 33 million cells. The scGPT efficiently distills critical biological insights concerning cells and genes. The model is useful but it can be optimized to obtain better performance across different downstream applications with further use of transfer learning. For example, transfer learning is used in [36] for data denoising in single-cell transcriptomics. Single cell RNA sequencing data are sparse and noisy. Therefore, authors used transfer learning across datasets to improve the quality of the data. A Bayesian model is coupled with a deep autoencoder to extract transferable gene-gene relationships across data from labs.

A data-integration infrastructure consists of deep generative models and adversarial training is proposed in [37] for both supervised and unsupervised integration of multiple batches. The authors employed six real bench-marking datasets to show that the proposed framework can address critical challenges including conservation of development trajectory, large number of batches, nested batch-effects, and skewed cell type distribution. Moreover, 1 million cells dataset are utilized to illustrate that the given framework can perform atlas-level cross-species. Similarly, authors in [38], consider the generalizability of another AI tool such as natural language processing in single-cell genomics. The authors found the necessity for careful consideration of data distribution and presented a subsampling method to overcome the effect of an imbalanced distribution.

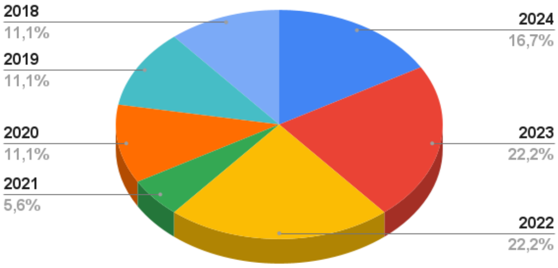


Fig. 4. Distribution of reported studies: Publication year

A symbolic data modeling approach “Evolutionary Polynomial Regression,” is adopted in [39]. The technique integrates genetic programming paradigm with regression capabilities to derive explicit analytical formulas. The authors have presented the main benefits of our multiscale numerical approach using spider silk case. Similarly, the authors in [40] suggest that integration of multiscale modeling and ML can offer new insights into disease mechanisms, assist in identification new targets and treatment plans, and help in decision making to benefit of human health.

In an interesting work of [41], the authors proposed a reduced order method for active force generation in cardiomyocytes. The model is designed by ML of a physics-based high-fidelity model. The authors trained an ANN within a grey-box technique. Then they validate the model under various pathological and physiological cell conditions.

IV. DISCUSSION

This section highlights the key points by discussing challenges and future directions. We have also reported in the Figure 4, a distribution of the studies that we analyzed in Section III according to their publication year

We found after an in-depth study that the AI technologies, particularly Reinforcement Learning (RL) techniques can be used for the development of optimal decision support agents able to assist in the definition of models’ features, such as the topology, and the prediction of some dynamics in single cell and cell aggregates (cell adhesion, growth, and remodeling). We observed that RL algorithms are much more focused on goal-oriented learning from interaction with the environment (system to control) than other machine learning approaches. Next we outline the list of our key findings:

- Deep RL (DRL) can be used to infer cell–cell interactions and collective cell behaviors in tissue morphogenesis from three-dimensional (3D) time-lapse images. In particular, Hierarchical DRL, which is known for multiscale learning and data efficiency, can be applied to examine cell migrations based on images with a ubiquitous nuclear label and simple rules formulated from empirical statistics of the images.
- Hybrid model-based RL can be employed for cell therapy manufacturing process control. Specially, hybrid model-based Bayesian RL approaches, accounting for both inherent stochasticity and model uncertainty are useful to guide optimal, robust, and interpretable dynamic decision making.
- Deep learning can be used to model the hierarchical structure and function of a cell.
- Cell movement in the early phase of Caenorhabditis elegans development is regulated by a highly complex process in which a set of rules and connections are formulated at distinct scales. We found that the DRL method can be utilized within an agent-based modeling system to characterize cell movement in the embryonic development of C.elegans.
- Cell migration modeling is a longstanding biological challenge, which is regulated by a highly complex set of regulatory mechanisms at multiple scales in a developmental system. We can use RL and convolutional neural networks to better study navigation rules and mechanisms during cell migration.
- Inverse RL can be integrated into data-driven mechanistic computational models. Inverse RL can be applied to datasets to infer how individual cells choose different actions based on heterogeneous states. Inverse RL techniques also have potential applications in three cell biology problems.
- We observed that including physical information in machine learning approaches can be useful to optimize the training phase and improve the output and prediction power of the methods. This is due to fact that sometimes established RL methods may have limitations in a scenario like: i) definition of the goal that in this case is a set of goals (i.e., structural equilibrium, local stiffness of structures, etc), ii) incorporation of initial knowledge such as a basis topology derived from biological images, iii) dimension of the state-space, and iv) interplay in cell aggregates.
- Deep RL approaches have application in studying cell–cell interactions and collective cell behaviors. Similarly, deep RL method can be utilized within an agent-based modeling system to characterize cell movement in the embryonic development of C.elegans.
- Bayesian RL techniques which account for both inherent stochasticity and model uncertainty are useful to guide optimal, robust, and interpretable dynamic decision making. Similarly, Deep learning and neural networks can be used to model the hierarchical structure and function of a cell.
- Moreover, RL and convolutional neural networks are suitable to study navigation rules and mechanisms during cell migration. Lastly, Inverse RL can be integrated into data-driven mechanistic computational models. Inverse RL can be applied to datasets to infer how individual cells choose different actions based on heterogeneous states.

V. CONCLUSION

This review has presented an overview of the use of artificial intelligence and machine learning for various aspects of cell mechanics. We also analyzed applying physical constraints on the AI and ML models in order to obtain better convergences during the training and validation phases. We observed that the AI technologies such as reinforcement learning techniques can be used for the development of optimal decision support agents able to assist in the definition of models' features, such as the topology, and the prediction of some dynamics in single cell and cell aggregates (cell adhesion, growth, and remodeling). We noted that RL algorithms are much more focused on goal-oriented learning from interaction with the environment (system to control) than other machine learning approaches.

ACKNOWLEDGMENT

This research has been partially supported by the Piano Nazionale di Ripresa e resilienza (PNRR) - Missione 4 "Istruzione e Ricerca" - Componente C2 - Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" - Grant number P2022MXCJ2, Settore PE1 – LINEA B – Sud, CUP F53D23010080001.

REFERENCES

[1] A. F. Pegoraro, P. Janmey, and D. A. Weitz, "Mechanical properties of the cytoskeleton and cells," *Cold Spring Harbor perspectives in biology*, vol. 9, no. 11, p. a022038, 2017.

[2] T. Svitkina, "The actin cytoskeleton and actin-based motility," *Cold Spring Harbor perspectives in biology*, vol. 10, no. 1, p. a018267, 2018.

[3] R. Phillips, J. Kondev, J. Theriot, and H. Garcia, *Physical biology of the cell*. Garland Science, 2012.

[4] C. Battle, C. P. Broedersz, N. Fakhri, V. F. Geyer, J. Howard, C. F. Schmidt, and F. C. MacKintosh, "Broken detailed balance at mesoscopic scales in active biological systems," *Science*, vol. 352, no. 6285, pp. 604–607, 2016.

[5] P. Romani, L. Valcarcel-Jimenez, C. Frezza, and S. Dupont, "Crosstalk between mechanotransduction and metabolism," *Nature Reviews Molecular Cell Biology*, vol. 22, no. 1, pp. 22–38, 2021.

[6] M. Naeem, S. T. H. Rizvi, and A. Coronato, "A gentle introduction to reinforcement learning and its application in different fields," *IEEE access*, vol. 8, pp. 209 320–209 344, 2020.

[7] M. Naeem, S. Bashir, Z. Ullah, and A. A. Syed, "A near optimal scheduling algorithm for efficient radio resource management in multi-user mimo systems," *Wireless Personal Communications*, vol. 106, no. 3, pp. 1411–1427, 2019.

[8] M. Naeem, A. Coronato, Z. Ullah, S. Bashir, and G. Paragliola, "Optimal user scheduling in multi antenna system using multi agent reinforcement learning," *Sensors*, vol. 22, no. 21, p. 8278, 2022.

[9] S. I. H. Shah, A. Coronato, M. Naeem, and G. De Pietro, "Learning and assessing optimal dynamic treatment regimes through cooperative imitation learning," *IEEE Access*, vol. 10, pp. 78 148–78 158, 2022.

[10] S. I. H. Shah, M. Naeem, G. Paragliola, A. Coronato, and M. Pechenizkiy, "An ai-empowered infrastructure for risk prevention during medical examination," *Expert Systems with Applications*, vol. 225, p. 120048, 2023.

[11] U. b. Khalid, M. Naeem, F. Stasolla, M. H. Syed, M. Abbas, and A. Coronato, "Impact of ai-powered solutions in rehabilitation process: Recent improvements and future trends," *International Journal of General Medicine*, pp. 943–969, 2024.

[12] M. Cinque, A. Coronato, and A. Testa, "Dependable services for mobile health monitoring systems," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 4, no. 1, pp. 1–15, 2012.

[13] C. J. Soelistyo, G. Vallardi, G. Charras, and A. R. Lowe, "Learning biophysical determinants of cell fate with deep neural networks," *Nature Machine Intelligence*, vol. 4, no. 7, pp. 636–644, 2022.

[14] M. Cinque, A. Coronato, and A. Testa, "A failure modes and effects analysis of mobile health monitoring systems," in *Innovations and advances in computer, information, systems sciences, and engineering*. Springer, 2012, pp. 569–582.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[16] M. Fiorino, M. Naeem, M. Ciampi, and A. Coronato, "Defining a metric-driven approach for learning hazardous situations," *Technologies*, vol. 12, no. 7, p. 103, 2024.

[17] M. Jamal, Z. Ullah, M. Naeem, M. Abbas, and A. Coronato, "A hybrid multi-agent reinforcement learning approach for spectrum sharing in vehicular networks," *Future Internet*, vol. 16, no. 5, p. 152, 2024.

[18] F. B. Mismar, J. Choi, and B. L. Evans, "A framework for automated cellular network tuning with reinforcement learning," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7152–7167, 2019.

[19] M. S. Schmitt, J. Colen, S. Sala, J. Devany, S. Seetharaman, A. Caillier, M. L. Gardel, P. W. Oakes, and V. Vitelli, "Machine learning interpretable models of cell mechanics from protein images," *Cell*, vol. 187, no. 2, pp. 481–494, 2024.

[20] J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker, "Using deep learning to model the hierarchical structure and function of a cell," *Nature methods*, vol. 15, no. 4, pp. 290–298, 2018.

[21] Z. Wang, Y. Xu, D. Wang, J. Yang, and Z. Bao, "Hierarchical deep reinforcement learning reveals a modular mechanism of cell movement," *Nature machine intelligence*, vol. 4, no. 1, pp. 73–83, 2022.

[22] H. Zheng, W. Xie, K. Wang, and Z. Li, "Opportunities of hybrid model-based reinforcement learning for cell therapy manufacturing process control," *arXiv preprint arXiv:2201.03116*, 2022.

[23] Z. Wang, D. Wang, C. Li, Y. Xu, H. Li, and Z. Bao, "Deep reinforcement learning of cell movement in the early stage of c. elegans embryogenesis," *Bioinformatics*, vol. 34, no. 18, pp. 3169–3177, 2018.

[24] ——, "Modeling cell migration with convolutional neural network and deep reinforcement learning," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), Tech. Rep., 2019.

[25] P. C. Kinnunen, K. K. Ho, S. Srivastava, C. Huang, W. Shen, K. Garikipati, G. D. Luker, N. Banovic, X. Huan, J. J. Linderman *et al.*, "Integrating inverse reinforcement learning into data-driven mechanistic computational models: a novel paradigm to decode cancer cell heterogeneity," *Frontiers in Systems Biology*, vol. 4, p. 1333760, 2024.

[26] A. Procopio, G. Cesarelli, L. Donisi, A. Merola, F. Amato, and C. Cosentino, "Combined mechanistic modeling and machine-learning approaches in systems biology–a systematic literature review," *Computer methods and programs in biomedicine*, p. 107681, 2023.

[27] A. Zaritsky, A. Jamieson, E. Welf, A. Nevarez, J. Cillay, U. Eskiocak, B. Cantarel, and G. Danuser, "Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma, cell systems 12 (2021)."

[28] D. Nguyen, L. Tao, H. Ye, and Y. Li, "Machine learning-based prediction for single-cell mechanics," *Mechanics of Materials*, vol. 180, p. 104631, 2023.

[29] H. Li, D. Matsunaga, T. S. Matsui, H. Aosaki, G. Kinoshita, K. Inoue, A. Doostmohammadi, and S. Deguchi, "Wrinkle force microscopy: a machine learning based approach to predict cell mechanics from images," *Communications biology*, vol. 5, no. 1, p. 361, 2022.

[30] K. S. Wong, X. Zhong, C. S. L. Low, and P. Kanchanawong, "Self-supervised classification of subcellular morphometric phenotypes reveals extracellular matrix-specific morphological responses," *Scientific Reports*, vol. 12, no. 1, p. 15329, 2022.

[31] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nature methods*, vol. 16, no. 12, pp. 1233–1246, 2019.

[32] S. Baar, M. Kuragano, K. Tokuraku, and S. Watanabe, "Towards a comprehensive approach for characterizing cell activity in bright-field microscopic images," *Scientific Reports*, vol. 12, no. 1, p. 16884, 2022.

[33] S. Rinner, A. Trentino, H. Url, F. Burger, J. von Lautz, B. Wolfrum, and F. Reinhard, "Detection of cellular micromotion by advanced signal processing," *Scientific Reports*, vol. 10, no. 1, p. 20078, 2020.

[34] H. Aragaki, K. Ogoh, Y. Kondo, and K. Aoki, "Lim tracker: a software package for cell tracking and analysis with advanced interactivity," *Scientific Reports*, vol. 12, no. 1, p. 2702, 2022.

[35] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, "scgpt: toward building a foundation model for single-cell multi-omics using generative ai," *Nature Methods*, pp. 1–11, 2024.

[36] J. Wang, D. Agarwal, M. Huang, G. Hu, Z. Zhou, C. Ye, and N. R. Zhang, "Data denoising with transfer learning in single-cell transcriptomics," *Nature methods*, vol. 16, no. 9, pp. 875–878, 2019.

[37] A. Shree, M. K. Pavan, and H. Zafar, "scdreamer for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier," *Nature Communications*, vol. 14, no. 1, p. 7781, 2023.

[38] S. A. Khan, A. Maillo, V. Lagani, R. Lehmann, N. A. Kiani, D. Gomez-Cabrero, and J. Tegner, "Reusability report: Learning the transcriptional grammar in single-cell rna-sequencing data using transformers," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1437–1446, 2023.

[39] V. Fazio, N. M. Pugno, O. Giustolisi, and G. Puglisi, "Physically based machine learning for hierarchical materials," *Cell Reports Physical Science*, vol. 5, no. 2, 2024.

[40] M. Alber, A. Buganza Tepole, W. R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W. W. Lytton, P. Perdikaris, L. Petzold *et al.*, "Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences," *NPJ digital medicine*, vol. 2, no. 1, p. 115, 2019.

[41] F. Regazzoni, L. Dedè, and A. Quarteroni, "Machine learning of multiscale active force generation models for the efficient simulation of cardiac electromechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 370, p. 113268, 2020.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 117–124 DOI: 10.15439/2024F5622 ISSN 2300-5963 ACSIS, Vol. 41

Analyzing the Privacy of a Healthcare RFID Authentication Protocol

Ștefana Gheorghiță
0009-0004-2925-3666
Faculty of Computer Science
"Alexandru Ioan Cuza" University of Iași, Iași, Romania
Email: gheorghitastefana@gmail.com

Anca-Maria Nica
0000-0002-3808-572X
Department of Computer Science
"Alexandru Ioan Cuza" University of Iași, Iași, Romania
Email: contact@ancamarianica.ro

***Abstract*—With the growing use of RFID systems in IoT environments, it is crucial for these systems to be highly efficient, reducing costs while also maintaining functionality. As technology evolves, adversaries' capabilities also increase, highlighting the necessity to consider all potential vulnerabilities that could be exploited, especially in terms of security and privacy. One particular case requiring attention is the use of temporary variables, which can inadvertently provide valuable information to an adversary. This scenario will be exemplified and addressed through the case of an RFID mutual authentication scheme designed for the healthcare field.**

I. INTRODUCTION

ONE essential concept that emerged in the last few years is the Internet of Things (IoT), which defines a revolutionary way of interacting with technology.

As the demand for appropriate devices that can implement IoT increased, RFID systems emerged as a potential solution due to their versatility based on specific architectures. RFID tags range from simple passive ones with minimal computational complexity (but also low production costs) to more complex, smart tags designed to perform various functions.

These considerations led to a significant increase in the number of patents and research papers on this subject [13]. One of the most exploited subdomains is represented by authentication protocols. As the number of proposed protocols is constantly growing, the technology and requirements change and it can be hard to maintain a detailed insight into the security and privacy requirements of such schemes. This led to the existence of many schemes incapable of achieving satisfactory privacy and/or security for real-life use. Key reason for this is the lack of suitable adversarial models in the analysis of the proposed protocols.

Contribution: The purpose of this paper is to draw attention to the privacy issues associated with RFID schemes and emphasize the necessity for a suitable model in analyzing security and privacy. This paper will focus on the vulnerabilities emerged from the use of global temporary variables. As such, we analyze the scheme presented in [8] using the Vaudenay model, one of the most notable security and privacy models for RFID systems. The vulnerabilities identified through this analysis will be highlighted. Additionally, we will address these problems and propose enhancements to improve the privacy of the scheme.

Paper structure: The current paper is structured in four sections. Section 1 contains the introduction. Section 2 summarizes one of the widely accepted security and privacy RFID models, the Vaudenay model. Section 3 focuses on the protocol to be analyzed. The protocol will be briefly described, followed by an analysis of potential attacks on the scheme. Finally, Section 3 will discuss enhancements aimed at improving the privacy of the scheme. The final section concludes the paper.

II. THE VAUDENAY MODEL

As RFID systems tend to be used on a larger scale, there is a need to find a balance between reducing costs in the manufacturing and utilization of tags and adhering to security requirements. Since implementing strong cryptographic algorithms would be costly and impractical for tags used at a larger scale, the focus has shifted towards finding solutions that take into account the inevitable risks to which such a system is exposed [7].

The most important attributes for an RFID system to maintain are privacy and security, as highlighted in various papers [1], [4]. Different concerns regarding these properties arise due to constraints imposed on the computational power of the tags.

From this perspective, the security and privacy model used in the analysis is crucial for defining an RFID scheme. Many models have been proposed to offer generality and suitability for simulating practical risks. Among these, two widely accepted models are the one proposed by Vaudenay in [1] and its extension for mutual authentication resulting from collaboration with Paise in [2], as well as the model proposed by Hermans et al. in [3] and its extension for multiple readers, included in [4]. The model proposed by Hermans et al. will be referred to as the HPVP model, based on the initials of its authors.

This section will conduct a summary of the Vaudenay model.

A. RFID System

An RFID system is defined by the existence of two main components:

- $\mathcal{T}$ - the set of tags (with the role of transponders), devices characterized by limited memory and computational

©2024, PTI

power; each tag has a corresponding unique identifier (ID) stored in the database of the reader; every tag stores its own state S, which may or may not include the associated ID;

- $\mathcal{R}$ - the set of readers, which have the role of transceivers (devices that both transmit and receive signals); in most cases, the focus is on the situation where there is only one reader;

As described in [1], we should also consider the three main algorithms which are necessary for an RFID system:

1) $SetupReader(1^k)$: using k (the security parameter), the pair (public key (K_P), secret key (K_S)) of the reader is generated;
2) $SetupTag_{K_P}(ID)$: for the tag with the identifier ID, the initial state is provided and stored in the memory of the tag and the corresponding secret of the tag is generated and added alongside the ID in the database of the reader if the tag is legitimate;
3) $Protocol$: the protocol between the reader and the tag, which ends with an **Output** from the reader ($\perp$ if the tag is not legitimate and ID if the tag is legitimate);

For an RFID scheme, the **Output** should be correct with overwhelming probability, meaning that for a legitimate tag the output is ID, otherwise being $\perp$.

B. Adversarial Model

The adversary is defined as an algorithm which can interact with the RFID system on the basis of the following oracles:

$CreateTag^b(ID)$**:** Based on the value of b, a legitimate ($b = 1$) or a illegitimate ($b = 0$) tag is created. This oracle calls $SetupTag$ and in the case of a legitimate tag, it is added in the central database. If the value of b is omitted, it means $b = 1$.

$Launch() \to \pi$**:** This oracle launches a new protocol session denoted by π.

$DrawTag(d) \to (vtag_1, b_1, ..., vtag_n, b_n)$**:** This oracle takes a probability distribution d as input and, based on that, creates n virtual tags, each of them being associated to a bit b, corresponding to the legitimacy of the tag. Already drawn tags cannot be drawn again (the oracle returns $\perp$). $\perp$ is also returned if the tags given as parameters do not exist. Together with this oracle, a hidden table $\mathcal{T}$ is created to store the real identity of the virtual tags: $\mathcal{T}(vtag_i) = ID_i$, where ID_i is the identifier of the real tag referenced by $vtag_i$.

$Free(vtag)$**:** This oracles frees the virtual tag $vtag$.

$SendTag(m, vtag) \to m^{'}$**:** This oracle is used to send a message m to the tag denoted by the virtual identity $vtag$ and to get its response in the form of $m^{'}$.

$SendReader(m, \pi) \to m^{'}$**:** In the protocol session π, a message m is sent to the reader and the response of it is returned.

$Execute(vtag) \to (\pi, transcript)$**:** This oracle simulates an entire protocol session by initially calling $Launch$ and then using subsequent calls of $SendReader$ and $SendTag$. It returns the pair containing the session π and the list of the successive protocol steps.

$Result(\pi) \to a$**:** This oracle is used for getting the result of the protocol session π regarding the authentication status of the tag. When the session is complete, if the tag is not considered legitimate (the output is $\perp$), it returns 0, otherwise it returns 1.

$Corrupt(vtag) \to \mathcal{S}$**:** This oracle is used for corrupting the tag referred by the virtual identity $vtag$. The returned value is the current state of the tag.

C. Adversary Classes

The Vaudenay model defines the following adversary classes:

- **weak adversary**: cannot access the $Corrupt$ oracle;
- **forward adversary**: if the adversary has used the $Corrupt$ oracle, then the only accessible oracle after that is $Corrupt$;
- **destructive adversary**: after a tag is corrupted, the tag is considered destroyed (the adversary cannot interact with it anymore);
- **strong adversary**: there are no restrictions imposed on the use of the oracles;

Additionally, **narrow adversaries** represent adversaries that do not have access to the $Result$ oracle. This notion can be combined with the ones mentioned above to construct the following classes: **narrow-weak**, **narrow-forward**, **narrow-destructive**, **narrow-strong**.

D. The Three Essential Properties

The paper [1] takes into account three cryptographic properties to be considered when analyzing an RFID scheme:

1) **Correctness**:
   A scheme ensures correctness if it outputs the correct result with overwhelming probability: if the tag with the identifier ID is legitimate, the scheme outputs ID, otherwise it outputs $\perp$.
2) **Security**:
   Definition 4 from [1] states that the security of a scheme is respected if a strong adversary has a negligible probability in determining a reader to identify an uncorrupted legitimate tag with which it has not engaged in any matching conversation. This is considered only in the case of tag authentication. When mutual authentication occurs, a tag should also only have a negligible probability of authenticating a legitimate reader with which it didn't have a corresponding conversation [2].

   Simple security refers to the same notion, but by restricting it to the situation in which the adversary cannot query the $Result$ oracle. Furthermore, the reader does not use the database for the messages it transmits and also there appear two new notions: a predicate R and a sampling algorithm S, which would simulate the $Result$ oracle. In addition, it is possible that the entry of the tag in the database to be updated by a different algorithm. For simple security to be achieved it is mandatory to have simple tag authentication and simple reader authentication. Simple security implies security.

3) **Privacy**:
The notion of privacy is presented in **Definition 6** and **Definition 7** from [1] and it is based on the notion of *a blinder for an adversary*.
A blinder $\mathcal{B}$ for an adversary $\mathcal{A}$ represents a PPT algorithm which has access to the messages to which $\mathcal{A}$ also has access and simulates the following oracles for $\mathcal{A}$: *Launch*, *SendReader*, *SendTag*, *Result*. $\mathcal{A}^{\mathcal{B}}$ denotes a blinded adversary, adversary that does not use the oracles named above, rather those oracles are simulated by the blinder.
A trivial adversary is an adversary for which exists $\mathcal{B}$ a blinder with the property that

$$\left|\Pr[\mathcal{A}\,\text{wins}] - \Pr[\mathcal{A}^{\mathcal{B}}\,\text{wins}]\right|$$

is negligible.
P-privacy: If every adversary in class P is trivial, then a RFID scheme is considered P-private.

III. ANALYSIS OF THE SHARIQ ET AL. PROTOCOL

In their 2021 publication, Shariq et al. introduced the *SR^2AP-DSC* mutual authentication protocol, specifically designed for use in the healthcare sector [8]. The authors asserted that this protocol offers robust privacy and security features. We selected this protocol for our analysis because it clearly illustrates the vulnerabilities associated with global temporary variables. Given its application in the medical field, where privacy is critically important, it is crucial to scrutinize any potential weaknesses. Highlighting these vulnerabilities is essential to ensure that the protocol meets the high privacy standards required in healthcare environments.

A. The Shariq et al. Protocol

1) Security Analysis: As the protocol is designed to be used in the healthcare field, it is necessary to consider that private patient data will be stored, thus requiring the modeling of a suitable adversary. With these considerations in mind, the Dolev-Yao model has been used as the baseline model for the adversary, as presented in the work [11]. In this model, the adversary is considered a legitimate user of the system, capable of being both initiator and receiver in message exchanges with a user A, while also having access to all messages passing through the network.

Considering this, the authors resorted to the security and privacy model based on indistinguishability proposed by Ouafi and Phan in paper [9]. This model is based on the Juels-Weis model [10], with certain differences, including those related to the restrictions imposed on the adversary.

In the model, communication will take place between two parties ($\mathcal{T}$ - the tag and $\mathcal{R}$ - the reader) during a session that will end with the *Accept* result for each party if it considers that the session has been conducted correctly, between the appropriate entities. We will denote by $\mathcal{S}$ the unit consisting of the server and the reader.

Two communicating entities are considered to be partners if and only if both parties provide the *Accept* response to each other, marking the completion of the session. (**Definition 1** [9])

One of the participating parties in the protocol is considered *fresh* at the end of the session if and only if: it has provided the *Accept* output (even if it has or does not have a partner) and neither the given instance nor its partner (if any) has received a *Corrupt* query. (**Definition 2** [9])
The adversary respects the described model and has access to the following oracles:

$\mathit{Execute}(\mathcal{S}, \mathcal{T}, i)$**:** Defines an eavesdropping attack, in which the adversary $\mathcal{A}$ has access to the messages exchanged in the session i between two honest parties $\mathcal{S}$ and $\mathcal{T}$, also having access to the shared secrets of the two parties.

$\mathit{Send}(U, V, m, i)$**:** Defines an active attack which takes place in the session i and it describes the situation when $\mathcal{A}$ impersonates entity U and sends the message m to the entity V (U and V are part of different categories: one is a reader and the other one is a tag).

$\mathit{Query}(\mathcal{T}, m_1, m_2)$**:** The adversary queries the tag $\mathcal{T}$ with the message m_1 and receives the message m_2.

$\mathit{Block}(\mathcal{A})$**:** It defines a *Denial of Service* attack with the purpose of stopping the protocol execution or desynchronizing the two parties.

$\mathit{Corrupt}(\mathcal{T}, K)$**:** Provides access to the secret key K' stored in the memory of the tag to the adversary, thereby allowing the adversary the opportunity to replace that key with another one, denoted as K. This oracle is used for modelling the property of *forward privacy*.

$\mathit{Test}(\mathcal{T}_0, \mathcal{T}_1, i)$**:** This oracle does not refer to any of the adversary's abilities or to any real event, but defines the notion of *UPriv* (untraceable privacy), a notion based on the property of indistinguishability. If the entity has responded with *Accept* and is offered a *Test* query, a random bit b from $\{0, 1\}$ will be chosen and the adversary will get $\mathcal{T}_b$, with the aim of discovering if they got $\mathcal{T}_0$ or $\mathcal{T}_1$. For the UPriv property to be relevant, the *Test* session should be fresh.

Definition 3 [9] describes the notion of untraceable privacy more in depth:

UPriv refers to the game $\mathcal{G}$ between an adversary $\mathcal{A}$ and different instances of tags or readers. The game contains 3 phases:

- **Learning phase**: the adversary can interact with $\mathcal{S}$ and the two randomly chosen tags ($\mathcal{T}_0$ an $\mathcal{T}_1$) and can use *Execute*, *Send* and *Corrupt* queries.
- **Challenge phase**: the adversary uses $\mathit{Test}(\mathcal{T}_0, \mathcal{T}_1, i)$ and sends it to the challenger who will choose the value $b \in \{0, 1\}$, corresponding to one of the tags. $\mathcal{A}$ can query *Execute* and *Send* oracles (also *Corrupt*, but without violating the freshness property) to help him guess which tag they received.
- **Guessing phase**: the adversary will present b', the index of the tag they think they got.

$\mathcal{A}$'s success in winning the game (and thus violating the *UPriv* property) is determined by the advantage the adversary has in distinguishing between the two tags (compared to a random choice between the two) and thus guessing the number b. According to the above, this advantage can be expressed mathematically as: $Adv_{\mathcal{A}}^{UPriv}(K) = \left|\Pr[b' = b] - \frac{1}{2}\right|$, where k is the security parameter.

2) The Authentication Protocol: The proposed protocol is intended to be used in implementing an intelligent healthcare system, in which tags are attached to both patients and medical equipment, and the reader will transmit the information obtained from the tags to the central server (trusted authority) which will store the received data and perform the necessary operations.

In the following parts, we will consider the reader and server as a unit (reader-server unit), which we will refer to more generally as the server and it will be denoted as $\mathcal{S}$.

The protocol consists of two phases: in the first phase, the system parameters are set, and in the second phase, authentication is performed.

In the setup phase the parameters of the system are initialized. The parameters are:

1) two large prime numbers: p (1024 bits) and q (160 bits);
2) the generator g with $g \in (1, p-1)$;
3) the shared secret key a with $a \in (0, q)$;
4) the public key v ($v = g^{-a} \bmod p$);
5) a_i - random number stored in encrypted form by the server;
6) ID_i (160 bits) - the ith tag identifier, stored in the memory of the tag and in the database of the server;

The authentication phase consists of four steps:

1) **Step 1**: $\mathcal{S} \rightarrow tag_i$: $\{C_1\}$
   In the first step, the server randomly chooses a non-zero integer C_1, which it sends to the tag with which communication is established. This tag will be denoted as tag_i.
2) **Step 2**: $tag_i \rightarrow \mathcal{S}$: $\{x_1, x_2, Auth_i\}$
   This step consists of operations performed by the tag: first, it will randomly choose two non-zero integers r_1 and r_2, and then these will be used to calculate the values that will be transmitted to the server: x_1, x_2, $Auth_i$.
   The two values x_1 and x_2 are computed as follows:
   - $x_1 = g^{r_1} \bmod p$
   - $x_2 = (r_2 \cdot v^{-r_1}) \bmod p$

   Now we will calculate e and y:
   - $e = h(r_2 \,||\, x_2 \,||\, x_1)$
   - $y = (r_2 + a_i \cdot e) \bmod q$

   The last value to be computed is $Auth_i$:
   - $Auth_i = ID_i \oplus h(r_2, C_1, e, y)$

   The tag sends $\{x_1, x_2, Auth_i\}$ to the server.
3) **Step 3**: $\mathcal{S} \rightarrow tag_i$: $\{C_2\}$
   In the third step, the authentication of the tag is performed.
   The server will carry out the necessary operations to extract the tag identifier from the received information and will search the database for the obtained value.
   Thus, we will calculate the values:
   - $S_1 = {x_1}^a \bmod p$
   - $S_2 = S_1 \bmod p$
   - $r_2 = (x_2 \cdot {S_2}^{-1}) \bmod p$
   - $e' = h(r_2 \,||\, x_2 \,||\, x_1)$
   - $y' = (r_2 + a_i \cdot e') \bmod q$

   Using the previously computed values, ID_i is computed:
   - $ID_i = Auth_i \oplus h(r_2, C_1, e', y')$

   The resulted value will be then searched in the database and if it is found, the tag will be authenticated. Now the server computes C_2 and sends it to the tag:
   - $C_2 = h(ID_i, r_2, C_1, e', y', Auth_i)$
4) **Step 3**: Authentication of the reader/server
   The tag will compute the value C_2' using the known values of the variables:
   - $C_2' = h(ID_i, r_2, C_1, e, y, Auth_i)$

   If $C_2' = C_2$, the reader-server unit is authenticated.

Thus, through these four steps and the properties of the Schnorr Cryptosystem, mutual authentication between the server and the tag is achieved. In the **Fig. 1**, the steps of the protocol are summarized.

B. Failing to Achieve Privacy in the Vaudenay Model

The authors of the $SR^2AP\text{-}DSC$ protocol analysed its security and privacy properties using the Ouafi-Phan privacy model and proved that their protocol achieves good security and privacy. The Vaudenay model, for which a summary was provided in the second section, represents one of the most influential models used in the analysis of RFID protocols, being a stronger model than the Ouafi-Phan model in terms of the abilities of the adversary in tag corruption.

This subsection will make an analysis of the protocol according to the Vaudenay model, proving that the scheme does not assure the necessary privacy properties.

Before moving forward, it's important to take two notes on the scheme. Firstly, the protocol exhibits linear complexity when identifying tags, as each a_i is unique for every tag and session and its value is not sent to the reader/server. Consequently, the server must iterate through all values in the database. Secondly, the protocol lacks specification regarding the update mechanism for a_i and ID_i after each session. The manner in which this update occurs is essential to the protocol, as failure to address it adequately could result in desynchronization issues.

In the article [5], there are described five cases in which an RFID protocol cannot assure privacy in the Vaudenay and HPVP models. Based on this situations, five lessons are formulated.

One scenario in which a scheme may fall is the use of global temporary variables, which means variables which are assigned at a certain step and then used at another step. Regarding this situation, **Theorem 3.1** [5] asserts that a

Reader + Server $\{g, p, q, v, a\}$	**Message**	**Tag** $\{ID_i, p, q, v, a_i\}$
Choose $k \in \mathbb{Z}^*$, $C_1 = k$	$\xrightarrow{\{C_1\}}$	
	$\xleftarrow{\{x_1,\, x_2,\, Auth_i\}}$	Generate two integers r_1, r_2 $x_1 = g^{r_1} \bmod p$, $x_2 = (r_2 \cdot v^{-r_1}) \bmod p$ $e = h(r_2 \Vert x_2 \Vert x_1)$, $y = (r_2 + a_i \cdot e) \bmod q$ $Auth_i = ID_i \oplus h(r_2, C_1, e, y)$
$S_1 = {x_1}^a \bmod p$, $S_2 = S_1 \bmod p$ $r_2 = (x_2 \cdot {S_2}^{-1}) \bmod p$ $e^{'} = h(r_2 \Vert x_2 \Vert x_1)$, $y^{'} = (r_2 + a_i \cdot e^{'}) \bmod q$ $ID_i = Auth_i \oplus h(r_2, C_1, e^{'}, y^{'})$ $C_2 = h(ID_i, r_2, C_1, e^{'}, y^{'}, Auth_i)$	$\xrightarrow{\{C_2\}}$	
		$C_2^{'} = h(ID_i, r_2, C_1, e, y, Auth_i)$ If $C_2^{'} = C_2$, the reader is authenticated

Fig. 1. The Shariq et al. scheme.

scheme, where the computations conducted by the tag to determine reader authentication are not based on PUFs (Physically Unclonable Functions) and the scheme uses global temporary variables, then the protocol cannot simultaneously achieve both mutual authentication and narrow-forward privacy within the Vaudenay model. This limitation arises only when the adversary has the capability to access global temporary variables, thus leading to a scenario of temporary state disclosure.

The demonstration of the theorem offers a comprehensive insight into the potential attack that can be mounted under these circumstances. We will consider $\mathcal{A}$ to be a narrow-forward adversary. The adversary will play the following privacy game:

1) $CreateTag(ID)$;
2) $(vtag, 1) \leftarrow DrawTag(distribution)$;
3) $\pi \leftarrow Launch()$;
4) $\{C_1\} \leftarrow SendReader(\emptyset, \pi)$;
5) $\{x_1,\, x_2,\, Auth_i\} \leftarrow SendTag(\{C_1\}, vtag)$;
6) $\{C_2\} \leftarrow SendReader(\{x_1, x_2, Auth_i\}, \pi)$;
7) $S \leftarrow Corrupt(vtag)$;
8) The adversary obtains all the necessary values for computing $h(ID_i, r_2, C_1, e, y, Auth_i)$ so they can decide if the reader will be authenticated by the tag or not.
9) If the adversary considers the reader to be authenticated, they return 0; otherwise, they return 1.

By corrupting the tag and considering that the temporary state is also returned when corrupting the tag, the adversary will obtain the following values: p, q, ID_i, a_i, r_2 (it is stored because it is used in two different steps), C_1. Furthermore, e, y and $Auth_i$ will be also probably present, but there is also the possibility that the tag recomputes them (if e and y are not directly stored, it means that x_1 and x_2 are or that r_1 is stored and used in the future computations). In all of these cases, the adversary has all the necessary values to compute $C_2^{'} = h(ID_i, r_2, C_1, e, y, Auth_i)$. If $C_2^{'} = C_2$, it means that the reader will be authenticated by the tag.

For the scheme to achieve narrow-forward privacy, there must exist a blinder $\mathcal{B}$ for the adversary $\mathcal{A}$ for which $\left|\Pr[\mathcal{A}\,\text{wins}] - \Pr[\mathcal{A}^{\mathcal{B}}\,\text{wins}]\right|$ is negligible. From the logic above, results that $\mathcal{A}$ will always output 0 in the case of a legitimate reader. On the other hand, in the blinded privacy game, because the blinder simulates $Launch$, $SendTag$ and $SendReader$, there is a high chance that this simulated reader will not be authenticated, making it obvious for the adversary which is the blinded game and which is the real one. For the adversary to not differentiate between the two games, the blinder should simulate the reader in such a manner to make it hard for the adversary to distinguish between a real reader and a fake one. This assumption would break the security of the scheme.

This means, for the scheme to still ensure narrow-forward privacy, it is necessary for the reader to not be authenticated. If reader authentication is performed, the narrow-forward privacy is lost.

As it was stated before, it can also be assumed that $Auth_i$ will also be present in the memory of the tag (as its value is also needed in the fourth step). This would offer the adversary the possibility to compare the stored $Auth_i$ with the one sent through the communication channel to differentiate between the real and the blinded privacy game. The same logic can also be applied to C_1, as it too will be saved between the steps, and it can also be used in the case when x_1 and x_2 are stored, instead of e and y.

This analysis proved that the scheme cannot achieve narrow-forward privacy and mutual authentication at the same time, the problem being the use of global temporary variables, which can be obtained by the adversary.

C. Improving the Protocol

As it was stated before, the use of global temporary variables can represent an impediment in achieving narrow-forward privacy at the same time as mutual authentication in the particularly case of temporary state disclosure in Vaudenay model, case which is plausible in the context of real-life

Reader + Server $\{g, p, q, v, a\}$	**Message**	**Tag** $\{ID_i, p, q, v, a_i, P, s_{C_2}\}$
Choose $k \in \mathbb{Z}^*$, $C_1 = k$	$\xrightarrow{\{C_1\}}$	
	$\xleftarrow{\{x_1,\, x_2,\, Auth_i\}}$	Generate two integers r_1, r_2 $K_{C_2} = P(s_{C_2})$ $x_1 = g^{r_1} \bmod p$, $x_2 = (r_2 \cdot v^{-r_1}) \bmod p$ $e = h(r_2 \,\|\, x_2 \,\|\, x_1)$, $y = (r_2 + a_i \cdot e) \bmod q$ $Auth_i = ID_i \oplus h(r_2, C_1, e, y)$ $C_2^{'} = h(ID_i, r_2, C_1, e, y, Auth_i) \oplus K_{C_2}$ erase K_{C_2}, r_1, r_2, e, y, x_1, x_2, $Auth_i$
$S_1 = {x_1}^a \bmod p$, $S_2 = S_1 \bmod p$ $r_2 = (x_2 \cdot {S_2}^{-1}) \bmod p$ $e^{'} = h(r_2 \,\|\, x_2 \,\|\, x_1)$, $y^{'} = (r_2 + a_i \cdot e^{'}) \bmod q$ $ID_i = Auth_i \oplus h(r_2, C_1, e^{'}, y^{'})$ $C_2 = h(ID_i, r_2, C_1, e^{'}, y^{'}, Auth_i)$	$\xrightarrow{\{C_2\}}$	
		$K_{C_2} = P(s_{C_2})$ If $C_2^{'} \oplus K_{C_2} = C_2 \implies$ the reader is authenticated erase K_{C_2}

Fig. 2. A PUF-protected variant of the Shariq et al. scheme.

conditions. One way to overcome this problem is presented in [6].

The solution mentioned above is based on the use of *Physically Unclonable Functions* (PUFs) to protect the temporary variables. PUFs can be described as having two components: the input (challenge) and the output (response) generated for the given input, a CRP (challenge-response pair) being modelled. The response relies not only on the value of the challenge, but also on the physical attributes of the object. The most essential properties of PUFs are unpredictability, physical unclonability and tamper-evidence. One problem that may appear is the different response for the same challenge, meaning that the PUF is not reliable/robust. There exist known practical solutions to overcome this issue, but the supplementary overhead needs to also be taken into consideration [12]. Based on the above-mentioned characteristics and considering the multiple analyses conducted on their properties [16], [17], PUFs represent a suitable choice in the case of RFIDs.

In [6] it is proven that if a protocol achieves X-privacy and mutual authentication in the Vaudenay model without temporary state disclosure, it can be modified by adding PUF computations to also achieve the same properties in the Vaudenay model with temporary state disclosure.

For this purpose, ideal PUFs are used, being defined as a function $P : \{0,1\}^p \to \{0,1\}^k$ (p, k - polynomial sized values in the security parameter). This function needs to respect two conditions: it is computationally indistinguishable from random functions and also tampering with the object means that P is destroyed. Supplementary, it is also considered that, after being corrupt, the tag is destroyed (according to the tamper-evident nature) [6].

Based on this and on the vulnerabilities highlighted in the previous subsection, the scheme can be modified in the form that is presented in **Fig. 2**.

For proving that this variant is secure, **Definition 5.2** [6] can be referred to. This definition states the condition for a PUF-protected variant of a scheme to be secure for some class of adversaries. It is assumed that if the adversary cannot obtain the PUF-protected variable without corruption, then it can obtain it in the case of corruption with temporary state disclosure only with negligible probability.

In the case of Shariq et al. scheme, the $C_2^{'}$ variable is not sent through the channel, meaning that the adversary cannot access it. For computing this value, the adversary would need the value of r_2, which they cannot obtain without accessing the temporary state (the other variables can be found in the permanent state or can be computed knowing r_2). $C_2^{'}$ is protected using $C_2^{'} \oplus P(s_{C_2})$ and the other temporary variables are erased, meaning that by corrupting the tag, $C_2^{'} \oplus P(s_{C_2})$ is the only value that the adversary can obtain, value which cannot be used in gaining any advantage as the tag is also destroyed after corruption. Based on the assumed security of the PUF, it can be concluded that the scheme is secure against a narrow-forward adversary (for which the attack was constructed). **Theorem 6.1** [6] summarizes the relationship between the original scheme and the PUF-protected variant of the scheme, stating that if a scheme achieves mutual authentication and X-privacy in the Vaudenay model without temporary state disclosure, then any PUF-protected variant of the scheme will also achieve mutual authentication and X-privacy in the Vaudenay model with temporary state disclosure.

From **Theorem 6.1** [6] and using the fact that the PUFs are considered secure, the constructed PUF-protected variant of the Shariq et al. scheme achieves narrow-forward privacy in the Vaudenay model with temporary state disclosure. This happens because, by corrupting the tag, besides the variables

Reader + Server $\{g, p, q, v, a\}$	**Message**	**Tag** $\{ID_i, p, q, v, a_i, P, s_{C_2}, s_{a_i}\}$
Choose $k \in \mathbb{Z}^*$, $C_1 = k$	$\xrightarrow{\{C_1\}}$	
	$\xleftarrow{\{x_1,\, x_2,\, Auth_i\}}$	Generate two integers r_1, r_2 $K_{C_2} = P(s_{C_2})$ $K_{a_i} = P(s_{a_i})$ $x_1 = g^{r_1} \bmod p$, $x_2 = (r_2 \cdot v^{-r_1}) \bmod p$ $e = h(r_2 \Vert x_2 \Vert x_1)$, $y = (r_2 + (a_i \oplus K_{a_i}) \cdot e) \bmod q$ $Auth_i = ID_i \oplus h(r_2, C_1, e, y)$ $C_2^{'} = h(ID_i, r_2, C_1, e, y, Auth_i) \oplus K_{C_2}$ erase K_{C_2}, r_1, r_2, e, y, x_1, x_2, $Auth_i$, K_{a_i}
$S_1 = {x_1}^a \bmod p$, $S_2 = S_1 \bmod p$ $r_2 = (x_2 \cdot {S_2}^{-1}) \bmod p$ $e^{'} = h(r_2 \Vert x_2 \Vert x_1)$, $y^{'} = (r_2 + a_i \cdot e^{'}) \bmod q$ $ID_i = Auth_i \oplus h(r_2, C_1, e^{'}, y^{'})$ $C_2 = h(ID_i, r_2, C_1, e^{'}, y^{'}, Auth_i)$	$\xrightarrow{\{C_2\}}$	
		$K_{C_2} = P(s_{C_2})$ If $C_2^{'} \oplus K_{C_2} = C_2 \implies$ the reader is authenticated erase K_{C_2}

Fig. 3. An Improved PUF-protected variant of the Shariq et al. scheme.

present in the permanent memory, the adversary only obtains $C_2^{'} \oplus P(s_{C_2})$, from which they cannot deduce any useful information.

The scheme still does not achieve narrow-destructive privacy (without temporary state disclosure), as the following attack can be mounted:

1) $\mathit{CreateTag}(ID)$;
2) $(vtag, 1) \leftarrow \mathit{DrawTag}(distribution)$;
3) $\pi \leftarrow \mathit{Launch}()$;
4) $\{C_1\} \leftarrow \mathit{SendReader}(\emptyset, \pi)$;
5) $\{x_1, x_2, Auth_i\} \leftarrow \mathit{SendTag}(\{C_1\}, vtag)$;
6) $(ID_i, p, q, v, a_i) \leftarrow \mathit{Corrupt}(vtag)$;
7) $\{C_2\} \leftarrow \mathit{SendReader}(\{x_1^{'}, x_2^{'}, Auth_i^{'}\})$;
8) If the adversary considers the reader to be authenticated, they return 0; otherwise, they return 1.

The adversary simulates the first two steps of the protocol and then corrupts the tag to obtain its permanent state.

Then, the adversary generates two random numbers r_1 and r_2 and computes $x_1^{'}$, $x_2^{'}$ and $Auth_i^{'}$ to be sent to the reader. As the adversary has access to ID_i, p, q, v, a_i, the computed $x_1^{'}$, $x_2^{'}$ and $Auth_i^{'}$ are valid. After that, the adversary can compute $C_2^{'}$ using the values obtained by corrupting the tag and the generated ones. As the response from the adversary is valid, if the adversary plays the real game, $C_2^{'} = C_2$ with overwhelming probability (if the reader would be authenticated).

In this way, if the reader is authenticated, the adversary can distinguish between the real privacy game and the blinded one (in which C_2 would be wrong with overwhelming probability). Considering this, when playing the real privacy game, the adversary will output 0 with overwhelming probability (when the reader is authenticated), while the result will be 1 with overwhelming probability, in the case of the blinded privacy game.

As a consequence, the scheme cannot achieve mutual authentication and narrow-destructive privacy, remaining limited to narrow-forward privacy.

For the scheme to also achieve narrow-destructive privacy, supplementary protection regarding the values from the permanent memory (ID_i, p, q, v and a_i) should be implemented.

In **Fig. 3**, a_i is protected using PUFs, meaning that, when the tag is corrupted, $a_i \oplus P(s_{a_i})$ is obtained (this is the stored value). As the PUFs are considered secure, the adversary cannot guess the real value of a_i, meaning that they cannot compute y and, subsequently, $Auth_i$. This type of protection could have been applied to ID_i too, depending on the specific requirements.

To ensure forward privacy, the scheme must be proven to be secure (to use the result included in **Lemma 8** from [1]). As forward privacy can be achieved using PKC, the security of the scheme depends on the security of the utilized Schnorr signature. Regarding this aspect, Schnorr signatures have been proven secure against an adaptive chosen-message attack in the Random Oracle Model, based on the complexity of the Discrete Logarithm Problem [14], [15]. Despite the numerous studies on the security of the Schnorr signature scheme, its security in the standard model has not been proven.

IV. CONCLUSIONS

As stated earlier in the paper, the disclosure of temporary state is a critical consideration for the practical implementation of RFID systems.

In the Vaudenay model, it is not explicitly stated whether the adversary also gains access to the temporary variables when corrupting the tag, but to consider this scenario as possible, is an essential requirement given the continuous advance of technology and the increasing capabilities of adversaries.

The analysis above was conducted to highlight vulnerabilities that may arise when using global temporary variables. As exemplified in the case of the Shariq et al. protocol, any scheme which uses global temporary variables fails to achieve narrow-forward privacy and mutual authentication in the Vaudenay model with temporary state disclosure, unless these variables are protected. Based on the results from [6], PUFs represent a suitable choice for protecting temporary variables.

Furthermore, an essential element to consider is the adversarial model used in the protocol analysis. This paper also seeks to highlight the differences that may arise in the analysis when employing different security and privacy models. The Ouafi-Phan model, despite considering the adversary's corruption capability, restricts the obtained values to the secret key, thus not providing a sufficiently strong model.

These considerations aim to emphasize the importance of adopting a suitable adversarial model in protocol analysis, one that accurately reflects the practical requirements of real-life RFID systems. This was illustrated through an examination of a scheme designed for use in the healthcare field, where privacy is a critical attribute.

ACKNOWLEDGMENT

Our sincere thanks to Prof. Dr. Ţiplea for his valuable suggestions, constant support, and insightful contributions to our research.

REFERENCES

[1] S. Vaudenay, "On Privacy Models for RFID," in *Advances in Cryptology – ASIACRYPT 2007*, K. Kurosawa (Ed.), Berlin, Heidelberg: Springer, 2007, pp. 68–87, http://dx.doi.org//10.1007/978-3-540-76900-2_5.

[2] R.-I. Paise and S. Vaudenay, "Mutual Authentication in RFID: Security and Privacy," *Proceedings of the 2008 ACM Symposium on Information, Computer and Communications Security, ASIACCS '08*, March 2008, http://dx.doi.org/10.1145/1368310.1368352.

[3] J. Hermans, A. Pashalidis, F. Vercauteren, and B. Preneel, "A New RFID Privacy Model," in *Computer Security – ESORICS 2011*, V. Atluri and C. Diaz (Eds.), Berlin, Heidelberg: Springer, 2011, pp. 568–587, https://doi.org/10.1007/978-3-642-23822-2_31.

[4] J. Hermans, R. Peeters, and B. Preneel, "Proper RFID privacy: Model and protocols," *IEEE Transactions on Mobile Computing*, vol. 13, no. 12, pp. 2888–2902, December 2014, http://dx.doi.org/10.1109/TMC.2014.2314127.

[5] F. L. Tiplea, "Lessons to be Learned for a Good Design of Private RFID Schemes," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 4, pp. 2384–2395, 2022, http://dx.doi.org/10.1109/TDSC.2021.3055808. http://dx.doi.org/10.1109/TMC.2014.2314127.

[6] F. L. Tiplea and C. Hristea, "PUF Protected Variables: A Solution to RFID Security and Privacy Under Corruption With Temporary State Disclosure," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 999–1013, 2021, http://dx.doi.org/10.1109/TIFS.2020.3027147.

[7] S. Weis, S. Sarma, R. Rivest, and D. Engels, "Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems," *Lect. Note. Comput. Sci.*, vol. 2802, April 2003, http://dx.doi.org/10.1007/978-3-540-39881-3_18.

[8] M. Shariq, K. Singh, M. Y. Bajuri, A. Pantelous, A. Ahmadian, and M. Salimi, "A Secure and Reliable RFID Authentication Protocol using Schnorr Digital Cryptosystem for IoT-enabled Healthcare in COVID-19 Scenario," *Sustainable Cities and Society*, vol. 75, pp. 103354, September 2021, http://dx.doi.org/10.1016/j.scs.2021.103354.

[9] K. Ouafi and R. C.-W. Phan, "Privacy of Recent RFID Authentication Protocols," in *Information Security Practice and Experience. ISPEC 2008*, Springer, Berlin, 2008, pp. 263-277, https://doi.org/10.1007/978-3-540-79104-1_19.

[10] A. Juels and S. A. Weis, "Defining Strong Privacy for RFID," *Cryptology ePrint Archive, Paper 2006/137*, 2006. Available: https://eprint.iacr.org/2006/137

[11] D. Dolev and A. Yao, "On the Security of Public Key Protocols," *IEEE Transactions on Information Theory*, vol. 29, no. 2, pp. 198-208, 1983, https://doi.org/10.1109/TIT.1983.1056650.

[12] Y. Gao, S. Al-Sarawi, and D. Abbott, "Physical unclonable functions," *Nature Electronics*, vol. 3, no. 2, pp. 81-91, February 2020, http://dx.doi.org/10.1038/s41928-020-0372-5.

[13] F. Costa, S. Genovesi, M. Borgese, A. Michel, F. A. Dicandia, and G. Manara, "A Review of RFID Sensors, the New Frontier of Internet of Things," *Sensors*, vol. 21, no. 9, article no. 3138, 2021, http://dx.doi.org/10.3390/s21093138. Available: https://www.mdpi.com/1424-8220/21/9/3138.

[14] D. Pointcheval and J. Stern, "Security Proofs for Signature Schemes," in *EUROCRYPT 1996*, vol. 1070, pp. 387-398, 1996, http://dx.doi.org/10.1007/3-540-68339-9_33.

[15] D. Pointcheval and J. Stern, "Security Arguments for Digital Signatures and Blind Signatures," *Journal of Cryptology*, vol. 13, pp. 361-396, 2000, https://doi.org/10.1007/s001450010003.

[16] B. Halak, "Physically Unclonable Functions: Design Principles and Evaluation Metrics," in *Physically Unclonable Functions*, Springer, Cham, 2018, https://doi.org/10.1007/978-3-319-76804-5_2.

[17] D. Yamamoto, M. Takenaka, K. Sakiyama, and N. Torii, "Security Evaluation of Bistable Ring PUFs on FPGAs Using Differential and Linear Analysis," *2014 Federated Conference on Computer Science and Information Systems*, Warsaw, Poland, 2014, pp. 911-918, http://dx.doi.org/10.15439/2014F122.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 125–130 DOI: 10.15439/2024F2586 ISSN 2300-5963 ACSIS, Vol. 41

Sensitivity Analysis in Air Pollution Modeling Supported by High Performance Supercomputers

Tzvetan Ostromsky, Silvi-Maria Gurova
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS),
Acad. G. Bonchev Str., Block 25-A,
1113 Sofia, Bulgaria
Email: {ceco,smgurova}@parallel.bas.bg

Meglena Lazarova
Faculty of Applied Mathematics and Informatics, Technical University of Sofia, 8 "St. Kliment Ohridski" Blvd., 1000 Sofia, Bulgaria
Email: { meglena.laz@tu-sofia.bg }

Abstract—**Environmental modeling (and air pollution modeling in particular) is one of the toughest problems of computational mathematics. All relevant physical, chemical, and photochemical processes in the atmosphere should be taken into account. These are mathematically represented by a complex system of partial differential equations (PDE). To simplify the original PDE system proper splitting procedure is applied. As a result, the initial system is replaced by several simpler submodels, connected with the main physical and chemical processes. Even in the case of a local study of the environment in a relatively small area, the model should be calculated in a large spatial domain, because the pollutants can be moved quickly over long distances, driven by the atmosphere dynamics, especially at high altitudes. One major source of difficulty is the high intensity of the atmospheric processes, which require a small time step to be used to get a stable numerical solution (at least in the chemistry submodel). All this makes the treatment of large-scale air pollution models a heavy computational task that requires efficient numerical algorithms. It has always been a serious challenge for the fastest and most powerful supercomputers of their time. Fortunately, Bulgaria is one of the leading countries in Eastern Europe concerning the supercomputer infrastructure development in recent years.**

Index Terms—**air pollution modeling, system of PDE, sensitivity analysis, algorithm, supercomputer, scalability.**

I. INTRODUCTION

THE IMPORTANCE of environmental security has grown rapidly in the last years of the 21st century. This has become a topical issue in the decision-making process of environmental issues for developed countries around the world. To create reliable scenarios of global environmental changes and their possible consequences, it is necessary to carry out many comprehensive scientific studies on various aspects of the environment and climate. From a scientific point of view, this means carefully analyzing the most important physical and chemical processes during the transport of pollutants, including all possible transformations and interactions with other species [12], in the air over long distances. The realization of this study requires collaboration between scientific experts in the field of environmental modeling, and numerical analysis, as well as the conduct of numerous studies, based on which results are obtained that will be useful to society.

The present work aims to help develop a new reliable and computationally efficient mechanism to investigate the sensitivity of estimated concentration levels for the most important air pollutants, such as nitrogen dioxide (NO2) and especially ozone (O3), to the change of the rates of chemical reactions involved (directly or indirectly) in their formation. Both the actual data estimates and some scenarios with a certain reduction in emissions from land transport in Europe calculated by the Unified Danish Euler Model (UNI-DEM) are used in these sensitivity analysis studies [5, 6, 13].

Sensitivity analysis (SA) [10, 17, 19] is called the study of how uncertainty in the output of a model can be allocated to different sources of uncertainty in the input data, and several sensitivity analysis techniques stand out in the literature [6, 13, 16, 18]. Most existing SA methods rely heavily on assumptions about some special properties related to the behaviour of the model (such as linearity, monotonicity, and additively of the relationship between the input factor and the output of the model). Among the quantitative methods, the methods based on variances are most often used [17]. The main idea of these methods is to estimate how the variance of an input or group of inputs contributes to the variance of the model's output.

The computational tasks involved in processing large-scale air pollution models are enormous. For this reason, it is advisable to simplify the considered large-scale tasks as much as possible, but to strive to preserve the high level of reliability of the obtained results of the considered model. On the one hand, this can be obtained using sensitivity analysis. On the other hand, it is necessary to study the influence of variations of the initial conditions, boundary conditions, and/or chemical rates on the model results. This will assist in creating more correct model simplifications. This kind of analysis can provide the researcher with valuable information about reasonable simplifications and/or identify reliable parameters

©2024, PTI

and mechanisms to improve. This is because changes in parameters and mechanisms affect model outputs.

The rest of the article is structured as follows. The second section provides an overview of the Danish Eulerian Model, encompassing its high-performance parallel code UNI-DEM, and its specialized sensitivity analysis version, SA-DEM. The third section presents numerical outcomes from scalability experiments conducted with SA-DEM on two of Europe's largest supercomputers – IBM MareNostrum III in Barcelona, Spain, and the petascale supercomputer Discoverer in Sofia, Bulgaria, part of the EuroHPC high performance computing network. Finally, concluding remarks from the presented results and some plans for future work are given at the end of the paper.

II. THE DANISH EULERIAN MODEL (DEM) AND ITS PARALLEL IMPLEMENTATION

A. Historical review

The Danish Eulerian Model (DEM), boasting a robust development history spanning over 30 years [2, 3, 14, 15, 21, 22, 23], stands as a formidable tool for assessing large-scale air pollution. Throughout its evolution, it has proven its efficacy in diverse, long-term environmental studies across various domains, including but not limited to environmental protection, human health care, agricultural production, forestry, wildlife, cultural heritage preservation and others. All relevant physical and chemical processes in the atmosphere should be taken into account, which are mathematically represented by a complex PDE system. To simplify it a proper splitting procedure is applied. As a result, the initial system is replaced by several simpler systems (submodels), connected with the main physical and chemical processes. These systems should be calculated in a large spatial domain, as the pollutants migrate quickly over long distances, driven by the atmosphere dynamics, especially at high altitudes. Here they are exposed to temperature, light, and other conditions changes in an extremely wide range, and so does the speed of most chemical reactions. One of the major sources of difficulty is the dynamics of the atmospheric processes, which require a small time step to be used (at least, for the chemistry submodel) to get a stable numerical solution of the corresponding system. All this makes the treatment of large-scale air pollution models a tuff and heavy computational task. It has always been a serious challenge, even for the fastest and most powerful state-of-the-art supercomputers. [7, 23].

B. Description of the Danish Eulerian Model (DEM)

The Danish Eulerian Model (DEM) [3, 20, 21] is mathematically represented by the following system of partial differential equations:

$$(1)\quad \frac{\partial c_s}{\partial t} = -\frac{\partial(uc_s)}{\partial x} - \frac{\partial(vc_s)}{\partial y} - \frac{\partial(wc_s)}{\partial z} + \frac{\partial}{\partial x}\left(K_x\frac{\partial c_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial c_s}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial c_s}{\partial z}\right) + + E_s + Q_s\left(c_{1,}c_{2,\dots,}c_q\right) - (k_{1s}+k_{2s})c_s \;;$$

$$s=1, 2, \dots, q$$

in which the parameters' meaning is as follows:

q - number of equations (equal to the number of chemical species),

c_s - concentrations of the chemical species considered,

u, v, w - components of the wind along the coordinate axes,

K_x, K_y, K_z - diffusion coefficients,

E_s - emissions in the space domain,

k_{1s}, k_{2s} - coefficients of dry and wet deposition respectively (s = 1, …, q),

$Q_s(c_1, c_2, \dots, c_q)$ - non-linear functions that describe the chemical reactions between the species.

C. Splitting into submodels

The above rather complex system is split into three subsystems (submodels), according to the major physical and chemical processes as well as the numerical methods applied in their solution. In particular, these are:

- (i) the horizontal advection and diffusion (2);
- (ii) chemistry, emissions, and deposition (3);
- (iii) vertical exchange (4).

The submodels are described rigorously by the next formulae:

$$(2)\quad \frac{\partial c_s^{[1]}}{\partial t} = -\frac{\partial(uc_s^{[1]})}{\partial x} - \frac{\partial(vc_s^{[1]})}{\partial y} + \frac{\partial}{\partial x}\left(K_x\frac{\partial c_s^{[1]}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial c_s^{[1]}}{\partial y}\right),$$

$$(3)\quad \frac{\partial c_s^{[2]}}{\partial t} = E_s + Q_s\left(c_1^{[2]}, c_2^{[2]}, \dots, c_q^{[2]}\right) - (k_{1s} + k_{2s})c_s^{[2]},$$

$$(4)\quad \frac{\partial c_s^{[3]}}{\partial t} = -\frac{\partial(wc_s^{[3]})}{\partial z} + \frac{\partial}{\partial z}\left(K_z\frac{\partial c_s^{[3]}}{\partial z}\right).$$

The discretization of the spatial derivatives in the right-hand sides of the sub-models results in forming three large systems of ordinary differential equations.

D. Numerical methods and domain decomposition techniques applied in the submodels

A selection of various numerical methods is used in the numerical solution of the submodels. High stability requirements apply to the selected methods, because of the stiffness of the equations to be solved.

In the *advection-diffusion* submodel Finite Elements method (FE) is applied, followed by predictor-corrector schemes with several different correctors.

In the submodel for *chemistry, emissions, and deposition*, an enhanced version of the Quasi Steady-State Approximation (QSSA) is employed [8, 9, 11]. Emphasizing the pivotal role of chemical reactions, the model introduces both nonlinearity and stiffness. This atmospheric chemistry model stands out for its meticulous and precise depiction of chemical processes. The chosen chemical scheme – the condensed CBM-IV (Carbon Bond Mechanism IV), was initially proposed in [8], with subsequent improvements, specifically addressing ammonia-ammonium transformations, documented in [21]. In its basic form (involving 35 chemical species), the model accounts for 116 chemical reactions, of which 69 are time-dependent and 47 are time-independent. This scheme proves to be well-suited and adequate for investigating scenarios involving elevated concentrations of various pollutants.

In the *vertical transport* submodel, the Finite Elements method is generally used, followed by θ-methods. It should be mentioned here that there is an option to switch off the vertical transport submodel, which results in a simplified 2D version of the model.

E. *Parallel implementation*

The primary tool for parallelization is the MPI standard library. Originally designed as a communication library for distributed memory computers, MPI (Message Passing Interface) has evolved into one of the most popular parallelization tools for application programming, known for its efficiency, portability, and user-friendliness. Its versatility extends to a wide range of parallel systems, including shared-memory computers and clustered systems (where each cluster node operates as a distinct shared-memory machine), ensuring a high level of code portability.

In the context of UNI-DEM, MPI parallelization relies on spatial domain partitioning (see [13, 14, 15] for more detail).

In SA-DEM code, built up on simultaneous execution of several UNI-DEM calls with perturbed internal parameters and/or input data, there is another (coarser) level of parallelism, based on these large potentially parallel tasks. The MPI library allows these to be organized as super-tasks, working on separate communicators.

For a more in-depth explanation of the main computational stages of the model, the numerical methods and techniques engaged in each stage, as well as the parallelization strategy used in the UNI-DEM and SA-DEM codes, one can refer to [1, 4, 14, 15, 21, 23].

III. NUMERICAL RESULTS

This section presents the outcomes of scalability experiments conducted on two of Europe's most powerful general purpose supercomputers, namely the IBM MareNostrum III at the Barcelona Supercomputing Centre (BSC) in UPC - Barcelona, Spain, and the Discoverer system at Sofia Tech Park in the capital of Bulgaria. Both systems are currently parts of the EuroHPC computational infrastructure, as MareNostrum III was upgraded and up-scaled to the more powerful MareNostrum V. Tables I and II below contain summarized results of these experiments. The primary user-defined parameters for the experiments with SA-DEM / UNI-DEM on both machines were configured as follows:

- Discretization grid size: 480 × 480 (two-dimensional fine-resolution grid with step 10 km in both directions);
- Number of chemical species (q): 35;
- Main time-discretization step: 90 seconds (applied to all stages, could be reduced on the chemistry stage, if necessary

TABLE I.
EXECUTION TIME T (IN SECONDS), SPEED-UP (SP), AND PARALLEL EFFICIENCY E (IN %) OF SA-DEM (FINEST GRID) ON THE SPANISH SUPERCOMPUTER IBM MARENOSTRUM III

Time **T** [seconds], speed-up ***(Sp),*** and parallel efficiency **E** [in %]
of SA-DEM (2D MPI version) on the IBM MareNostrum III supercomputer
(480×480) mesh, 35 chem.species, NSIZE = 32

PE / MPI tasks	nodes	STAGE						TOTAL		
		Advection-diffusion			Chemistry					
		T [s]	*(Sp)*	*E* [%]	T [s]	*(Sp)*	*E* [%]	T [s]	*(Sp)*	*E* [%]
10	1	82927	(*10*)	100 %	74835	(*10*)	100 %	168847	(*10*)	100 %
40	3	19578	(*42*)	108 %	16624	(*45*)	113 %	40282	(*42*)	105 %
80	5	10178	(*81*)	103 %	8953	(*84*)	104 %	22471	(*75*)	94 %
160	10	5188	(*160*)	98 %	4632	(*162*)	101 %	13273	(*127*)	80 %
320	20	2945	(*282*)	88 %	2352	(*318*)	99 %	8331	(*203*)	63 %
640	40	1537	(*540*)	84 %	1204	(*622*)	97 %	5387	(*313*)	49 %
960	60	1206	(*688*)	72 %	815	(*918*)	96 %	4118	(*410*)	43 %
1600	100	869	(*954*)	60 %	493	(*1517*)	95 %	3373	(*501*)	31 %

in order to ensure numerical stability, as well as for more accurate calculations of the extreme values of concentrations for some unstable species, involved in very quick chemical reactions);

- Chunk size parameter NSIZE (used for improving data locality of the parallel computations and better utilization of the fastest cache memory): 32;
- Period of simulations: one year.

A. Numerical results from scalability experiments with SA-DEM by using the IBM MareNostrum III supercomputer at BSC

In Table I results of running SA-DEM on the most powerful Spanish supercomputer – IBM MareNostrum III at BSC (located in the Technical University of Catalonia (UPC) - Barcelona) are presented. The main technical parameters of this system are described on its webpage [24]. The size of the test problem makes it impractical to use less than 10 parallel tasks in the simulations. That’s why the table starts with experiment with 10 MPI tasks on one node (capable of executing up to 16 tasks in parallel on its 16 cores), assuming also a speed-up (Sp) of 10 in this case and measuring the efficiency E with respect to this initial experiment. We should mention also that not always the number of calculations determines the computing time, the amount of data transfers between different levels storage must also be taken into account. The limited size of cache (the fastest access storage) makes it more efficient in processing smaller tasks (their size is inversely proportional to their number), which explains the super-linear speed-up for the few initial rows of the table (the same applies, as well, to Table II).

B. Numerical results from scalability experiments with SA-DEM and UNI-DEM by using the largest Bulgarian supercomputer Discoverer

The outcomes from scalability experiments utilizing SA-DEM / UNI-DEM on Bulgaria's largest supercomputer (Discoverer), are reported in this section (Table 2 and Table 3). The machine was installed three years ago at Sofia Tech Park by Atos company. In November 2023, the new ranking of the top 500 supercomputers in the world was released, where Discover took place 166 [27]. This supercomputer is a crucial component of a novel network comprising eight new high-performance supercomputers across the European Union, established and supervised by the European High-Performance Computing Joint Undertaking (EuroHPC JU).

IV. CONCLUSIONS

Analyzing the results from the numerical experiments presented in the previous section the following insights were made:

- The parallel MPI implementation of SA-DEM demonstrates excellent balance, portability, and efficiency on some of leading Europe's supercomputers (including Bulgaria’s most powerful supercomputer Discoverer), integrated into the EuroHPC computing infrastructure.
- The efficiency and speed-up are higher in the computationally-intensive stages. In particular, the chemistry-deposition stage (which does not need any communication

TABLE II.
EXECUTION TIME **T** (IN *SECONDS)*, SPEED-UP ***(SP),*** AND PARALLEL EFFICIENCY **E** (*IN* %) OF EXPERIMENTS WITH SA-DEM (2D FINE GRID VERSION) ON THE BULGARIAN PETASCALE SUPERCOMPUTER DISCOVERER (PART OF THE EUROHPC INFRASTRUCTURE)

Time T [seconds], speed-up (Sp), and parallel efficiency E [in %]
of SA-DEM (2D MPI version) on DISCOVERER supercomputer
(480×480) mesh, 35 chem.species, NSIZE = 32

PE / MPI tasks	nodes	STAGE						TOTAL		
		Advection-diffusion			Chemistry					
		T [s]	*(Sp)*	*E* [%]	T [s]	(*Sp*)	*E* [%]	T [s]	(*Sp*)	*E* [%]
10	1	73028	*(10.0)*	100 %	64817	*(10.0)*	100 %	148225	*(10)*	100%
20	2	36842	*(19.8)*	99 %	30296	*(21.4)*	107 %	72341	*(20.5)*	102%
40	3	18458	*(39.6)*	99 %	15465	*(41.9)*	105%	37949	*(39.1)*	98%
80	5	9634	*(75.8)*	95 %	7981	*(81.2)*	102 %	20636	*(71.8)*	90%
120	10	4807	*(152)*	95 %	3925	*(165)*	103 %	11795	*(126)*	79%
320	20	2525	*(289)*	90 %	2037	*(318)*	99 %	6861	*(216)*	68%
640	40	1290	*(566)*	88 %	1034	*(627)*	98 %	4810	*(308)*	48%
960	60	908	*(805)*	84 %	697	*(930)*	97 %	3363	*(441)*	46%
1600	100	764	*(956)*	60 %	445	*(1457)*	91 %	2781	*(533)*	33%

between the tasks) has almost linear overall speed-up, even super-linear concerning the number of MPI tasks (a cache-size effect due to decreasing size of processed data per task with increasing the number of MPI tasks).

- The advection-diffusion stage scales pretty well too, taking into account that there is some unavoidable computational overhead due to overlapping subdomain boundaries of the partitioning. In the chemically-intensive stages, the hybrid MPI-OpenMP code leverages a reduction in time, thanks to the involvement of multiple threads with the OpenMP lower level of parallelism activated at the core level within a node. However, this optimization has minimal impact on other stages.
- To maintain comparability, the size of chunks at the chemistry-deposition stage (NSIZE=32) has been consistent across all experiments. Although this might not be optimal in every scenario, particularly when the number of MPI tasks is high, using a lower value for NSIZE can potentially yield better results.
- As computations span a substantial number of nodes, communications between MPI processes on different nodes become costly, resulting in some speed-up degradation and reduced overall efficiency. To address this, efforts should focus on optimizing I/O processes to mitigate this negative impact, forming part of our future tasks.

ACKNOWLEDGMENT

This research of the authors was partially supported by the following projects:

The work of Tzvetan Ostromsky has been supported by the Bulgarian National Science Fund via the project "Efficient methods for modeling, optimization and decision making" (KP-06-N52/5) as well as by the Ministry of Education and Science (MES) of Bulgaria (Agreement № Д01-27/06.02.2024**).**

The work of Silvi-Maria Gurova was carried out using the infrastructure purchased under the National Roadmap for RI, financially coordinated by the MES of the Republic of Bulgaria (Grant No. DO1-325/01.12.2023) and supported also by Grant No 30 from CAF America.

The work of Meglena Lazarova was supported by the Bulgarian National Science Fund under Project KP-06-M62/1 "Numerical deterministic, stochastic, machine and deep learning methods with applications in computational, quantitative, algorithmic finance, biomathematics, ecology and algebra" from 2022.

The European High-Performance Computing Joint Undertaking (Euro-HPC JU) is kindly acknowledged for granting us access and computer time on the DISCOVERER supercomputer.

REFERENCES

[1] Alexandrov V. and et al., Numerical integration of chemical ODE problems arising in air pollution models. Env. Modeling and Assessment **2,** 1997, 365–377.

[2] Christensen J., The Danish Eulerian Hemispheric Model. In: Gryning SE. Schiermeier F.A. (eds.) Air Pollution Modeling and Its Application XI NATO Challenges of Modern Society 21 Springer Boston MA, 1996.

[3] Christensen J., The Danish Eulerian hemispheric model — a three-dimensional air pollution model used for the arctic. Atmospheric Environment, **31(24),** 1997**,** 4169-4191.

[4] Dimov I., Monte Carlo Methods For Applied Scientists. World Scientific, 2007.

[5] Dimov I. et al., A Super-Convergent Stochastic Method Based on the Sobol Sequence for Multidimensional Sensitivity Analysis in Environmental Protection. *Axioms*, 2023, **12(2):**146.

[6] Dimov I. et al., Studying the sensitivity of pollutants' concentrations caused by variations of chemical rates. J. Comput. Appl. Math. **235,** 2010**,** 391–402.

[7] Dimov I. et al., Computational challenges in the numerical treatment of large air pollution models. Ecological Modelling **179**, 2004, 187–203.

[8] Gery M. et al., A photochemical kinetics mechanism for urban and regional scale computer modelling. J. Geophys Res. **94(D10),** 1989, 12925–12956.

[9] Hesstvedt E. et al., Quasi-steady-state approximations in air pollution modeling: comparison of two numerical schemes for oxidant prediction. Int. Journal of Chemical Kinetics **10**, 1978, 971–994.

[10] Homma T., Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering and System Safety **52**, 1996, 1–17.

[11] Hov Ø. et al., Comparison of numerical techniques for use in air pollution models with non-linear chemical reactions. Atmospheric Environment **23**, 1988, 967–983.

[12] Hvidtfeldt U. A. et al., Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon. Environmental Epidemiology **2(2**), 2018, pe014.

[13] Ostromsky Tz. et al., Air pollution modeling, sensitivity analysis, and parallel implementation. Int. Journal of Environment and Pollution **46 (1-2),** 2011, 83–96.

[14] Ostromsky Tz. and Zlatev Z., Parallel Implementation of a Large-scale 3-D Air Pollution Model. Large Scale Scientific Computing (S. Margenov, J. Wasniewski, P. Yalamov, Eds.) Spinger **2179**, 2001, 309–316.

[15] Ostromsky Tz. and Zlatev Z., Flexible Two-level Parallel Implementations of a Large Air Pollution Model. Numerical Methods and Applications (I.Dimov, I.Lirkov, S. Margenov, Z. Zlatev eds.) LNCS **2542**, 2002, 545–554.

[16] Ostromsky Tz. et al., Advanced Sensitivity Analysis of the Danish Eulerian Model in Parallel and Grid Environment. AIP Conf. Proc**. 1404**, 2011, 225–232.

[17] Saltelli A. et al., Sensitivity Analysis. John Wiley & Sons publishers, Probability and Statistics series, 2000.

[18] Saltelli A. et al., Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, Halsted Press New York, 2004.

[19] Sobol I., Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates Mathematics and Computers in Simulation, **55 (1-3**), 2001, 271–280.

[20] WEB-site of the Danish Eulerian Model: https://www2.dmu.dk/AtmosphericEnvironment/DEM/ (Last accessed December 2023).

[21] Zlatev Z., Computer treatment of large air pollution models Kluwer, 1995.

[22] Zahari Z., Comprehensive Air Pollution Studies with the Unified Danish Eulerian Model, 2003, 1125-1137.

[23] Zlatev Z. and Dimov I., Computational and Numerical Challenges in Environmental Modelling Elsevier Amsterdam, 2006.

[24] Technical documentation of MareNostrum III: https://www.bsc.es/marenostrum/marenostrum/mn3 (Last accessed April 2024)

[25] The official web page of the supercomputer Discoverer: https://sofiatech.bg/en/petascale-supercomputer/ (Last accessed April 2024)

[26] Discoverer HPC Doc: https://docs.discoverer.bg/index.html (Last accessed April 2024)

[27] Top500 list of supercomputers – November 2023: https://www.top500.org/system/179948/ (Last accessed April 2024)

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 131–138 DOI: 10.15439/2024F7665 ISSN 2300-5963 ACSIS, Vol. 41

Handling Lot Sizing/Job Scheduling Synchronization through Path Search Algorithms

Alain Quilliot and Aurélien Mombelli
LIMOS Lab.
UCA, CNRS and EMSE
Clermont-Ferrand, France
Email: alain.quilliot@uca.fr
LIMOS Lab.
Clermont-Ferrand, France

***Abstract*—We deal here with the synchronization of a resource production process and the consumption of those resources by machines used in order to run a set of jobs. Resources are produced by a *Resource Lot Sizer* player, under some economic costs and technical constraints. Both the *Resource Lot Sizer* and the *Job Scheduler* interact through *transfer transactions*, that involve specific constraints, making the synchronization between both processes become an issue. Resulting decision problem appears as a multi-stage *Lot Sizing* problem, arising in contexts where the resource is some renewable energy (hydrogen, electricity, ...) required by the jobs and stored inside tanks or batteries embedded into the machines (vehicles or robots). We first cast it into the MILP format, perform a structural analysis of the transfer transactions and handle resulting MILP_SLSS model through Branch and Cut before reformulating it as a path search problem set in a specific *Transfer* space and handling through a filtered adaptation of the A* algorithm.**

I. INTRODUCTION

LOT SIZING arises when one schedules production while taking care of avoiding strong activation costs (see [3]). A *Lot Sizing* strategy consists in concentrating production on well-chosen periods in order to minimize those activation costs. Resulting problems look like Knapsack problems involving a temporal dimension. One often handles them through dynamic programming, that may yields FPTAS: *Fully Polynomial Time Approximation Scheme*. Multi-dimensional Lot Sizing (see [**?**]) means a heterogeneous production and different kinds of capacity constraints (weight, volume, ...). Multi-level Lot Sizing means several players acting on a same system (a main resource producer and several subcontractors). Those players may be partially independent from each others, raising the collaborative issue and leading to cooperative game or multi-objective reformulations (see [2]). In any case, resulting problems are NP-Hard.

Now it may happen that such a process takes place in the context of an interaction between the *Lot Sizer* player, that is a *Resource Lot Sizer*, and a (several) *Job Scheduler(s)* who uses resources in order to run jobs (production or services) on parallel machines (vehicles or industrial robots). It is specifically the case when the resource is energy (power, hydrogen, biofuel, ...) loaded into batteries or tanks embedded into the machines (see [4]). In such a case, both the resource lot sizer and the job scheduler interact through *transfer transactions*, consisting in transferring resources (the energy) from the resource production plant into the storage facilities embedded into the machines. Both players must agree about the production periods i and the job dates t when the energy producer transfers some energy amount m to the job scheduler. Such a *transfer transaction* $\omega = (i, t, m)$ usually involves its own constraints and costs. According to this, we get a full description of our process if we know, besides the production vector z and the schedule of the jobs, the collection of all *transfer transaction* $\omega = (i, t, m)$ which make both players meet.

As the *Lot Sizing* problems, job scheduling problems have been widely investigated in the past (see [3]). It may arise in contexts related not only to energy management (see [2], [8]), but also to real time cooperation between sensors and robots, to the interaction between electric vehicles and recharge facilities (see [4], [5]). It raises the *centralized versus collaborative* issue (see [2]) since involved players may be independent. However, though both the *Lot Sizing* and job scheduling problems have been extensively studied, few authors addressed the synchronization issue that is at the core of the situation that we just described above (see [1]). Our main purpose here is to address it.

We stick here to the centralized paradigm and focus on the combinatorial issues related to synchronization and suppose that the processing order of the jobs $j = 0, \dots, M-1$ on the machine have already been decided. We denote by **SLSS**: *Synchronized Lot Sizing/Scheduling* resulting problem that may be viewed as containing the *core* of the synchronization issue. We first make appear the central role played here by the *transfer transactions*: We check that, for any feasible solution of **SLSS**, related transfer transactions define a chain inside some partially ordered set. This leads us to next design a MILP: *Mixed Integer Linear Programming* setting of **SLSS** that is centered on those transfer transactions and involve complex *no-antichain* constraints. We check that, though those *no-antichain* constraints are exponential, they may be separated in polynomial time, opening the way to the handling of **SLSS** through Branch and Cut. Yet, this approach remains time consuming. Keeping on with the idea of taking advantage of the specific structure of the transfer transactions as a partially

©2024, PTI

ordered set, we reformulate **SLSS** problem as a path search problem inside some large size *Transfer* network. We deal with this path search reformulation while designing a *Path_SLSS* algorithm that proves itself far more efficient than the MILP setting and allows us to state that **SLSS** is pseudo-polynomial. So the paper is organized as follows. We describe in Section II the **SLSS** problem. In Section III, we study the structure of the transfer transactions and set a transfer transaction driven MILP formulation **MILP_SLSS** of **SLSS**, that involves *no-antichain* constraints. We explain in Section IV the way those *no-antichain* constraints may be separated, allowing us to deal with **MILP_SLSS** through branch and cut. Section V is devoted to our main contribution, that consists in a path search reformulation of **SLSS** and its more efficient handling through the design of a filtered A* like path search algorithm ([7]). This algorithm allows us to check that **SLSS** is pseudo-polynomial. We conclude with numerical experiments.

II. THE SYNCHRONIZED LOT SIZING/SCHEDULING PROBLEM

We consider:

- **A Resource Lot Sizer:** The time horizon of the resource lot sizer is divided into N periods $i = 0, \ldots, N-1$, all with equal duration p. Thus the starting time of period i is equal to $p.i$. During a period i, this resource producer may be active or idle. If it is active, then it produces R_i resources, under a variable *running* cost $Cost_i^R$. If it is idle, then turning it into active induces an additional *activation* cost $Cost_i^A$. It is provided with a resource storage facility, with capacity C^P and initial load H_0^P. It cannot produce if resulting load exceeds C^P. This last hypothesis will be discussed in Section II.*B*. It must end with a load at least equal to H_0^P and while minimizing its production cost.
- **A Job Scheduler:** It must sequentially perform M jobs $j = 0, \ldots, M-1$ on a single machine, according to this order, within a time horizon $[0, N.p]$. Running a job j requires e_j resource units and t_j time units. It is provided with a resource storage facility, with capacity C^S and initial load equal to H_0^S. It must end with a load at least equal to H_0^S while minimizing that makespan, that means the ending time of job $M-1$.
- **Transfer Transactions:** Because of its limited storage capacity C^S, the job scheduler must periodically receive resources from the resource producer. Such a *transfer transaction* (see Figure 1) takes place during a whole period i, between the end of some job j and the beginning of its successor $j+1$, and involves the transfer of m resources. It is denoted as $\omega = (i, j, m)$, where $m \leq Inf(C^P, C^S)$. Performing $\omega = (i, j, m)$ requires a full period i, during which production is forbidden. It also requires from the job scheduler ϵ_j additional resources, τ_j additional time and a potential waiting time to be spent before the transfer transaction takes place. One must understand those requirements as related to a kind of detour (if the machine is a vehicle) or set up (if it is a robot). The consequence is that if $\omega = (i, j, m)$ defines a transfer transaction and if we denote by T_j the starting time of j, by V_j^S the resource load of the job scheduler at the time when j starts and by V_i^P the resource load of the resource lot sizer at the beginning of period i, then:
  - The starting time $p.i$ of period i must be at least equal to $T_j + t_j + \tau_j$ (augmented with a possible waiting time), while the ending time $p.(i+1)$ of i will coincide with the starting time T_{j+1} of job $j+1$.
  - The load V_{i+1}^P of the resource lot sizer at the beginning of period $i+1$ must be equal to $V_i^P - m$ and must be non negative.
  - The load V_{j+1}^S of the job scheduler at the beginning of job $j+1$ must be equal to $V_j^S - \epsilon_j + m$ must not exceed C^S; Its load $V_j^P - \epsilon_j$ at the beginning of period i must be non negative.

  Since those additional coefficients τ_j and ϵ_j most often corresponds to some kind of detour, we suppose that the following triangle inequalities hold for any $j \leq M-1$: $\tau_j \leq \tau_{j+1} + t_{j+1}$ and $\epsilon_j \leq \epsilon_{j+1} + e_{j+1}$.
  These inequalities mean that if the job scheduler has no time left for a transfer transaction at the end of job j, then keeping on with another job will not provide it with the missing time.
  In order to allow resources to be transferred before job 0 or after job $M-1$ we introduce a fictitious job -1, such that $t_{-1} = 0, e_{-1} = 0$, τ_{-1} and ϵ_{-1} being non null additional coefficients, and another fictitious job M, with $t_M = e_{-M} = \tau_M = \epsilon_M = 0$.

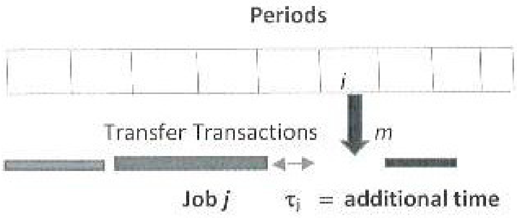


Fig. 1. A Transfer Transaction

Then solving the **SLSS:** ***Synchronized Lot Sizing/Scheduling*** problem means, in a natural way, computing the resource production periods, the starting times of the jobs, and the transfer transactions that link both players in such a way that:

- All jobs are done.
- Capacity constraints, job resource requirements and final state constraints are met.
- Some cost $\alpha.\sum_i PCost + \beta.T$ is minimized, where $PCost$ is the production cost, T is the ending time of the fictitious job M and α, β are *time versus money* coefficients.

III. A *Transfer Transactions* DRIVEN A MILP MODEL

As told in the introduction, transfer transactions (i, j, m) will be the leader component of any **SLSS** algorithmic solution.

A. The Transfer Transaction Partially Ordered Set

Let us first introduce some additional definitions:

Resource and Time Delays.

Let us recall that a transfer transaction $\omega = (i, j, m)$ starts at time $p.i$ and lasts one period. It starts just after the ending time of job j, augmented with delay τ_j and with a possible waiting time. Its ending time $p.(i+1)$ coincides with the starting time of $j+1$. Besides, it requires an additional resource ϵ_j, that the job scheduler consums just before the transaction takes place. A transfer transaction may take place just after the end of job $M-1$, at the very end of the process. It may also occur just before the beginning of job 0 or, in other words, between the fictitious job -1 provided with values ϵ_{-1} and τ_{-1} and job 0.

For any job pair (j_1, j_2) s.t. $0 \leq j_1 < j_2 \leq M-1$ we denote by $\mu(j_1, j_2) = \epsilon_{j_2} + \sum\limits_{j_1+1 \leq j \leq j_2} e_j$ the amount of resources required between two transfer transactions respectively related to j_1 (just after j_1) and j_2 (just after j_2). We call this quantity the *resource_delay* between j_1 and j_2. This definition applies to $j_1 = -1$. By the same way, we set, for any $j_1 = -1, \dots, M-1$:

- $\mu^{Start}(j_1) = (\sum_{j \leq j_1} e_j) + \epsilon_{j_1}$ that expresses the resource consumed between the beginning of the process and the first transfer transaction in case this transfer occurs just after j_1. We call it the *start_ resource_delay* induced by j_1. Notice that if $j_1 \neq -1$, then $\mu^{Start}(j_1) = \mu(-1, j_1)$.
- $\mu^{End}(j_1) = (\sum_{j > j_1} e_j)$ that expresses the resource consumed between the last transfer transaction in case this transfer transaction occurs just after j_1 and the end of the process. We call it the *end_resource_delay* induced by j_1. Notice that $\mu^{End}(j_1)$ might also be written $\mu(j_1, M)$ if we extend the definition of $\mu(j_1, j_2)$.

Considering the time instead of the resource, we set $\Delta(j_1, j_2) = \tau_{j_2} + \sum\limits_{j_1+1 \leq j \leq j_2} t_j$, that means the time required between two transfer transactions respectively related to j_1 and j_2. This definition applies to $j_1 = -1$. We call this quantity the *time _delay* between j_1 and j_2. By the same way, we set, for any $j_1 = -1, \dots, M-1$:

- $\Delta^{Start}(j_1) = (\sum_{j \leq j_1} t_j) + \tau_{j_1}$ that expresses the minimal time between the beginning of the process and the first transfer transaction in case this transaction occurs just after j_1. We call it the *start_time_delay* induced by j_1.
- $\Delta^{End}(j_1) = (\sum_{j > j_1} t_j)$ that expresses the time required between the last transfer transaction in case this transfer transaction occurs just after j_1 and the end of the process. We call it the *end_time_delay* induced by j_1.

We derive from those *time_delays* the following informations:

- We set $\pi_m(j) = \left\lceil \frac{\Delta^{Start}(j)}{p} \right\rceil$. We easily check that $\pi_m(j)$ means the first possible period for a transfer transaction involving a job $j \in \{-1, \dots, M-1\}$.
- We set $\pi_M(j) = (N-1) - \left\lceil \frac{\Delta^{End}(j)}{p} \right\rceil$. We easily check that $\pi_M(j)$ means the last possible period for a transfer transaction involving $j \in \{-1, \dots, M-1\}$.

A consequence of the triangle inequalities $\tau_j \leq \tau_{j+1} + t_{j+1}$ is that the interval $\{\pi_m(j), \dots, \pi_M(j)\}$ provides us with the periods when a transfer transaction involving j may occur.

Also:

- For any j such that $M-1 \geq j \geq 0$, we denote by $\Phi(j)$ the unique job j_1 (it may not exist) such that $j < j1 \leq M-2$ and $\mu(j, j_1) \leq C^S$ and that $\mu(j, j_1+1) > C^S$.
- By the way, we denote by $\Phi(-1)$ the largest job j_1 such that $\mu^{Start}(j_1) \leq H_0^S$ and by $\Phi(M)$ the smallest j_1 such that $\mu^{End}(j_1) \leq C^S - H_0^S$.

A consequence of the inequalities $\epsilon_j \leq \epsilon_{j+1} + e_{j+1}$ is that if $\Phi(j)$ exists, then it tells us that at least one transfer transaction must take place that involves j_1 belonging to the period interval $\{j+1, \dots, \Phi(j)\}$. If such a transaction were not existing, then moving ahead of $\Phi(j)$ or moving backward from j would not allow make appear any feasible transfer transaction, and the whole process would become infeasible.

Finally we denote by $C^{TR} = min(C^P, C^S)$ the maximal value m that may be involved into a transfer transaction.

The *Transfer Transaction* Partially Ordered Set $(\Omega, \ll)$.

We define Ω as the set of all (period/job) pairs (i, j) such that $\tau_m(j) \leq i \leq \tau_M(j)$, augmented with 2 fictitious pairs $Source$ and $Sink$. It contains all pairs (i, j) that may be involved into a transfer transaction.

We set $(i_1, j_1) \ll (i_2, j_2)$ if and only if a transfer transaction related to period i_2 and job j_2 can be preceded by another transfer transaction between period i_1 and station j_1 in the sense of the time delays, that means if $j_2 > j_1$ and $(i_2 - i_1 - 1).p \geq \Delta(j_1, j_2)$. We set $Source \ll (i, j)$ and $(i, j) \ll Sink$ for any (i, j). We easily check that this relation $\ll$ is transitive and anti-reflexive and so defines a partial order relation on Ω.

The order relation $\ll$ induces on the set Ω an oriented graph structure $G^{\ll}$. Any arc $((i_1, j_1), (i_2, j_2))$ of $G^{\ll}$ may be provided with a length $L^{\ll}((i_1, j_1), (i_2, j_2))$ as follows:

- If both (i_1, j_1) and (i_2, j_2) are different from $Source$ and $Sink$, then $L^{\ll}((i_1, j_1), (i_2, j_2)) = \mu(j_1, j_2)$;
- If $(i_1, j_1) = Source$ then $L^{\ll}((i_1, j_1), (i_2, j_2)) = \mu^{Start}(j_2)$;
- If $(i_2, j_2) = Sink$ then $L^{\ll}((i_1, j_1), (i_2, j_2)) = \mu^{End}(j_1)$.

Lemma 1: *Given a feasible solution of* **SLSS**. *Then the transfer transactions related to this solution define a path* Γ *in the acyclic graph* $G^{\ll}$*, that connects* $Source$ *to* $Sink$ *and is such that, for any arc* $((i_1, j_1), (i_2, j_2))$ *of* Γ*:*

- *If* (i_1, j_1) and (i_2, j_2) *are both different from* $Source$ *and* $Sink$*, then:* $L^{\ll}((i_1, j_1), (i_2, j_2)) \leq C^S$;
- *If* $(i_1, j_1) = Source$ *then* $L^{\ll}((i_1, j_1), (i_2, j_2)) \leq H_0^S$;

- *If* $(i_2, j_2) = Sink$ *then* $L^{\ll}((i_1, j_1), (i_2, j_2)) \leq C^S - H_0^S$.

Proof: TTw consecutive transfer transactions $(i_1, j_1, m_1), (i_2, j_2, m_2)$ involved into a feasible solution of **SLSS** define an arc in the sense of $(\Omega, \ll)$ because the time between the end of i_1 and the beginning of i_2 must be large enough in order to make possible processing the jobs $j_1 + 1, \ldots, j_2$. The bound imposed to the length of $((i_1, j_1), (i_2, j_2))$ means that the resources needed in order to process those jobs under the additional resource ϵ_{j_2} requirement cannot exceed the storage capacity C^S. □

B. A MILP Formulation of ***SLSS***

Let us recall that a *≪-Antichain* of the partially ordered set $(\Omega, \ll)$ is any subset of Ω made of pairs (i, j) pairwise incomparable in the sense of $\ll$. Then we set **SLSS_MILP** by first introducing the following variables:

- **Transfer Transaction Variables**
  - {0, 1}-valued vector $U = (U_{i,j}, i = 0, \ldots, N-1, j = -1, \ldots, M-1)$: $U_{i,j} = 1$ means that some transfer transaction (i, j, m) is part of the solution.
  - Rational non negative vector $m = (m_{i,j}, i = 0, \ldots, N-1, j = -1, \ldots, M-1)$: If $U_{i,j} = 1$ then $m_{i,j}$ means related resources.
- **Resource Lot Sizing Variables**
  - {0, 1}-valued vector $y = (y_i, i = 0, \ldots, N-1)$: $y_i = 1$ means that the production is activated at the beginning of period i.
  - {0, 1}-valued vector $z = (z_i, i = 0, \ldots, N-1)$: $z_i = 1$ means the production active at period i.
  - {0, 1}-valued vector $\delta = (\delta_i, i = 0, \ldots, N-1)$: $\delta_i = 1$ means a transfer transaction at period i.
- **Job Variables**
  - T: $T =$ the ending time of the job M.
  - {0, 1}-valued vector $\gamma = (\gamma_j, j = -1, \ldots, M-1)$: $\gamma_j = 1$ means that a transfer transaction takes place at the end of job j. (If $j = -1$ then it means that a transfer transaction takes place just before job 0).

Then, the constraints of the **SLSS_MILP** come as follows:

- **Structural Constraints**

$$\text{Minimize} \sum_i (Cost_i^R . z_i + Cost_i^A . y_i) + \alpha . T \qquad (*)$$

$$\forall j : \quad T \geq \sum_i U_{i,j} . (p(i+1) + \Delta^{End}(j)) \qquad (C.1)$$

$$y_0 - z_0 \geq 0 \qquad (C.2)$$

$$\forall i \geq 1 : \quad y_i \geq z_i - z_{i-1} \qquad (C.3)$$

$$\forall i \geq 0 : \quad z_i + \delta_i \leq 1 \qquad (C.4)$$

$$\forall i : \quad \delta_i = \sum_j U_{i,j} \qquad (C.5)$$

$$\forall j : \quad \gamma_j = \sum_i U_{i,j} \qquad (C.6)$$

$$\forall (i, j) \ \text{ s. t.} (i > \tau_M(j)) \vee (i < \tau_m(j)) : \quad U_{i,j} = 0 \quad (C.7)$$

$$\forall A, A \ \ll\text{-Antichain} : \quad \sum_{(i,j) \in A} U_{i,j} \leq 1$$

$(C.8$: *No-Antichain Constraints*$)$

$\forall j$ s. t. $-1 \leq j \leq M-1$:

$$\sum_{k \in \{j+1, \ldots, \Phi(j)\}} \gamma_k \geq 1 \qquad (C.9)$$

$$\sum_{k \in \{-1, \ldots, \Phi(-1)\}} \gamma_k \geq 1 \qquad (C.10)$$

$$\sum_{k \in \{\Phi(M), \ldots, M-1\}} \gamma_k \geq 1 \qquad (C.11)$$

$$T \geq \sum_j (t_j + \tau_j . \gamma_j) \qquad (C.12)$$

$$\forall j : \quad H_0^P + \sum_{k < \tau_M(j)} z_k . R_k \geq \sum_{k \leq j} (\gamma_k . \epsilon_k + e_k) - H_0^S (R.1)$$

- **Resource Amount Constraints**

$$\forall i, j : \quad m_{i,j} \leq U_{i,j} . C^{TR} \qquad (TL.1)$$

$$\forall i : \quad \sum_{j, k \leq i} m_{k,j} - \sum_{k < i} z_k . R_k \leq H_0^P \qquad (TL.2)$$

$$\forall i : \quad \sum_{k \leq i} z_k . R_k - \sum_{j, k \leq i} m_{k,j} \leq C^P - H_0^P \qquad (TL.3)$$

$$\sum_i z_i . R_i - \sum_{i,j} m_{i,j} \geq 0 \qquad (TL.4)$$

$\forall j$ s. t. $0 \leq j \leq M-1$:

$$\sum_{i, k < j} m_{i,k} + H_0^S \geq \sum_{k \leq j} (e_k + \gamma_k . \epsilon_k) \qquad (TL.5)$$

$$\sum_{i,j} m_{i,j} \geq \sum_{j \leq M-1} (e_j + \gamma_j . \epsilon_j) \qquad (TL.6)$$

$\forall j$ s. t. $-1 \leq j \leq M-1$:

$$\sum_{i, k \leq j} m_{i,k} \leq C^S - H_0^S + \sum_{k \leq j} (e_k + \gamma_k . \epsilon_k) \quad (TL.7)$$

Theorem 1. *The MILP program* ***SLSS_MILP*** *model solves the* ***SLSS*** *problem in an exact way.*

Proof. (*) clearly means the objective function of **SLSS**, while one easily checks that (C.1, C.2, C.3, C.4, C.5, C.6, C.7, C.8, C9, C10, C11, TL.1, TL.2, TL.3, TL.6) are necessary. (C.12): T is at least equal to the running times of the jobs and the transfer transactions. (R.1): For any job j, the resources produced before $j+1$ augmented with the initial resources, must be enough for jobs 0 to $j+1$. (TL.4): We must globally produce at least as much as we transfer. (TL.5): For any j, the resources needed in order to perform jobs $0, \ldots, j$ together with the related transfer transactions must not exceed available resources (the triangle inequalities are involved). (TL.7): For any j, the resources of the job scheduler after running j must not exceed C^P.

Conversely, we get sufficiency by first checking that constraints (C.5, ..., C.8) imply that the transfer transactions deriving from U define a chain in $(\Omega, \ll)$. Constraints (C.9, C.10, C11) make this chain consistent with Lemma 1. Let us denote it by $\Gamma = \{(i_1, j_1), \dots, (i_Q, j_Q)\}$ and let us introduce variables $T_j, V_j^S, j = 0, \dots, M-1, \hat{V}_q^S, q = 1, \dots, Q, V_i^P, i = 0, \dots, N-1, \hat{V}_q^P, q = 1, \dots, Q$, respectively representing the starting time of job j, the resources stored by the job scheduler at time T_j, the resources stored by the job scheduler just before the qth transfer transaction takes place, the resources stored by the resource lot sizer at the beginning of period i and the resources stored by the resource lot sizer just before the qth transfer transaction takes place (at the beginning of period i_q). Then it is enough to follow Γ in order to provide those variables with values defining a feasible schedule. □

IV. A BRANCH AND CUT FOR THE **SLSS_MILP** MODEL

Applying a MILP library to **SLSS_MILP** means designing a procedure in order to separate the *No_Antichain* constraints (C.8), that means in order to dynamically generate and insert new *No_Antichain* constraints as soon as a current solution of the rational relaxation of **SLSS_MILP** appears, that does not meet those constraints. Let U be some rational (or integral) vector satisfying (C4, C5, C6, C7). Separating (C.8) means searching for an antichain A in $(\Omega, \ll)$ such that $\sum_{(i,j) \in A} U_{i,j} > 1$. Theoretically (see [6]) it can be done in polynomial time by application of a feasible network flow procedure. However, the arcs of $G^{\ll}$ are too many for an efficient application of such a separation procedure. Instead, we proceed in a heuristic way, through partial tree search. Related *backtracking* nodes correspond to sequences $\{(i_1, j_1), \dots, (i_k, j_k)\}$ ordered according to decreasing $U_{i,j}$ values and whose elements define an antichain in Ω. In order to speed the process, we arbitrarily impose an upper bound on the number of backtracking nodes created this way and use this upper bound as a control parameter. If U is integral, then we only need to scan the transfer transactions and check that 2 consecutive transfer transactions agree with $\ll$.

V. HANDLING SLSS AS A PATH SEARCH PROBLEM

Let us come now to our main contribution. We are going to show that **SLSS** may be reformulated as a path search problem in an acyclic large state network. This will allow us to handle it while adapting the well-known A* algorithm (see [7]) for the computation of robot trajectories inside state spaces. This approach will prove itself to be far more efficient that the MILP based approach.

A. The SLSS Path Search Formulation

We saw in Section III.*A* that the transfer transactions involved in any feasible **SLSS** solution define in the graph $G^{\ll}$ a path Γ from $Source$ to $Sink$ such that if (i_1, j_1) and (i_2, j_2) are consecutive in Γ, then: (P0)

- If $(i_1, j_1) \neq Source$ and if $(i_2, j_2) \neq Sink$ then $\mu(j_1, j_2) \le C^S$
- If $(i_1, j_1) = Source$ then $\mu^{Start}(j_2) \le H_0^S$
- If $(i_2, j_2) = Sink$ then $\mu^{End}(j_1) \le C^S - H_0^S$

Such a path $\Gamma = \{Source, (i_1, j_1), (i_2, j_2), \dots, (i_Q, j_Q), Sink\}$ does not fully characterize a **SLSS** feasible solution. But we are going to show that it will do it as soon as we are able to provide it with well-fitted sequences $w = \{w_q, \quad q = 0, 1, \dots, Q, Q+1\}$ and $W = \{W_q, q = 0, 1, \dots, Q\}$, with the meaning:

- For any $q = 1, \dots, Q$, w_q means the resources globally stored together by the resource lot sizer and the job scheduler at the end of period i_q. For any $q = 1, \dots, Q-1$, W_q denotes the resource produced between period i_q and period i_{q+1}.
- $w_0 = H_0^P + H_0^S =$ the initial resources and W_0 means the resources produced before period i_1. $W_Q =$ the resources produced after period Q.
- $w_{Q+1} \ge H_0^P + H_0^S$ means the final global resources.

Then we may characterize the **SLSS** solutions according to the following statements:

Theorem 2. *Given a path $\Gamma = \{Source, (i_1, j_1), (i_2, j_2), \dots, (i_Q, j_Q), Sink\}$ that meets (P0), together with non negative values $w_q, q = 0, 1, \dots, Q, Q+1$ and $W_q, q = 0, 1, \dots, Q$. They may be extended into a feasible solution of* **SLSS** *iff:*

1) $w_1 = w_0 + W_0 - \mu^{Start}(j_1)$;
   $w_0 = H_0^P + H_0^S$;
   $W_0 \le C^P - H_0^P$;
   W_0 *is a feasible production for periods in* $\{0, \dots, i_1 - 1\}$. *(P1)*
2) *For any* $q = 1, \dots, Q-1$: *(P2)*
   - $W_q \le min(C^P, C^S + C^P - w_q)$ *is a feasible production for periods in* $\{i_q + 1, \dots, i_{q+1} - 1\}$;
   - $w_{q+1} = w_q + W_q - \mu(j_q, j_{q+1})$;
   - $\mu(j_q, j_{q+1} \le Inf(w_q, C^S)$.
3) $W_Q \le min(C^P, C^S + C^P - w_q)$ *is a feasible production for periods in* $\{i_Q + 1, \dots, N-1\}$;
   $w_{Q+1} = (w_Q + W_Q - \mu^{End}(j_Q) \ge H_0^P + H_0^S$; *(P3)*

Proof. Lemma 1 tells us that we may derive from path Γ starting times values $T_j, j = 0, \dots, M-1$ for the jobs j, that are consistent with the time requirements of **SLSS**. Let us suppose that we know the values $\hat{w}_q^S, q = 1, \dots, Q, \hat{w}_q^P, q = 1, \dots, Q$ denoting the resources available for respectively the job scheduler and the resource lot sizer just after the qth transfer transaction takes place (at the end of period i_q). Then we easily deduce the resources $V_j^S, j = 0, \dots, M-1$ available for the job scheduler at time T_j, and the resources $V_i^P, i = 0, \dots, N-1$ available for the resource producer at time $p.i$. So the key point becomes distributing global resource quantities w_q among the job scheduler and the resource lot sizer, and get quantities $\hat{w}_q^S, q = 1, \dots, Q, \hat{w}_q^P, q = 1, \dots, Q$ that mean the resources available for respectively the job scheduler and the resource lot sizer at the end of period i_q, in such a way that resulting quantities $V_j^S, j = 0, \dots, M-1$

and $V_i^P, i = 0, \dots, N-1$ meet the resource requirements of **SLSS**. We get it by applying the following rule: For every q, we complete the pair (i_q, j_q) with a transfer value m_q in such a way that either we completely fill the storage device of the job scheduler ($\hat{w}_q^S = C^S$) or, if it is not possible, we make the storage facility of the resource lot sizer become empty ($\hat{w}_q^P = 0$). Then, it becomes a matter of routine to check that resulting $V_j^S, j = 0, \dots, M-1$ and $V_i^P, i = 0, \dots, N-1$ meet the resource requirements of **SLSS**. □

A Path Search Formulation of SLSS.

Theorem 2 suggests us the construction of the following *Augmented Transfer* oriented graph $G^{Augment,\ll}$:

- The nodes of $G^{Augment,\ll}$ are pairs (x, w), where x is a node of $G^{\ll}$ and w is a resource value such that $0 \le w \le C^S + C^P$; We define the node $(Source, H_0^S + E_0^P)$ as a *source* node and any node $(Sink, w \ge H_0^S + H_0^P)$ as a *sink* node.
- We say that $((x_1, w_1), (x_2, w_2))$ defines an arc of $G^{Augment,\ll}$ if (x_1, x_2) is an arc of $G^{\ll}$ that meets (P0) and if w_2 may be written $w_2 = w_1 + W - \mu(j_1, j_2)$, W being a feasible production for periods in $i_1+1, \dots, i_2-1$, in a way that meets (P1, P2, P3) of Theorem 2.
- The cost $Cost^{Augment}((x_1, w_1), (x_2, w_2))$ of such an arc is $\alpha.PCost(W) + \beta.(i_2 - i_1)$, where $PCost(W)$ is the production cost induced by W. In case $x_1 = Source$ or $x_2 = Sink$, then this formula has to be respectively adapted as $\alpha.PCost(W) + \beta.\Delta^{Start}(j_2)$ and $\alpha.PCost(W) + \beta.\Delta^{End}(j_1)$.

Then we get a **SLSS Path Reformulation**: {Compute a shortest path (in the sense of $Cost^{Augment}$) in $G^{Augment,\ll}$ from the *source* node to the *sink* nodes.}

B. An A* Based Algorithm

Above **SLSS** path reformulation suggests us to deal with **SLSS** while relying on a standard path search algorithm for acyclic graphs. However, we must take care on the 2 following specific features of our path model:

1) Every time we are working with some node (i_1, j_1, w_1) of $G^{Augment,\ll}$, we must generate arcs $((i_1, j_1, w_1), (i_2, j_2, w_2))$, together with values $W = w_2 - w_1 + \mu(j_1, j_2)$ and related cost values $PCost(W)$. But computing these cost values requires the resolution of some local lot sizing problem.
2) Depending on the context, values w may be large, possibly infinite. So we need filtering devices.

Computing the Cost Values through a Pre-Process.

We deal with the first issue while performing a dynamic programming pre-process that yields a table TAB, indexed on the pair $i_1, i_2, -1 \le i_1 < i_2 \le N$. For any such a pair, $TAB[i_1, i_2]$ contains a list of 3-uples $(W, PCost(W), Sol)$ where W means the resources produced during periods $i = i_1 + 1, \dots, i_2 - 1$, $PCost(W)$ means their optimal production cost, and Sol some related production schedule. This dynamic programming pre-process is implemented according to a backward driven strategy: It involves a main loop indexed on the 2-uples (i_2, w_2^P) and an internal loop indexed on the 2-uples (i_1, w_1^P) such that $i_1 < i_2$, and works in pseudo-polynomial time. We may speed it by noticing that if $i_2 \le N - 2$, then we may restrict ourselves to pairs $i_1, i_2)$ such that it is possible to compute j_1, j_2 with $i_1 \ge \tau^m(j_1)$ and $i_2 \le \tau M(j_2)$. So, every time we generate a decision value W, we get its cost value $PCost(W)$ through a direct access to the table TAB. We notice that if C^S and C^P are bounded by polynomial functions of N and M then the construction of TAB can be performed in polynomial time, which also means that in the general case, this construction can be performed in pseudo-polynomial time.

Introducing Filtering Devices: A Lower Bound.

As for the second issue, we may reduce the number of nodes explored during our path search by performing some kind of rounding (for instance modulo the first k bits, as usually done in order to design FPTAS algorithms). But the most natural way is to rely on a lower bound procedure, that, with some node (i_1, j_1, w_1) of $G^{Augment,\ll}$ is going to associate a lower bound $LB(i_1, j_1, w_1)$ of a shortest path from (i_1, j_1, w_1) to the *sink* nodes in $G^{Augment,\ll}$. We get such a lower bound by first computing a lower bound $Trans_{j_1}^{Min}$ on the number of transfer transactions that will remain to be performed after period i_1. We see that we may define $Trans_{j_1}^{Min}$ as follows:

$$Trans_{j_1,w_1}^{Min} = \lceil \frac{(H_0^S - w_1 + \sum_{j > j_1} e_j)}{C^{TR}} \rceil,$$

Next we notice that we get a lower bound on the additional time $T\_Add_{j_1,w_1}^{Min}$ that will to be spent by the job scheduler because of those transfer transactions by setting:

- $\tau_{j_1}^{Min} =$ the smallest value $\tau_j, j \ge j_1 + 1$;
- $T\text{-}Add_{j_1,w_1}^{Min} = \tau_{j_1}^{Min} \cdot Trans_{j_1,w_1}^{Min}$.

We deduce a lower bound T_{j_1,w_1}^{Min} on the time that the job scheduler will have to spend after the end of period i_1 before achieving its own process by setting:

$$T_{j_1,w_1}^{Min} = (\sum_{j > j_1} t_j) + T\text{-}Add_{j_1,w_1}^{Min}.$$

By the same way, we check that achieving this process will require during the periods $i_1, \dots, N-1$ the production of at least W_{j_1,w_1}^{Min} resource, where W_{j_1,w_1}^{Min} is defined by:

- $\epsilon_{j_1}^{Min} =$ the smallest value $\epsilon_j, j \ge j_1 + 1$;
- $W_{j_1,w_1}^{Min} = H_0^P + H_0^S + \mu^{End}(j_1) + \epsilon_{j_1}^{Min} \cdot Trans_{j_1,w_1}^{Min} - w_1$.

Then we may retrieve from $TAB[i_1, N]$ a value $PCost_{i_1,j_1,w_1}^{Min}$ equal to the smallest cost value related to some production W such that $W \ge W_{j_1,w_1}^{Min}$;

Lemma 3: $LB(i_1, j_1, w_1) = \beta \cdot T_{j_1,w_1}^{Min} + \alpha.PCost_{i_1,j_1,w_1}^{Min}$ *is a lower bound for the cost of a shortest path from* (i_1, j_1, w_1) *to the sink nodes in* $G^{Augment,\ll}$.

Proof. It derives in a straightforward way from the construction of the quantities T_{j_1,w_1}^{Min} and $PCost_{i_1,j_1,w_1}^{Min}$. □

Exact and Heuristic Dominance Rules.

We enhance our algorithm by introducing dominance rules:

- **Exact Dominance Rule:** If at some time during the search process we deal with two nodes $\sigma_1 = (i_1, j_1, w_1)$ and $\sigma_2 = (i_2, j_2, w_2)$ of $G^{Augment,\ll}$ such that $(i_1, j_1, w_1) = (i_2, j_2, w_2)$ and $Cost^{G\_Augment}(\sigma_1) + LB(\sigma_1) \leq Cost^{G\_Augment}(\sigma_2) + LB(\sigma_2)$, then we may *kill* σ_2.
- **Heuristic Dominance Rule:** We may reinforce above rule by setting that $\sigma_1 = (i_1, j_1, w_1)$ dominates $\sigma_2 = (i_2, j_2, w_2)$ if:
  - $Cost^{G\_Augment}(\sigma_1) + LB(\sigma_1) \leq Cost^{G\_Augment}(\sigma_2) + LB(\sigma_2)$
  - $i_1 \leq i_2$, $j_2 \leq j_1$ and $w_1 \leq w_2$.

  If at some time during the search process we deal with two nodes $\sigma_1 = (i_1, j_1, w_1)$ and $\sigma_2 = (i_2, j_2, w_2)$ of $G^{Augment,\ll}$ such that σ_1 dominates σ_2 , then we *kill* σ_2. Yet this rule is only a heuristic dominance rule. Yet, numerical experiments will show its efficiency.

The Algorithm *A*_SLSS*: So we design our algorithm *A*_SLSS* as an adaptation of the well-known *A** (see [7]) algorithm for path search in very large networks. At any time during the *A*_SLSS* resolution process, we are provided with an expansion list LS of nodes $\sigma = (i, j, w)$of $G^{Augment,\ll}$, given together with related cumulative costs $Cost^{G\_Augment}(\sigma)$ and future cost estimations $LB(\sigma)$. Those nodes are ordered according to increasing $Eval(\sigma) = Cost^{G\_Augment}(\sigma) + LB(\sigma)$ estimation values. Then we pick up the first element $\sigma_1 = (i_1, j_1, w_1)$ in LS, called the *pivot* node, and we expand it: we generate all decisions (i_2, j_2, W) such that $(i_1, j_1) \ll (i_2, j_2)$, (P0) holds and $W \in TAB[i_1, i_2]$ that are valid in the sense that they meet (P1, P2, P3) of Theorem 2. For every such a decision, we generate resulting state $\sigma_2 = (i_2, j_2, w_2)$ and insert it into LS while meeting the dominance rules (depending on the rules, it will yield an exact or heuristic algorithm). We stop when the *pivot* state σ is a *sink* node.

Resulting *A*_SLSS* algorithm may be summarized:

```
Algorithm A*_SLSS
Initialization:
  • $LS = \{source\}$, $sink$ being the source node of $G^{Augment,\ll}$, provided with $Eval(sink) = LB(sink)$ estimation value; $NotStop$;
  • $Curr\_Sol$ = current partial solution set = $Nil$;
While $NotStop$ and $LS \neq Nil$ do
  1) Denote by $Pivot$ the head of $LS$;
  2) Remove it from $LS$and insert it into $Curr\_Sol$;
  3) If $Pivot = (i_1, j_1, w_1)$ is a sink node then $Stop$
     Else
        Generate all valid decisions $Dec = (i_2, j_2, W)$ that apply to $Pivot$;
        For any such a decision $Dec$ do
          • Compute resulting node $\sigma_2 = (i_2, j_2, w_2)$, together with its value $Eval(i_2, j_2, w_2)$;
          • If $\sigma_2$ is dominated by no node $\sigma_0$ in $LS$ then insert it into $LS$, while
            keeping $LS$ from containing a node $\sigma_0$ dominated by $\sigma_2$;
  If $Stop$ then Retrieve from $Pivot$ and $Curr\_Sol$ a full path $\Gamma$ solution of SLSS;
```

Depending on the dominance rules that we apply here, we obtain two algorithms *A*_SLSS* and *Heur_A*_SLSS* .

Theorem 3. *When implemented with the weak dominance rules, above algorithm A*_SLSS solves* **SLSS** *in an exact way. In any case, it works in pseudo-polynomial time.*

Proof. The first part of this statement derives from Theorem 2 in a straightforward way. This algorithm is nothing more than the $A*$ algorithm applied to the state network whose nodes are all states $\sigma = (i, j, w)$, and arcs corresponds to transitions $(\sigma 1 = (i_1, j_1, w_1) \leftarrow (\sigma 2 = (i_2, j_2, w_2))$ according to decisions (i_2, j_2, W). As for the second part, we see that if C^S and C^P are bounded by polynomial functions of N and M then the number of nodes of the graph $G^{Augment,\ll}$ is also bounded by a polynomial function of N and M, while TAB may be computed in pseudo-polynomial time. We conclude. □

VI. NUMERICAL EXPERIMENTS

Purpose: Evaluating the behavior of the A^* algorithm, with respect to **SLSS_MILP**, considered as a benchmark.
Technical Context: Algorithms were implemented in C++ on an Intel i5-9500 CPU at 4.1GHz. CPU times are in seconds. We used the CPLEX20 library for the MILP models.

A. Instance generation

Production and consumption coefficients: In order to mimic what may be a power market, we cluster production periods into $\#SP$ super-periods of same length. Each super-period is assigned symbolic mean production and cost values $\overline{R_{cl}}$, $\overline{Cost_{cl}}$ in $\{Low, Medium, High\}$. Then, integral production and variable cost values are randomly generated for every period according to those mean super-period values. We generate activation costs in such a way that the activation cost represents around a third on the global production cost.
Storage capacities and scaling coefficients α, β: In order to control the relation between the number of transfer transactions and the number of activation decisions, we impose the quotient $\frac{C^P}{C^S}$ to remain inside an interval $[0.5, 3]$. We do in such a way that $\frac{\mu(-1,M)}{C^{TR}}$ evolves like M, so that we may control the number of jobs between transfer transactions (in average close to 5). We do in such a way that the respective weights of the production and the scheduling parts of the global cost remain integral and comparable.
Tables I and II present a package of 12 instances.

B. Outputs

- **Tables III and IV** : This table is devoted to the **SLSS_MILP** model. It provides the objective value obj of **SLSS_MILP**, its linear relaxation $relax$, the number $A - Cuts$, of *No-Antichain* cuts generated during the process, related CPU times CPU.

TABLE I
INSTANCE PARAMETERS TABLE

id	1	2	3	4	5	6
N	20	30	40	60	70	80
M	10	10	15	20	20	25
p	2	3	4	2	3	4
$\#SP$	2	2	3	4	5	5

TABLE II
INSTANCE PARAMETERS TABLE

id	7	8	9	10	11	12
N	100	110	120	140	150	160
M	30	30	35	40	40	40
p	2	3	4	2	3	4
$\#SP$	5	5	5	4	5	5

- **Table V and VI** : This table is devoted to the path search approach. It provides resulting number TT of transfer transactions, together with CPU times $Path - CPU$ induced by the A^* algorithm. It also provides the same values $HF - TT$ and $HF - Path - CPU$, together with the cost value $HF - obj$, obtained by introducing the strong (heuristic) dominance rules of Section V, that means by applying the *Heur_A*_SLSS* Algorithm.

Comments: The MILP model is time consumming, due to the gap induced by the relaxation of its integrality constraints. The A^* algorithm significantly outperforms the MILP model. The heuristic dominance rule of Section V.*B* induces a very small gap with respect to optimality (it reach optimality in 10 among the 12 instances, and the gap for the 2 remaining instances 8 and 12 hardly reaches 2%, while speeding in average the search process by 40%.

VII. FUTURE WORK

Future research will be oriented towards: 1) the **SLSS** extensions that make the parallel machine scheduling decisions be part of the problem; 2) the collaborative issue, when several *job schedulers* interact with the *Lot Sizer* player; 3) the management of uncertainty.

TABLE III
MILP RESOLUTION

id	1	2	3	4	5	6
obj	186	202	238	386	402	456
$A - Cuts$	158	179	326	595	554	818
$relax$	103.6	131.8	147.8	270.8	285.6	321.6
CPU	31.8	20.0	32.6	998	1997	1024

TABLE IV
MILP RESOLUTION

id	7	8	9	10	11	12
obj	558	596	650	684	758	781
$A - Cuts$	357	882	1071	17431	1928	2235
$relax$	376.6	388.2	360.0	452.5	564.0	536.1
CPU	3079	2889	5906	10058	11647	7945

TABLE V
A* ALGORITHM RESOLUTION

id	1	2	3	4	5	6
TT	3	4	3	5	7	6
$Path - CPU$	2.1	3.2	1.9	15.8	57.4	45.6
$HF - obj$	186	202	238	386	402	456
$HF - TT$	3	4	3	5	7	6
$HF - Path - CPU$	1.8	2.6	1.5	10.9	43.5	25.6

TABLE VI
A* ALGORITHM RESOLUTION

id	7	8	9	10	11	12
TT	5	8	6	9	7	10
$Path - CPU$	38.3	148.7	61.8	598.9	265.6	677.7
$HF - obj$	558	608	650	684	758	802
$HF - TT$	5	9	6	9	7	10
$HF - Path - CPU$	28.1	83.2	42.6	395.8	137.4	400.6

ACKNOWLEDGMENT

Present work was funded by French ANR: National Agency for Research, Labex IMOBS3, and PGMO Program.

REFERENCES

[1] F. Bendali, J. Mailfert, E. Mole-Kamga, A. Quilliot, H. Toussaint, *Pipelining dynamic programming process in order to synchronize energy production and consumption*, Proc. 2020 FEDCSIS WCO Conf., p 303-306, 2020. doi.org/10.15439/978-83-955416-7-4.

[2] Biel K., Glock C. H.: Systematic literature review of decision support models for energy-efficient production planning. *Computers and Industrial Engineering*, 101, pp. 243-259, (2016). https://doi.org/10.1016/j.cie.2016.08.021.

[3] Clark A. Almada-Lobo B., Almeder C.: , J.: Lot sizing and scheduling: Industrial extensions and research opportunities. *International Journal of Production Research* 49-9, p 2457-2461 (2011). https://doi.org/10.1080/00207543.2010.532908.

[4] Erdelic T., Caric T., Lalla-Ruiz E.: A Survey on the Electric Vehicle Routing Problem: Variants and Solution Approaches.*Journal of Advanced Transportation* Volume 2019. Article ID 5075671; https://doi.org/10.1155/2019/5075671.

[5] S. Fidanova, O. Roeva, M. Ganzha, *Ant colony optimization algorithm for fuzzy transport modelling*, Proc. 2020 FEDCSIS WCO Conference, p 237-240, 2020. doi.org/10.15439/978-83-955416-7-4

[6] Franck A.: On chain and antichain families of a partially ordered set. *Journal of Combinatorial Theory. Series B* 29-2, p 176-184 (1980). https://doi.org/10.1016/0095-8956(80)90079-9.

[7] Hart P. E., Nilsson N. J., Bertram R.: A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4-2, p 100-107 (1968). https://doi.org/10.1109/TSSC.1968.300136.

[8] Irani S, Pruhs K.: Algorithmic problems in power management. *ACM SIGACT News* 36-2, p 63-76 (2005). DOI:10.1145/1067309.1067324.

[9] K. Stoilova, T. Stoilov, *Bi-level optimization application for urban traffic management*, Proc. 2020 FEDCSIS WCO Conf., p 327-336, 2020. doi.org/10.15439/978-83-949419-5-6.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 139–149 DOI: 10.15439/2024F8851 ISSN 2300-5963 ACSIS, Vol. 41

Converting German Historical Legal Documents to TEI XML including challenges with Table Extraction

Thomas Reiser* , Petra Steiner†
* University of Koblenz, Germany,
Email: treiser@uni-koblenz.de
† Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Email: steiner@bibb.de

Abstract—**The occupations archive at the Federal Institute for Vocational Education and Training contains thousands of historical German Vocational Education and Training (VET) and Continuing Vocational Education and Training (CVET) regulations from the last 100 years. However, these are hardly accessible because they are currently only available in their original paper form. We present a workflow that transcribes images of these regulations into the TEI XML format which preserves the logical document structure and stores metadata. This paper addresses issues caused by poor page segmentation of the applied optical character recognition (OCR) methods and presents rules that can reconstruct a large part of the documents' hierarchy. A straightforward table recognition method for tables with borders is presented, as well as a metadata extraction procedure for the selected data set. While our approach is generic and functional, further research is necessary to develop a fully automated and more robust workflow.**

I. INTRODUCTION

IT IS evident that legal texts constitute an indispensable component of labor market research. The occupations archive at the Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung, BIBB) encompasses a multitude of historical regulations pertaining to vocational education and training (VET) and continuing vocational education and training (CVET). The regulations in the occupations archive are from the 1920s, the Third Reich, German Democratic Republic, and also the Federal Republic Germany. However, these documents are currently accessible to only a select few individuals with access to this archive. The objective is to digitize the archive by generating transcripts of the documents in the Text Encoding Initiative (TEI) XML format. This format is capable of capturing all logical text elements and layout elements such as page beginnings, footnotes, page headers, and line breaks.

A feasibility analysis and a preliminary draft of a pipeline for this process have been developed using a data set that is already available as digital images. The selected data set comprises 600 VET and 383 CVET regulations from 1969 to 2022. As the Vocational Training Act (Berufsbildungsgesetz, BBiG) of 1969 established a framework for the majority of these regulations and they are all published in the Federal Gazette (Bundesanzeiger), these documents are open to the public and allow for a well-structured rule-based approach. Once the documents in the archive are available as digital images, we aim to optimize the transcription process for them by applying state of the art layout analysis and structure recognition methods.

In this paper, our research questions are:

1) How can VET and CVET regulations from 1969 to 2022 that have been published in the Federal Gazette be digitized into fully structured TEI XML documents?
2) How are the selected documents structured?
3) How can errors in these documents be detected when there is hardly any ground truth available?

This paper presents the results of the feasibility analysis and is divided into five sections. The first section provides an introduction into related works about document digitization efforts and the selected document collection, consisting of over 900 German training regulations. The second section presents a brief overview of the state of the art in the generation of text hierarchy from unstructured texts, including layout analysis, OCR, and classification models for text hierarchy recognition. The third section outlines the methodological background and the utilized pipeline, including scan preprocessing methods, OCR, postprocessing steps from the layout analysis, recognition of different text elements such as lists, headings, etc., the transcription into the target data format, TEI XML, and table recognition. The fourth section presents the experimental results and evaluation of this novel approach by comparing extracted metadata to the small existing ground truth and analyzing properties of the generated transcripts. The last section presents the conclusions and an outlook for further research on documents of the occupations archive, once its regulations are available as digital images.

A. *Motivation*

The digitization of historical documents has gained particular interest in recent years, with numerous approaches emerging to address this task. Optical character recognition (OCR) represents a foundational technology for digitization, with a significant research focus and a range of established tools. This paper describes a basic transcription pipeline for German

©2024, PTI

training regulations that have been published in the Federal Gazette, to generate fully structured TEI XML documents, maintaining the logical document structure and text hierarchy of the digitized documents.

There are numerous methodologies to address the task of document digitization. One basic approach is the detection of the entire text within the document images, as exemplified by the methodology employed in the case of the Finnish newspaper digitization project[1], where the objective is to generate an ALTO XML document that contains all recognized text, utilizing the Tesseract OCR engine. An alternative approach is shown in [2], where the authors model the text structure of legal texts in Austria and align the recognized text to this predefined structure, thereby improving the structured recognition of text.

More advanced methods employ the OCR results to construct structured data from the text images. The authors of [3] use OCR to digitize invoice papers and to structure the recognized information, such as product description, quantity, and price. Similarly, in the study by [4], the names of judges at German federal courts from 1950 to 2019 were extracted from the Federal Gazette by applying OCR to these publications.

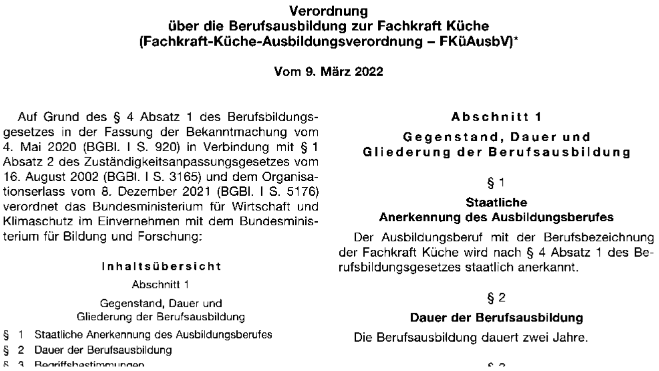
Not only table-structured data, but also graph data can be extracted from text images. T2KG, an NLP tool that can construct knowledge graphs from text, is presented in [5]. Another way to generate knowledge graphs from text using Open Information Extraction is introduced in [6]. Other approaches include contextual extraction and representation methods based on knowledge graphs, ontologies and taxonomies, see [7], [8], [9], [10], [11], [12].

For better accessibility, some OCR workflows are embedded into web applications. One of the largest efforts for this is OCR4all [13] which allows the usage of different preprocessing steps, segmentation methods, and OCR models. It also enables the interaction with each of the process steps so users can do corrections at intermediate results to improve the overall outcome. However, there are also less extensive tools which allow the management of digitized document collections, for example in [14], [15]

B. Data Set

The data set comprises 600 VET and 383 CVET regulations from 1969 to 2022. All of these regulations are published in the Federal Gazette and follow a similar layout structure. Each regulation begins with a short preamble, which is sometimes followed by a table of contents. A regulation page that illustrates most of the layout elements described here is shown in Figure 1.

Every regulation is comprised of multiple paragraphs. These paragraphs commence with a section sign (§) and are followed by the section number. One line below, the section headline is displayed. The paragraphs are further structured into sections. These sections can be either a single block of text or a set of enumerated sections, with the section number between parentheses ((1), (2),...).



Verordnung
über die Berufsausbildung zur Fachkraft Küche
(Fachkraft-Küche-Ausbildungsverordnung – FKüAusbV)*

Vom 9. März 2022

Auf Grund des § 4 Absatz 1 des Berufsbildungsgesetzes in der Fassung der Bekanntmachung vom 4. Mai 2020 (BGBl. I S. 920) in Verbindung mit § 1 Absatz 2 des Zuständigkeitsanpassungsgesetzes vom 16. August 2002 (BGBl. I S. 3165) und dem Organisationserlass vom 8. Dezember 2021 (BGBl. I S. 5176) verordnet das Bundesministerium für Wirtschaft und Klimaschutz im Einvernehmen mit dem Bundesministerium für Bildung und Forschung:

Inhaltsübersicht

Abschnitt 1

Gegenstand, Dauer und
Gliederung der Berufsausbildung

§ 1 Staatliche Anerkennung des Ausbildungsberufes
§ 2 Dauer der Berufsausbildung
§ 3 Begriffsbestimmungen

Abschnitt 1

Gegenstand, Dauer und
Gliederung der Berufsausbildung

§ 1

Staatliche
Anerkennung des Ausbildungsberufes

Der Ausbildungsberuf mit der Berufsbezeichnung der Fachkraft Küche wird nach § 4 Absatz 1 des Berufsbildungsgesetzes staatlich anerkannt.

§ 2

Dauer der Berufsausbildung

Die Berufsausbildung dauert zwei Jahre.

Fig. 1: Excerpt of the first page of the VET regulation for kitchen qualified professionals from 2022.

These enumerations are hierarchically structured, with each element capable of containing a further enumeration (a), b), ...), which itself can be further nested (aa), bb), ...). In some regulations, the paragraphs are grouped in sections (1. Abschnitt or Abschnitt 1) or parts (Erster Teil or Teil 1). There are also a few exceptions where the sections are grouped in parts. The following combinations of the aforementioned structure elements are possible:

1) paragraphs
2) paragraphs with table of contents
3) paragraphs in sections
4) paragraphs in sections with table of contents
5) paragraphs in parts
6) paragraphs in parts with table of contents
7) paragraphs in sections in parts with table of contents

II. RELATED LITERATURE

In order to get a large digitized text corpus, many approaches can be considered. Structuring the documents takes text digitization a step further because not only is text recognized, but also logical units, such as paragraphs, section, listings, tables, and figures need to be detected to make the output more meaningful.

While we plan to develop a more advanced approach for documents of the occupations archive, we conducted a feasibility and usability analysis of such a pipeline a first subset of VET and CVET regulations that have been openly available in the Federal Gazette. The results of these are discussed in this paper.

In first approaches for document structuring from the 2000s, human knowledge about the document has been used to define text- and layout-based rules to extract the text structure [16], [17], [18]. The first version of our digitization workflow also uses predefined rules, but we aim to recognize patterns using layout and text features automatically in future research to structure texts with different layouts. Similar to [2], a certain structure is defined to recreate the text hierarchy. There are a few different possible structures consisting of the same structure elements which are also described in subsection I-B.

Although some standard conversion tools, like Vertopal, to convert files that contain text into markup languages exist [19], they assume that the text in the documents that need to be converted contain correct structural information, which cannot be guaranteed by default OCR models. While they can be used to create files in HTML or TEI XML, the output files often do not represent the text hierarchy or logical document structure but are rather another representation of recognized text areas, lines, and text.

For metadata extraction, machine learning models like they are used in GROBID [20] can be used to extract metadata like title and authors. It also allows the recognition of references and citations, and the detection of the abstract. Despite it having a lot of useful features, it was trained on scientific articles and performs especially well on them. To use GROBID on legal texts like the training regulations in this article, it would be necessary to fine tune the model, for which training data would be required. Besides this, it again depends on good page segmentation and reliable text recognition.

While some other tools like PdfPig[1] or PDFMiner[2] also allow layout analysis from words in a PDF file, they are based on heuristics that do not match all document layouts, and again, depend on previously recognized words.

In a more recent publication [21], a tree structured document hierarchy is generated using the HELD model where text areas are inserted at a specific level in the tree hierarchy depending on the output of a binary classification function. This function compares features of the element that needs to be inserted and elements that have already been classified, e.g., format features or consecutive numbers in this text level. While this is a promising approach, it again requires training data and is not straightforward to implement. Besides this, it again depends on correct page segmentation which is at the moment not guaranteed. Because, in this article, we focus on a smaller feasibility analysis and the classification of elements is done by iterating over each element, it is easier and more efficient to use predefined rules, as mentioned above. This approach may become more relevant on the target data set of the occupations archive.

Our method aims at using out of the box OCR models and preprocessing techniques with rule based postprocessing. While OCR-D [22] is a tool that combines preprocessing with OCR, it was not fast enough on the available hardware to be efficiently used within the given time constraints for our use case. Because it has shown comparable results to other commerical alternatives in [23], [24], the Tesseract OCR engine has been selected for this work.

Although Tesseract delivers state of the art OCR results, it is still no one fits all solution. As mentioned before, many tools depend on correct layout analysis. To improve the results of layout analysis, many researchers currently employ object detection methods to pretrain models such as Faster R-CNN, YOLO, etc. to recognize text regions and classifying them into heading, list item, title, and more [25], [26], [27], [28].

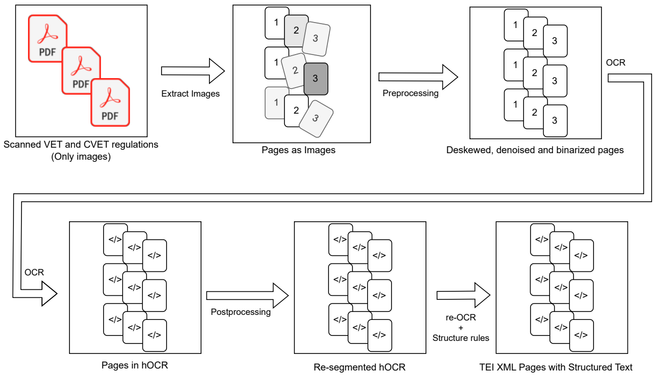


Fig. 2: Overview of the presented pipeline.

Similar to the layout analysis, also tables are recognized in a similar manner, using deep learning object detection methods [29], [30]. However, again, due to the lack of labeled training data for this data set, these are not applied in the feasibility analysis in this paper, but will likely be used on the regulations in the occupation archive.

TEI XML has been selected as the target data format because it is recommended by the German Research Foundation (DFG) as a good standard for long time archives of documents [31]. While PDF is a proprietary standard that is stored in binary files, markup languages like XML can be read by almost any computer without installing additional software that can read PDF files. Besides that, these files can be efficiently stored in XML databases like eXist-db [32] to manage the document collection. eXist-db also allows the addition of plugins, such as a versioning plugin which allows memory-efficient saving of different versions of the same documents and a fast restore, if an older version is required. Especially in this automated setting, where errors need to be expected, this can be very helpful. Another useful plugin is TEI Publisher [33] which allows a user friendly management of the XML database as well as viewing the documents in a well human-readable ways and editing uploaded TEI XML files. Besides this, it allows viewing the text along with the original images which can facilitate the error correction process. Therefore, TEI XML has been found to be a suitable choice for a target data format when digitizing larger text corpora.

III. METHODOLOGY

In this section, the developed pipeline is presented step by step. An illustration is depicted in Figure 2. The pipeline starts with a list of PDF or image files. If PDF files are given, the pages are extracted as images at 300 dots per inch (DPI) which is considered to be a good trade-off between image quality and storage efficiency [31].

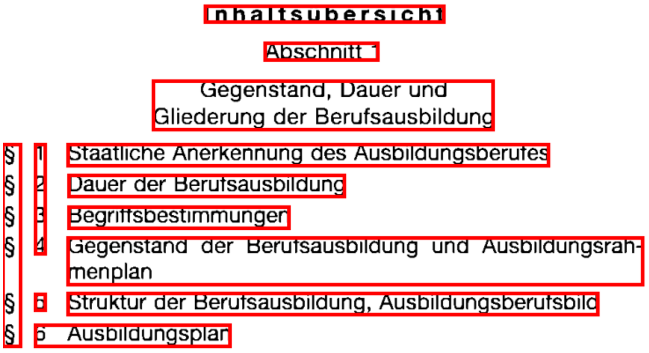
A. Preprocessing

A common step when it comes to document digitization with OCR is image preprocessing [34]. The scanned pages

---

[1] https://github.com/UglyToad/PdfPig

[2] https://github.com/pdfminer/pdfminer.six



Inhaltsübersicht

Abschnitt 1

Gegenstand, Dauer und
Gliederung der Berufsausbildung

§ 1 Staatliche Anerkennung des Ausbildungsberufes
§ 2 Dauer der Berufsausbildung
§ 3 Begriffsbestimmungen
§ 4 Gegenstand der Berufsausbildung und Ausbildungsrahmenplan
§ 5 Struktur der Berufsausbildung, Ausbildungsberufsbild
§ 6 Ausbildungsplan

Fig. 3: Bounding boxes of lines recognized by Tesseract in the table of contents of the VET regulation for kitchen qualified professionals in 2022.

can be skewed, contain noise such as water damages or ink stains, white margins from the paper, and are usually in color. For most OCR engines, a clean black and white image with no noise and straight pages is required. To achieve this from the original images, scantailor[3] is used, an open-source tool that supports many different preprocessing steps, including image binarization, page deskewing, content detection, and noise removal. These steps are applied and afterwards, OCR is executed.

B. OCR

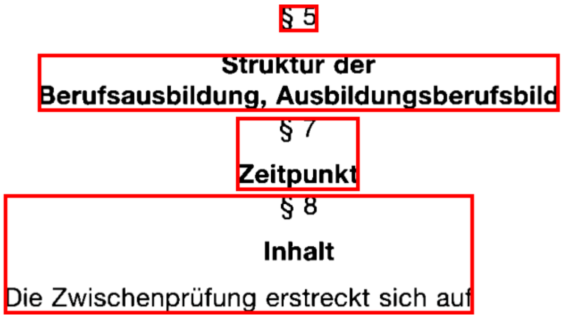
The character recognition is done by an OCR engine. One of the most frequent used OCR engines in literature is Tesseract which is open-source and can in some cases give comparable results to commercial alternatives like ABBYY FineReader [23], [24]. It is able to recognize most text areas and characters correctly, but sometimes has trouble with text that is aligned on the same width across multiple lines as shown in Figure 3.

Another problem is the uneven indentation of text in the document. As shown in Figure 4, Tesseract paritally recognizes the paragraph headlines as paragraph number and headline in two text areas, one text area, or even including a part of the paragraph's text.

To fix these errors, a postprocessing step is introduced that fixes the most frequent OCR errors that have been found by investigating the results by hand.

C. Post-Correction

To solve the aforementioned issues, conditions on when and how to restructure bounding boxes are introduced. Recognized text areas are defined by bounding boxes (bboxes), consisting of left, upper, right, and lower border. For the issue shown in Figure 3, a threshold was defined and all text areas on the same height are found and sorted ascending by their right border. Afterwards, if the distance between the left border of a text element in one of the sorted clusters to the right border of its predecessor is smaller than the predefined threshold or



§ 5

Struktur der
Berufsausbildung, Ausbildungsberufsbild

§ 7

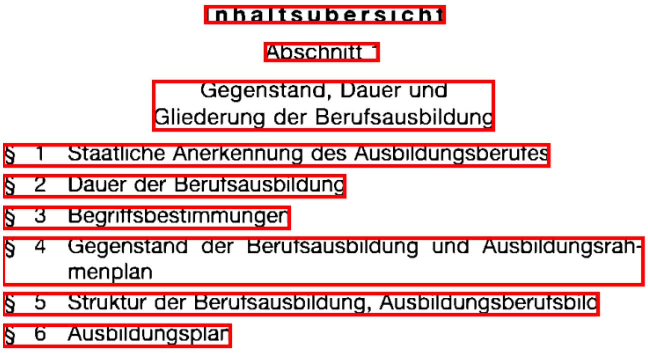
Zeitpunkt

§ 8

Inhalt

Die Zwischenprüfung erstreckt sich auf

Fig. 4: For each of the three paragraph headlines, Tesseract recognized the text area in a different way.



Inhaltsübersicht

Abschnitt 1

Gegenstand, Dauer und
Gliederung der Berufsausbildung

§ 1 Staatliche Anerkennung des Ausbildungsberufes
§ 2 Dauer der Berufsausbildung
§ 3 Begriffsbestimmungen
§ 4 Gegenstand der Berufsausbildung und Ausbildungsrahmenplan
§ 5 Struktur der Berufsausbildung, Ausbildungsberufsbild
§ 6 Ausbildungsplan

Fig. 5: Bounding boxes of lines after correcting the errors of the Tesseract output shown in Figure 3.

the elements overlap, they need to be merged. Once all areas that need to be merged have been identified, the area with the furthest right border in each of these clusters is expanded to the border that is the furthest left in the cluster. After all merges have been completed, each element that has been used to expand another text area, is removed. The result is shown in Figure 5.

For the inconsistent assignment of lines to text areas as shown in Figure 4, a measure is defined for when to split text areas. First, the line distances for subsequent lines in each text area are computed. These distances for an example page are plotted in Figure 6.

As shown in this example, there seem to be two mass centers of line distances: One for lines that belong into the same text area, another for lines that should be within two different text areas. A common way to detect a fixed number of clusters is k-Means Clustering [35]. Setting $k = 2$ will detect both clusters in the line distances. If the line distance is larger than the largest value of the cluster with the lower centroid, marked by the red line in the plot, the text area is split between the two respective text lines. The result will cause all paragraph headlines to be split in two boxes, *§ paragraph number* and *paragraph headline*. The resulting bboxes of the text areas are now as shown in Figure 7. The text areas now only contain coherent text segments. This

---

[3] https://github.com/trufanov-nok/scantailor-universal

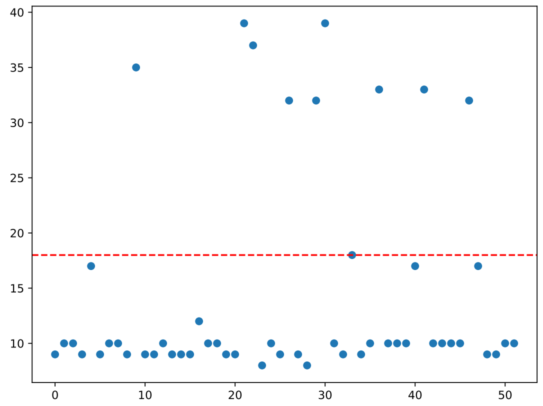


Fig. 6: Line distances within text areas of the second page of the VET regulation for kitchen qualified professionals in 2022.

procedure is applied for each document and formalized as follows:

Let

- c be the set of text areas that have been recognized by Tesseract. Here, a text area is defined by its bounding box and the list of lines in this area.
- l_{c_i} be the list of detected lines in the text area $c_i \in c$. A line is defined by its bounding box and the list of words in it. Words are defined by their bounding box and text.
- $l_{c_i}^j$ be the j-th element of the list of lines l_{c_i} that is ordered by the y_1-coordinate of the bounding box of each line, i.e., its upper border.
- $bbox(l)$ be the bounding box of a line l, defined by x_1, y_1, x_2, y_2 where x_1 defines the left border, y_1 the upper border, x_2 the right border, and y_2 the lower border of the line.
- $d = \bigcup_{c_i \in c} \left\{ bbox(l_{c_i}^j)_{y_1} - bbox(l_{c_i}^{j-1})_{y_2} \,\middle|\, j \in \{1 \ldots |l_{c_i}|\} \right\}$ be the set of line distances between lines that are in the same text area.
- $kmeans_k(D) = \{D_1, ..., D_k\}$ be the result of k-Means clustering where k determines the number of clusters to detect and $D_i \subseteq D$ are the detected clusters.

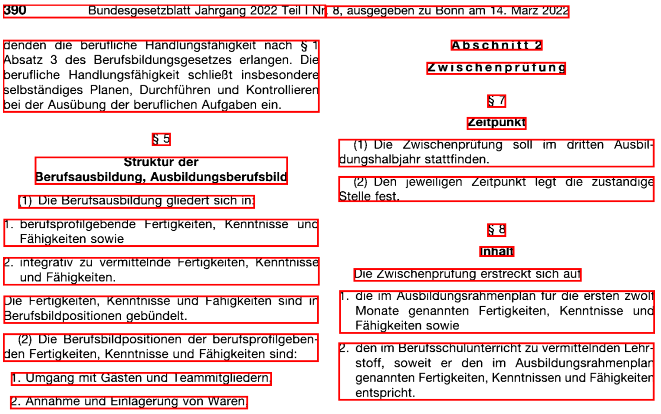
The threshold to split text areas for this document is then calculated by:

$$split\_thresh = max \left\{ argmin_{d_i \in kmeans(d)} \left\{ min\left(d_i\right) \right\} \right\}$$

After these postprocessing steps have been applied to improve text segmentation, OCR is applied a second time for each resegmented text area to resolve errors that have been caused by bad segmentation.

D. Metadata Extraction

The texts in the selected corpus contain some metadata about themselves: title, release date, place of publication, and



390 Bundesgesetzblatt Jahrgang 2022 Teil I Nr. 8, ausgegeben zu Bonn am 14. März 2022

denden die berufliche Handlungsfähigkeit nach § 1 Absatz 3 des Berufsbildungsgesetzes erlangen. Die berufliche Handlungsfähigkeit schließt insbesondere selbständiges Planen, Durchführen und Kontrollieren bei der Ausübung der beruflichen Aufgaben ein.

§ 5

Struktur der Berufsausbildung, Ausbildungsberufsbild

(1) Die Berufsausbildung gliedert sich in:

1. berufsprofilgebende Fertigkeiten, Kenntnisse und Fähigkeiten sowie
2. integrativ zu vermittelnde Fertigkeiten, Kenntnisse und Fähigkeiten.

Die Fertigkeiten, Kenntnisse und Fähigkeiten sind in Berufsbildpositionen gebündelt.

(2) Die Berufsbildpositionen der berufsprofilgebenden Fertigkeiten, Kenntnisse und Fähigkeiten sind:

1. Umgang mit Gästen und Teammitgliedern
2. Annahme und Einlagerung von Waren

Abschnitt 2

Zwischenprüfung

§ 7

Zeitpunkt

(1) Die Zwischenprüfung soll im dritten Ausbildungshalbjahr stattfinden.

(2) Den jeweiligen Zeitpunkt legt die zuständige Stelle fest.

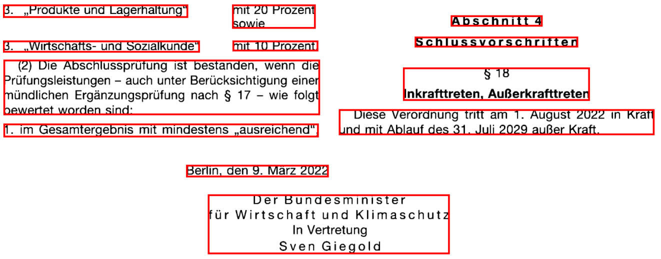
§ 8

Inhalt

Die Zwischenprüfung erstreckt sich auf

1. die im Ausbildungsrahmenplan für die ersten zwölf Monate genannten Fertigkeiten, Kenntnisse und Fähigkeiten sowie
2. den im Berufsschulunterricht zu vermittelnden Lehrstoff, soweit er den im Ausbildungsrahmenplan genannten Fertigkeiten, Kenntnissen und Fähigkeiten entspricht.

Fig. 7: The bounding boxes of Figure 4 after splitting the text areas according to the described procedure. Each text area now contains a single text element.



3. „Produkte und Lagerhaltung“ mit 20 Prozent sowie

3. „Wirtschafts- und Sozialkunde“ mit 10 Prozent.

(2) Die Abschlussprüfung ist bestanden, wenn die Prüfungsleistungen – auch unter Berücksichtigung einer mündlichen Ergänzungsprüfung nach § 17 – wie folgt bewertet worden sind:

1. im Gesamtergebnis mit mindestens „ausreichend“,

Abschnitt 4

Schlussvorschriften

§ 18

Inkrafttreten, Außerkrafttreten

Diese Verordnung tritt am 1. August 2022 in Kraft und mit Ablauf des 31. Juli 2029 außer Kraft.

Berlin, den 9. März 2022

Der Bundesminister
für Wirtschaft und Klimaschutz
In Vertretung
Sven Giegold

Fig. 8: All regulation bodies end with a centered text area containing the responsible federal minister and his ministry.

responsible federal ministry and minister. Since the layout is the same for all of the selected texts, the corresponding elements can be detected by patterns and the gathered information can be stored. For TEI XML documents, all metadata is stored in a `fileDesc` within a `teiHeader` element. The first element giving information about the document itself is the title which is the first element after the page header. In the OCR result, the title is extracted and encoded in a `title` tag. For the author, the last element of the body text is considered. This element contains the responsible federal ministry and minister. An example is shown in Figure 8. To extract the name of the federal minister, his ministry and, in some cases, his substitute, again, a regular expression is used: `Der\s*Bundesminis[ftl]er\s*(.*?)(?: In\s*Vertretung|$)` for male and `Die\s*Bundesminis[ftl]erin\s*(.*?)(?: In\s*Vertretung|$)` for female ministers. `title` and `author` are stored in the `titleStmt`, with the `author` containing `persName` for his name and `orgName` for his organization's name.

Finally, the element above the author text area contains the city and publication date, as also seen in Figure 8. The text is split at the comma to separate city from date and both information is stored in the target document as `pubPlace` and `date` within the `publicationStmt` of the `fileDesc`. As these elements are not relevant for the document content, they are no longer considered after being encoded in the metainformation part.

The `teiHeader` for the VET regulation for kitchen qualified professionals from 2022 is shown here:

```
<TEI version="3.3.0"
 ↪  xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Verordnung
         <lb/>über die
          ↪  Berufsausbildung zur
          ↪  Fachkraft Küche
         <lb/>(Fachkraft- Küche-
          ↪  Ausbildungsverordnung
          ↪  - FKüAusbV)*
        </title>
    <author>
     <persName>Sven
      ↪  Giegold</persName>
     <orgName>Bundesministerium für
      ↪  Wirtschaft und
      ↪  Klimaschutz</orgName>
    </author>
   </titleStmt>
   <publicationStmt>
    <publPlace>Berlin</publPlace>
    <date>9. März 2022</date>
   </publicationStmt>
  </fileDesc>
 </teiHeader>
 ...
</TEI>
```

E. Transcription to TEI XML

The regulations always have a page header which should be represented in the encoded document but not part of the text. To recognize these headers, the uppermost element and all elements on the same height are selected and their text is concatenated. Because the text is in a two column layout, Tesseract is sometimes not able to recognize the page header as a single line and splits it to fit into the two column layout. Therefore, all elements on the same height as the uppermost element on a page are also considered to be part of the header.

As target format for the image transcripts, TEI XML version P5 [36] has been selected which is the latest version at the time this article was written. For the transcription into fully structured TEI XML files, text based rules, i.e., regular expressions, have been defined in order to classify text elements that start a new paragraph, section or part. These rules are:

- parts (*Teil N* or *N-ter Teil*): `^Teil\s*\d\s*$` or `^\b\w+er\b\s*Teil\s*$`
- sections (*Abschnitt N* or *N. Abschnitt*): `^Abschnitt\s*\d+$` or `^\d+\.\s*Abschnitt\s*$`
- paragraphs (*§ N Headline*): `^(§|5|8|S|s|&amp;|\$)*\s*\d+\s*$`

Text within a paragraph is also structured in a given hierarchy, the patterns for these categories are:

1) (1) Enumerated section: `^[\(\}\]]\d+[\)\}\]]`
2) 1. First level enumeration: `^\d+\s?[\.,]\s?.*`
3) a) Second level enumeration: `^[a-z]\s*[\]\)\}]\s`
4) aa) Third level enumeration: `^([a-z])\1{1}\s*[\]\)\}]\s`
5) aaa) Fourth level enumeration: `^([a-z])\1{2}\s*[\]\)\}]\s`
6) aaaa) Fifth level enumeration: `^([a-z])\1{3}\s*[\]\)\}]\s`
7) aaaa) Sixth level enumeration: `^([a-z])\1{4}\s*[\]\)\}]\s`
8) Everything that does not match any of the specified patterns is considered to be a non-enumerated paragraph section.

All patterns are designed in a way that there is space for errors in the text recognition. If none of the mentioned patterns matches the recognized text, the text is considered as a non-enumerated paragraph section. The classification with these regular expressions is used to recognize headlines, paragraphs, and list items on different levels to recreate the original text hierarchy.

An excerpt of the generated TEI XML document of the VET regulation for kitchen qualified professionals from 2022 is shown here:

```
<TEI version="3.3.0"
 ↪  xmlns="http://www.tei-c.org/ns/1.0">
 ...
  <div n="1" type="abschnitt">
   <head>Abschnitt 1<lb/>Gegenstand,
    ↪  Dauer und<lb/>Gliederung der
    ↪  Berufsausbildung</head>
   <div n="1" type="paragraph">
    <head>8 1<lb/>Staatliche</head>
    <div type="section">
     <p>Anerkennung des
      ↪  Ausbildungsberufes</p>
    </div>
    <div type="section">
     <p>Der Ausbildungsberuf mit der
      ↪  Berufsbezeichnung<lb/>der
      ↪  Fachkraft Küche wird nach 8
      ↪  4 Absatz 1 des
      ↪  Be-<lb/>rufsbildungsgesetzes
      ↪  staatlich anerkannt.</p>
    </div>
```

```
      </div>
      <div n="2" type="paragraph">
        <head>82<lb/>Dauer der
          ↪ Berufsausbildung</head>
        <div type="section">
          <p>Die Berufsausbildung dauert
            ↪ zwei Jahre.</p>
        </div>
      </div>
        ...
    </div>
  ...
</TEI>
```

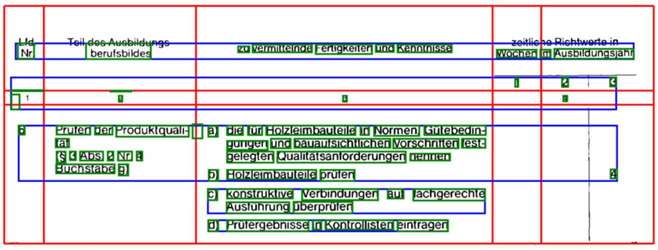
F. Table Recognition

While the appendix of VET regulations from the late 1970s and later consists mostly of the tabular training schedules, in most of the CVET regulations' appendices, example certificates are included. This part focuses on encoding the tables in the appendices. All text areas in the appendix that have not been found as part of a table are encoded as a p-tag with the area's text in the appendix without any further structuring.

Table recognition is a common challenge in document digitization with a lot of ongoing research. As all tables in the selected data set have borders which are also recognized by Tesseract, these lines are used to detect table cells. An example on how the lines are recognized is shown by the red lines in Figure 10. If a line has higher width than height, it is considered to be a horizontal line. If its height is larger than its width, it is considered as a vertical line. Vertical lines are sorted ascending by their horizontal position and horizontal lines are sorted ascending by their vertical position. Then, each line in the sorted set of vertical lines is used with its predcessor to get the left and right border for a table box. The same method is used for horizontal lines to get the lower and upper border. In the detected cells, OCR is applied again and the recognized text is used to fill the table content. Starting from the second level enumeration pattern from subsection III-E, the patterns can be reused to recognize enumerations and plain text in the table cells.

Although this is a very simple approach, it mostly works on this data set. There are, however, some issues with this method. Like most of the pipeline, this procedure depends on the OCR results. In some cases, Tesseract was unable to recognize all lines, as seen in Figure 9, or only fractions of them.

To get the entirety of vertical lines that are only detected in fractions, all vertical lines are expanded to the lowermost and uppermost border of all lines. Horizontal lines are expanded to the left- and rightmost borders of all lines. Afterwards, all lines are compared on how close their borders are to each other to remove duplicate lines. Although this ensures that all lines are recognized, this method does not allow combined cells like they are found in the data set and also assumes that the upper, left, right, and lower border of the tables have been identified correctly. The detection of cells that need to be combined remains an open issue for now. Another issue with this procedure that needs to be addressed is caused by lines that are not recognized at all. In this case, two neighboring rows or columns will be considered as one.



Lfd. Nr	Teil des Ausbildungsberufsbildes	zu vermittelnde Fertigkeiten und Kenntnisse	zeitliche Richtwerte in Wochen im Ausbildungsjahr		
			1	2	3
1	2	3		4	
6	Prüfen der Produktqualität (§ 3 Abs. 2 Nr. 4 Buchstabe g)	a) die für Holzleimbauteile in Normen, Gütebedingungen und bauaufsichtlichen Vorschriften festgelegten Qualitätsanforderungen nennen b) Holzleimbauteile prüfen c) konstruktive Verbindungen auf fachgerechte Ausführung überprüfen d) Prüfergebnisse in Kontrollisten eintragen			4

Fig. 9: In this case, the line between the last two table columns has not been detected at all.

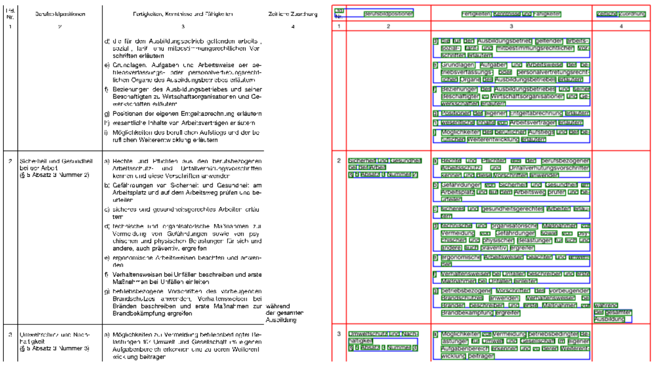


Fig. 10: The table lines and text on the left side are recognized as shown on the right side. Two of the horizontal table lines are detected longer than they actually are.

Once we digitize the entire occupations archive, we will use more advanced table recognition methods because the older regulations, especially from the German Democratic Republic, contain borderless tables with complex structure and many empty cells.

IV. EVALUATION

As there is hardly any ground truth available for a proper evaluation, a usability analysis as well as a small corpus analysis is conducted in order to find anomalies and weaknesses of the proposed procedure.

The web service used for document collection offers some metadata such as title, number of pages, and release year. To test the validity of the extracted metadata, extracted titles have been compared to the titles given in the original data. To allow some OCR errors when comparing the headlines, a maximum Levenshtein distance of eight was allowed between two headlines to be considered from the same regulation. Because there are many long titles, this distance allowed a reasonable amount of differences between the actual and detected titles.

For VET regulations, only three of the 600 regulation titles could not be mapped. This was due to the selected Levenshtein distance being too small or bad text segmentation by Tesseract.

With similar errors, six of the 383 CVET regulations could not be mapped to any of the titles in the ground truth. However, another error where the expected layout was not properly recognized by Tesseract was found such that the preamble text was considered to be the title.

Because the release years are also given in the ground truth, the regulations per year in the ground truth and the digitized collection are counted and compared. The results are shown in Figure 11. Dates that have been recognized by the pipeline but do not lie in the specified interval between 1969 and 2022 have been discarded. Additionally, in some documents, no release date was recognized at all. These two circumstances result in less digitized than original regulations that are considered in the plot.

Although there are some years with as many regulations in the digitized data set as in the ground truth, in most years, there are more documents in the ground truth. This issue has been caused by damages files that had metadata but could not be downloaded because the files were damaged. As shown in Figure 11b, in 2012, there have been more documents in the digitized data set than in the ground truth as a result of bad OCR. Due to less damaged files in the CVET regulations, there are more matches to the ground truth than for the VET regulations.

In subsection I-B, the different layout types have been presented. How these layout types are distributed among the VET regulations is shown in Figure 12.

As seen in Figure 12a, most VET regulations consist only of paragraphs that are not part of sections (Abschnitt) or parts (Teil). Also, only 12.6% of all regulations have a table of contents. All VET regulations with a table of contents consist of parts or sections which contain the paragraphs. Parts mostly appear in regulations for an industrial sector and regulations that cover multiple professions, e.g., the regulation on vocational training in the laboratory field of chemistry, biology and coatings (Verordnung über die Berufsausbildung im Laborbereich Chemie, Biologie und Lack), or in the regulation on vocational training in the weaving industry (Verordnung über die Berufsausbildung in der Weberei-Industrie). It becomes visible that some of the layout types can be found in only few regulations.

Like the VET regulations, most CVET regulations only consist of paragraphs. However, the second most relevant text structure type, consisting of sections and paragraphs without a table of contents, makes up almost a third of all CVET regulations. Less than 5% of all CVET regulations have a table of contents. Any other layout type is hardly present in the CVET regulations. The entire layout type distribution can be seen in Figure 12b.

For an insight how many parts, sections, and paragraphs most regulations consist of and to detect anomalies, bar charts with the number of them and how frequent this number occurs have been created. The number of parts for CVET regulations is shown in Figure 13.

There are two noticeable regulations with only one part. This shows one of the weaknesses with a rule-based approach that relies on text: In these regulations, a text area that contains only the line *written part* (schriftlicher Teil) exists. This matches the regular expression that matches patterns of n-th part (Erster Teil, ...) and causes the workflow to recognize it as a headline for a part although the regulations themselves contain no parts.

A similar anomaly can be seen when counting sections (Abschnitt) in the VET regulations, as seen in Figure 14.

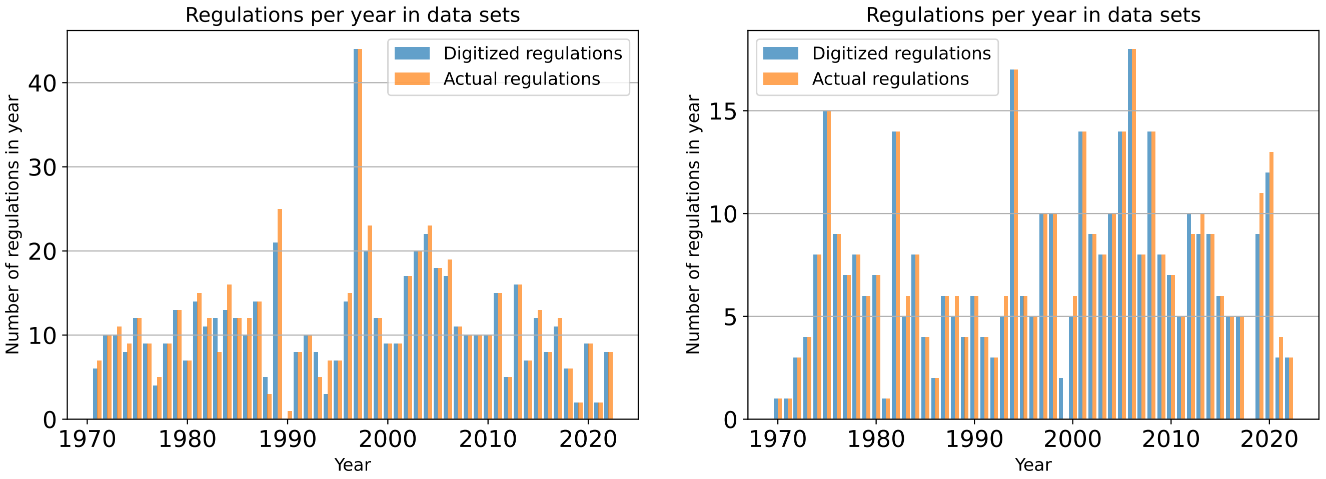
There are three regulations with only one section. In these three cases, however, only one section has been properly detected. This either hints at bad OCR results or a bug in the pipeline implementation that needs to be fixed.

V. CONCLUSIONS AND OUTLOOK

In this article we proposed a basic pipeline that takes images as input, applies image preprocessing, OCR, and postcorrection to process the OCR output into fully structured TEI XML. These transcripts have been used for a first insight into the digitized corpus of 600 VET and 383 CVET regulations that have been published in the Federal Gazette from 1969 to 2022 to detect anomalies which helped identifying weaknesses and bugs in the pipeline and its implementation. Although the basic approach looks promising on the selected documents, it strongly depends on the corpus-specific layout and good OCR results.

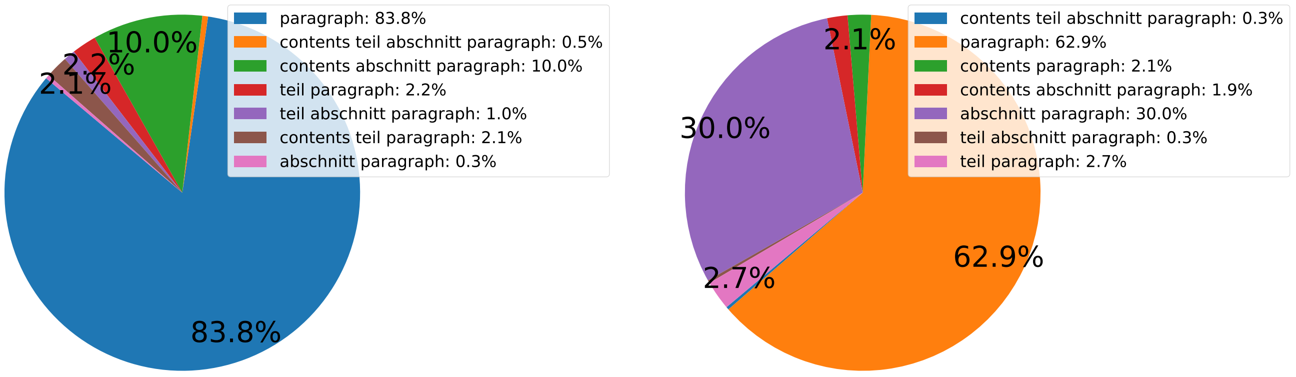
Although Tesseract performs well in many domains, it shows some limitations for the selected documents which make the described postprocessing necessary. It remains to try out different OCR engines, as some commercial tools like ABBYY FineReader not only recognize text, but also some text style information such as underlined, bold, and italic text. This work also shows how much information about a text can be gained using OCR and domain knowledge. Besides that, other tools such as layout-parser [25] or models trained on the annotated DocLayNet dataset [28] that also contain more information than just recognized text will be considered. There are also more advanced table recognition methods such as CascadeTabNet [30] and Table Transformer [37] that need to be evaluated on our data set in future research.

Structured text contains more information and features than plain text that is mapped to regions on an image. Without having to label a large amount of training data, document structure information has been be extracted to get first insights into the digitized text corpus. In future research, we aim at the automated recognition of layout- and text-based rules in order to find these rules for any input document automatically and more flexibility for other documents. Once all regulations of the occupations archive have been digitized, the application of NLP and text mining methods will be much easier and more targeted because each paragraph, section and table are already identified, making a mapping to certain areas of a training program more feasible. Besides that, large, structured text corpora are also useful for the training of large language models and can support researchers in different domains to retrieve information from historical documents much faster.



(a) (b)

Fig. 11: VET (left) and CVET (right) regulations per year according to the ground truth and the extracted data.



(a) (b)

Fig. 12: Distribution of layout types across all VET (left) and CVET (right) regulations in the data set. The percentage describes how many VET (or CVET) regulations consist of these elements. E.g., in the left plot, “paragraph” describes the percentage of VET regulations that consist only of paragraphs, but are not grouped in sections and do not contain a table of contents.

REFERENCES

[1] M. Koistinen, K. Kettunen, and J. Kervinen, “How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine,” *Proc. of LTC*, pp. 279–283, 2017.

[2] A. Nabizai and H.-G. Fill, “Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetzestexten,” *Jusletter IT*, February 2017. [Online]. Available: http://eprints.cs.univie.ac.at/5131/

[3] V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu, and L. Mary Gladence, “Digitization of Data from Invoice using OCR,” in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022. doi: 10.1109/ICCMC53470.2022.9754117 pp. 1–10.

[4] H. Hamann, “The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data,” *Journal of empirical legal studies*, vol. 16, no. 3, pp. 671–688, 2019. doi: https://doi.org/10.1111/jels.12230

[5] N. Kertkeidkachorn and R. Ichise, “T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text,” in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, “OpenIE-based approach for Knowledge Graph construction from text,” *Expert Systems with Applications*, vol. 113, pp. 339–355, 2018. doi: https://doi.org/10.1016/j.eswa.2018.07.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418304329

[7] J. Dörpinghaus and A. Stefan, “Knowledge extraction and applications utilizing context data in knowledge graphs,” in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2019. doi: 10.15439/2019F3 pp. 265–272.

[8] J. Dörpinghaus, A. Stefan, B. Schultz, and M. Jacobs, “Context mining and graph queries on giant biomedical knowledge graphs,” *Knowledge and Information Systems*, vol. 64, no. 5, pp. 1239–1262, 2022. doi: https://doi.org/10.1007/s10115-022-01668-7

[9] Y. Fettach, M. Ghogho, and B. Benatallah, “Knowledge graphs in education and employability: A survey on applications and techniques,” *IEEE Access*, vol. 10, pp. 80 174–80 183, 2022. doi: 10.1109/ACCESS.2022.3194063

[10] J. Dörpinghaus, S. Klante, M. Christian, C. Meigen, and C. Düing, “From social networks to knowledge graphs: A plea for interdisciplinary approaches,” *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100337, 2022. doi: https://doi.org/10.1016/j.ssaho.2022.100337

[11] J. Dörpinghaus, V. Weil, and J. Binnewitt, “Towards the analysis of longitudinal data in knowledge graphs on job ads,” in *The Workshop on Computational Optimization*. Springer, 2022. doi:

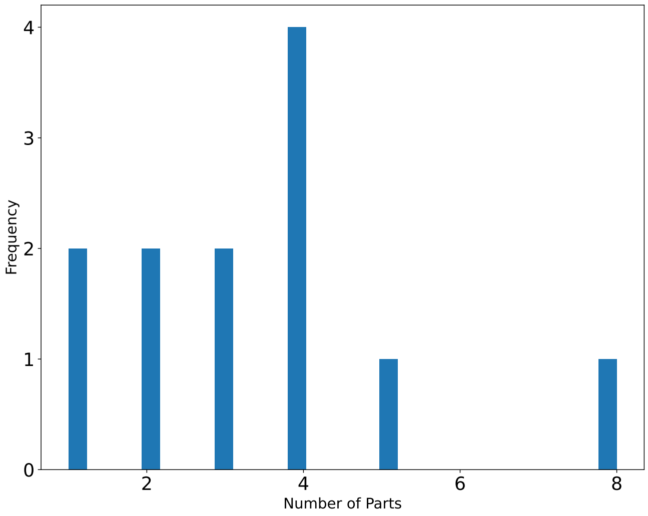


Fig. 13: The number of parts (Teil) in a CVET regulation on the x-axis and the frequency number of regulations with it on the y-axis.

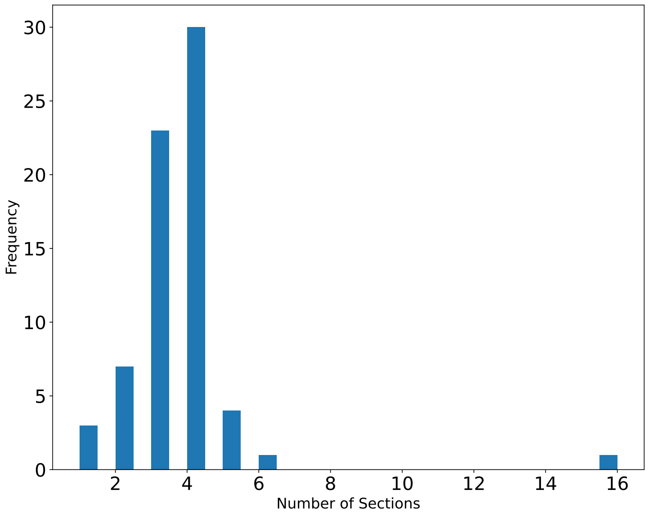


Fig. 14: The number of sections (Abschnitt) in a VET regulation on the x-axis and the frequency number of regulations with it on the y-axis.

https://doi.org/10.1007/978-3-031-57320-0_4 pp. 52–70.

[12] A. Fischer and J. Dörpinghaus, “Web mining of online resources for german labor market research and education: Finding the ground truth?” *Knowledge*, vol. 4, no. 1, pp. 51–67, 2024. doi: https://doi.org/10.3390/knowledge4010003

[13] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, “OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings,” *Applied Sciences*, vol. 9, no. 22, p. 4853, 2019. doi: https://doi.org/10.3390/app9224853

[14] J. M. Jayoma, E. S. Moyon, and E. M. O. Morales, “OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines,” in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2020. doi: 10.1109/HNICEM51456.2020.9400000 pp. 1–6.

[15] S. Van Nguyen, D. A. Nguyen, and L. S. Q. Pham, “Digitalization of Administrative Documents A Digital Transformation Step in Practice,” in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021. doi: 10.1109/NICS54270.2021.9701547 pp. 519–524.

[16] S. Tsujimoto and H. Asada, “Major components of a complete text reading system,” *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1133–1149, 1992. doi: 10.1109/5.156475

[17] J. v. Beusekom, D. Keysers, F. Shafait, and T. Breuel, “Example-based logical labeling of document title page images,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007. doi: 10.1109/ICDAR.2007.4377049 pp. 919–923.

[18] S. Klink and T. Kieninger, “Rule-based document structure understanding with a fuzzy combination of layout and textual features,” *International Journal on Document Analysis and Recognition*, vol. 4, no. 1, pp. 18–26, 2001. doi: https://doi.org/10.1007/PL00013570

[19] P. Pathirana, A. Silva, T. Lawrence, T. Weerasinghe, and R. Abeyweera, “A comparative evaluation of pdf-to-html conversion tools,” in *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 6, 2023. doi: 10.1109/SCSE59836.2023.10214989 pp. 1–7.

[20] P. Lopez, “Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” in *Research and Advanced Technology for Digital Libraries*, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. doi: https://doi.org/10.1007/978-3-642-04346-8_62. ISBN 978-3-642-04346-8 pp. 473–474.

[21] 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: https://doi.org/10.1007/s11390-021-1076-7

[22] C. Neudecker, K. Baierer, M. Federbusch, M. Boenig, K.-M. Würzner, V. Hartmann, and E. Herrmann, “Ocr-d: An end-to-end open source ocr framework for historical printed documents,” in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, ser. DATeCH2019. New York, NY, USA: Association for Computing Machinery, 2019. doi: 10.1145/3322905.3322917. ISBN 9781450371940 p. 53–58. [Online]. Available: https://doi.org/10.1145/3322905.3322917

[23] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, “OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym,” in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds. Cham: Springer International Publishing, 2016. doi: https://doi.org/10.1007/978-3-319-50835-1_66. ISBN 978-3-319-50835-1 pp. 735–746.

[24] M. Lundqvist and A. Forsberg, “A comparison of OCR methods on natural images in different image domains,” 2020.

[25] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li, “Layoutparser: A unified toolkit for deep learning based document image analysis,” pp. 131–146, 2021. doi: https://doi.org/10.1007/978-3-030-86549-8_9

[26] 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.

[27] Z. Shen, K. Zhang, and M. Dell, “A large dataset of historical japanese documents with complex layouts,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. doi: 10.1109/CVPRW50498.2020.00282 pp. 2336–2343.

[28] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. W. J. Staar, “Doclaynet: A large human-annotated dataset for document-layout segmentation,” p. 3743–3751, 2022. doi: 10.1145/3534678.353904. [Online]. Available: https://doi.org/10.1145/3534678.3539043

[29] X. Zhong, E. ShafieiBavani, and A. J. Yepes, “Image-based table recognition: data, model, and evaluation,” 2020. [Online]. Available: https://arxiv.org/abs/1911.10683

[30] D. Prasad, A. Gadpal, K. Kapadni, M. Visave, and K. Sultanpure, “Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents,” 2020.

[31] R. Altenhöner, A. Berger, C. Bracht, P. Klimpel, S. Meyer, A. Neuburger, T. Stäcker, and R. Stein, “DFG-Praxisregeln "Digitalisierung". Aktualisierte Fassung 2022.” Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7435724

[32] W. Meier, “exist: An open source native xml database,” in *Web, Web-Services, and Database Systems*, A. B. Chaudhri, M. Jeckle, E. Rahm, and R. Unland, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. doi: https://doi.org/10.1007/3-540-36560-5_13. ISBN 978-3-540-36560-0 pp. 169–183.

[33] e editiones, “Tei publisher,” accessed: 2024-07-15. [Online]. Available: https://teipublisher.com

[34] L. O’Gorman and R. Kasturi, *Document Image Analysis*. IEEE Computer Society Press Los Alamitos, 1995, vol. 39.

[35] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*. Morgan kaufmann, 2022.

[36] TEI Consortium, eds., *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, TEI Consortium, 2024, last modified 2024-07-08. [Online]. Available: http://www.tei-c.org/Guidelines/P5/

[37] B. Smock, R. Pesala, and R. Abraham, “PubTables-1M: Towards comprehensive table extraction from unstructured documents,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. doi: https://doi.org/10.48550/arXiv.2110.00061 pp. 4634–4642.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 151–158
DOI: 10.15439/2024F8384
ISSN 2300-5963 ACSIS, Vol. 41

Smart Assistants for Enhancing System Security and Resilience

Andrey Sadovykh
0000-0003-2384-5447
Softeam, Paris, France
andrey.sadovykh@softeam.fr

Dragos Truscan, Tanwir Ahmad
0000-0002-4367-6225 0000-0003-3416-2422
Åbo Akademi University, Turku, Turku
firstname.lastname@abo.fi

Martin A. Schneider
0000-0002-8864-6492
Fraunhofer FOKUS, Berlin, Germany
martin.schneider@fokus.fraunhofer.de

Wissam Mallouli, Ana Cavalli
0000-0003-2548-6628 0000-0003-2586-9071
Montimage
Paris, France
firstname.lastname@montimage.org

Cristina Seceleanu
0000-0003-2870-2680
Mälardalen University
Västerås, Sweden
cristina.seceleanu@mdu.se

***Abstract*—Security and resilience have become paramount concerns for integrated system manufacturers as the number of vulnerabilities continues to increase annually. Cyber threats pose significant risks with substantial potential impacts on both manufacturers and end users. New regulations, such as the EU Cybersecurity Act and Cyber Resilience Act, mandate stricter practices and thorough verification throughout development and operations. Implementing a holistic DevSecOps process encompassing threat analysis, requirements engineering, development practices, verification, and operations management is challenging for large enterprises and SMEs. This complexity arises from the need for specialized expertise, knowledge of various techniques and tools, rigorous principle application, and thorough verification at each step, making the process costly, time-consuming, and potentially stifling innovation and time-to-market.**

Our proposal introduces a suite of smart assistants designed to work collaboratively with engineers. These assistants recommend best practices and tools, suggest context-specific regulatory requirements, analyze design architecture, generate tailored code and configurations, and conduct resilience tests. This comprehensive approach aims to ensure the correctness and completeness necessary for security and regulatory compliance.

***Index Terms*—DevSecOps, Smart assistants, Security, Resilience, Requirements Engineering, Security by Design, Testing, Monitoring, Anomaly Detection**

I. INTRODUCTION

IN THE digital age, where software underpins virtually every aspect of modern life, security has emerged as a paramount concern [1], [2]. However, the relentless pursuit of fast deployment often takes precedence over robust security practices, leading to the proliferation of vulnerabilities and insecure applications [3]. This paper investigates the urgent need for a paradigm shift towards integrated hardware and software security engineering to address this pressing issue.

At the heart of this effort lies the recognition that software forms the backbone of IT infrastructures, services, and products. Yet, despite its pervasive influence, the current software development practices prioritize speed over security, leaving systems vulnerable to attacks. Compounding this challenge is the fact that a significant portion of the software and hardware utilized within the European Union (EU) is developed outside its borders, necessitating stringent security requirements and their verification to comply with EU standards.

Central to our investigation is the imperative for the EU to ensure the verifiability and auditability of software and hardware concerning their security. This includes a comprehensive analysis of the security implications associated with using open-source software and hardware, as well as strategies for enhancing security auditability within this context. The latest supply chain attacks on open source software projects, such as the xz utils backdoor [4], underpin this imperative, moreover, in a digital ecosystem characterized by perpetual updates and evolving regulatory frameworks. Hence, there arises a critical need for methodologies and tools that facilitate continuous security assessments to adapt to the dynamic nature of modern software and hardware landscapes.

Several initiatives have proposed a holistic cybersecurity view under the DevOps paradigm. Among them, the VeriDevOps project [5] proposed integrating DevOps principles with early verification, test automation, and monitoring to ensure software security and reliability. VeriDevOps provides a systematic approach to embedding security requirements throughout the software development lifecycle. Key technologies include Natural Language Processing (NLP) for analyzing and formalizing security specifications and automated tools for quality assurance, system testing, and runtime monitoring. VeriDevOps automates the configuration of trace monitors based on security requirements and employs continuous monitoring to detect anomalies and vulnerabilities. It also generates attack tests to identify invalid states and security weaknesses. Additionally, it performs automated design and code checks using semi-structured and structured formalisms, either through model simulation or formal verification, to ensure compliance with security standards.

Thanks to AIDOaRt, grant agreement No 101007350 under JU ECSEL and Horizon 2020 funding programs

©2024, PTI

Applying VeriDevOps may pose several challenges. While it proposes more than 20 tools, it also requires expertise across multiple domains, including Threat Analysis, Security Requirements Engineering, Testing, Monitoring, and Incident Analysis. Despite automating many steps, significant manual input is still necessary, making applying the methodology tedious. Additionally, integrating specific security methods for hardware development and verification into the workflow is complex. While VeriDevOps covers many scenarios, it cannot encompass all possible security situations, necessitating further expansion to address diverse use cases comprehensively. Thus, applying VeriDevOps effectively demands both broad expertise and ongoing adaptation to cover more specific scenarios.

The proposal of this paper is to elevate the intelligence and automation of cybersecurity in system development by incorporating AI-based assistance. This assistance would provide recommendations for configuring and formalizing diverse security properties, selecting suitable testing techniques, and interpreting the results. Additionally, it would offer boilerplates, examples, and detailed explanations for various methods and tools, thereby streamlining the implementation process. One of the significant challenges is ensuring the AI-based system effectively suggests design approaches and coding practices that enhance system resilience, maintaining operation even amidst attacks, including hardware integration aspects. This necessitates the AI is able to understand and apply resilience requirements accurately, guiding developers to meet these requirements and verify compliance effectively. Furthermore, AI assistance must facilitate informed decision-making by recommending appropriate methods for satisfying resilience requirements. This includes providing insights into the application of these methods to ensure thorough verification of compliance. The challenge lies in the AI's capacity to interpret complex security specifications and translate them into actionable guidance that aligns with industry standards and best practices.

Overall, integrating AI-based assistance within VeriDevOps and enriching it with hardware aspects aims to overcome these challenges. By providing intelligent, context-aware advice and support along with facilitating comprehensive verification processes, this approach seeks to enhance the resilience and security of systems in a systematic and scalable manner.

The paper is structured as follows: Section II presents the background, including the achievement of the project and a review of related works, setting the context for SecDevOps and automation with smart assistants. Section III introduces the concept of employing smart assistants throughout different phases of the SecDevOps cycle, demonstrating its potential to enhance security practices through main scenarios and flows in real-world development contexts. Finally, Section IV concludes the paper by summarizing key contributions and outlining future research and development avenues.

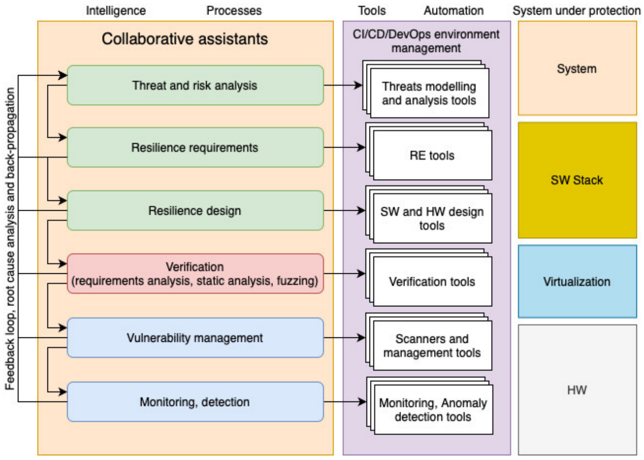


Fig. 1. The Concept of Smart Assistants for Continuous Holistic Security Verification

II. BACKGROUND AND RELATED WORK

A. Cybersecurity Engineering Process

Cybersecurity implementation within a system must commence at the foundational level of requirements specification. These requirements guide system design, component selection, implementation, integration, and subsequent verification and validation processes. The efficacy of requirements specification hinges upon various parameters, notably clarity, atomicity, and verifiability. Formal and semi-formal specification techniques play a pivotal role in automating the verification process, aiding in identifying and mitigating cybersecurity threats throughout the system's lifecycle. Initiating with a comprehensive understanding of the system's scope and the assets requiring protection against cybersecurity threats, the requirements specification informs critical design decisions. These decisions encompass the selection of appropriate hardware and software platforms and the configuration of the application stack. Moreover, the design phase necessitates strategic considerations regarding access control mechanisms and internal restrictions, which dictate the architectural blueprint of the system. After design, adherence to coding practices becomes imperative to minimize the injection of vulnerabilities and fortify mechanisms for secure storage of sensitive information.

Beyond the developmental phase, systems are inevitably exposed to many attacks aiming at exploiting vulnerabilities within the application, software, and hardware infrastructure. Given the omnipresent nature of vulnerabilities, incessantly discovered at a mass scale, preemptive measures must be in place to detect and manage these vulnerabilities through timely patches and other protective measures. Anticipating vulnerabilities before disclosure underscores the importance of anomaly detection mechanisms capable of preemptively identifying potential attack vectors.

Moreover, hardware resilience assessment is essential in ensuring the overall security and reliability of computing systems, as it complements software resilience assessment by addressing vulnerabilities inherent in the physical components. Techniques for hardware resilience assessment include rigorous testing for fault tolerance, stress testing, and examining supply chain integrity. These assessments face significant challenges, such as detecting and mitigating hardware backdoors, counterfeit components, and vulnerabilities introduced during manufacturing. Given that hardware can originate from untrusted sources, it is critical to validate its security through thorough inspection and verification processes. This layered approach helps prevent the software from inheriting hardware-related vulnerabilities, thereby fortifying the system resilience.

Resilience in software and hardware systems is fundamentally the ability of a system to resist, absorb, recover from, and adapt to adverse conditions, particularly in the face of physical and sophisticated cyber-attacks [6]. Systems, characterized by the integration of physical components and software elements, are inherently susceptible to a range of threats aimed at disrupting system availability, perturbing performance, and other malign objectives. The repercussions of such attacks extend beyond system disruption and can trigger multilevel consequences, including economic, social, and environmental impacts. Moreover, the interconnected nature of components means that an attack on one part of the system can negatively affect other parts. This interconnectedness is even more pronounced in our world of interlinked critical infrastructures, where an attack on one system can trigger cascading failures across others. Therefore, it is imperative that systems are engineered with resilience in mind, adopting proactive designs and reactive countermeasures to effectively mitigate these threats [7]. These resilience techniques are multifaceted and can be classified into several categories as proposed by NIST.

B. The VeriDevOps Framework

VeriDevOps proposed a methodology [8] that merges DevOps principles with early verification, test automation, and monitoring, ensuring the security and reliability of systems. It provides a structured approach to software development, emphasizing the continuous integration of security requirements throughout the development lifecycle. At its core, VeriDevOps automates key aspects of software development with security in mind, including defining and analyzing security requirements, conducting testing and monitoring, and integrating these processes into established VeriDevOps practices.

The process begins with analyzing and formalizing text-based security requirements gathered from various sources. NLP and pattern recognition technologies play a crucial role in maintaining consistency and clarity in these specifications. Additionally, patterns are translated into temporal logic for better understanding. Another essential aspect is the automated configuration of trace monitors, which are based on formalized security requirements using structured formalisms. These monitors are continuously adjusted and monitored over time to detect anomalies and vulnerabilities during runtime. Furthermore, VeriDevOps automatically generates attack tests based on security requirements, aiming to expose potential vulnerabilities by pushing the system into invalid states. These tests complement positive testing methods and reveal insecure behaviors that may go unnoticed otherwise. To further improve testing, guidelines can be established for testers to propose scenarios that evaluate both security and energy properties, often overlooked areas. Finally, VeriDevOps automates design and code checks according to specified security requirements using semi-structured and structured formalisms. These verification activities can be carried out through simulation or formal verification of system descriptions.

The VeriDevOps Methodology encompasses a suite of interconnected tool sets designed for Security Requirements Generation, Reactive Protection at Runtime, and Prevention at Design and Development. These tool sets are closely integrated to align security requirements with design analysis, code-level verification, and runtime system analysis. They comprise concrete tool components provided and developed by VeriDevOps partners, varying in licensing policies and maturity levels. While some tools are well-established commercial or open-source solutions, others are more experimental. However, all tools must adhere to the interfaces and features outlined in the VeriDevOps Methodology and be interchangeable to a certain extent. Case studies combine these tools in a specific industry context based on their alignment with requirements and compatibility with industry practices.

1) Requirements Specification: Security requirements undergo examination and formalization, sourced from diverse textual descriptions. To ensure consistency and clarity while avoiding inconsistencies and ambiguities, we leverage NLP alongside established patterns or boilerplates. Additionally, techniques are employed to automatically translate these patterns into temporal logic, further enhancing requirement clarity and consistency. Various techniques, such as PROPAS and RQCODE, can be integrated into the VeriDevOps methodology for requirements formalization. Manual and semi-automatic translation methods are also employed to optimize this process. Furthermore, verification and analysis tasks can be executed by either simulating the final model or verifying the system's description. Using natural languages and model smells, we've established indicators (e.g., NALABS) for security requirement flaws and defined metrics to automatically detect these flaws in security artifacts.

2) Prevention at Development: In this phase, multiple techniques are employed for test modeling (e.g., UPPAAL, PyLC, Modelio, GW2UPPAAL), automated test generation (e.g., MetaTester, CompleteTest, Graphwalker), and vulnerability localization (e.g., Localizer, RCA). This information aids in generating both positive and negative tests intended to push the system into specific states to expose potential vulnerabilities. To enhance this process, establishing guidelines and a format enabling testers to design scenarios evaluating not only security aspects but also energy properties would be beneficial, as energy properties are often overlooked in testing.

3) Protection at Operations: An automated setup of monitoring tools (e.g., MMT, THOE, EARLY), based on the specification of security requirements in natural language, semi-structured, or structured formalisms is available. Over time, these traces are automatically configured and continuously observed using formal or semi-formal specifications. Runtime monitoring, which observes system behavior during operation, is implemented to detect errors, monitor performance, ensure compliance, and maintain system health. This serves as a foundation for potential preemptive countermeasures.

VeriDevOps represents a departure from traditional software development methodologies by placing security at the forefront of the development process. By integrating security requirements analysis, automated testing, vulnerability localization, and continuous monitoring into the DevOps pipeline, VeriDevOps ensures a holistic approach to software security, thus enabling organizations to proactively identify and mitigate security threats throughout the software development lifecycle, thereby enhancing the resilience of software systems.

C. Challenges

Addressing security challenges within the DevOps framework involves navigating several complex issues. Integrating security requirements into the DevOps pipeline presents a significant challenge. It requires that security considerations are seamlessly woven into existing development and deployment processes without impeding agility or efficiency.

Achieving clarity and detailing security requirements is crucial yet challenging. It necessitates expressing security needs in a clear, unambiguous manner that leaves no room for interpretation. However, achieving this clarity becomes more complex when utilizing diverse formal methods and tools across different stages of the development lifecycle. Moreover, integrating security requirements analysis and verification throughout the entire DevOps process is essential but challenging. It involves overcoming barriers to incorporating security considerations at every stage, ensuring consistency and accuracy in specifying and analyzing requirements across diverse environments. Additionally, automating security test generation and selection poses its own set of challenges. Identifying appropriate tools and techniques for generating automated security tests and ensuring their seamless integration within the DevOps pipeline can be a daunting task, requiring careful consideration of compatibility and effectiveness.

Implementing robust security monitoring systems presents another challenge. This entails establishing comprehensive monitoring across all critical components of the DevOps pipeline, detecting and responding to security threats in real-time while maintaining system performance and reliability. Another significant set of challenges involves supporting and guiding developers through the implementation of the DevSecOps methodology to facilitate their ability to select appropriate methods and tools, configure them effectively, and utilize them proficiently. These challenges encompass various aspects such as formal specifications, static analysis, testing, monitoring, root cause analysis, and the management of vulnerabilities. Addressing these challenges is essential for empowering developers to seamlessly integrate security practices into the software development lifecycle to ensure the reliability and security of the resulting software products.

Finally, integrating methodologies, tools, and technologies within Continuous Integration, Deployment, DevOps and CyberOps practices is indispensable. By embedding security checks within automated pipelines and establishing real-time monitoring mechanisms, organizations can ensure adherence to quality standards and resilience practices across the system lifecycle's developmental and operational phases. This holistic approach enhances system security and fosters a culture of proactive cybersecurity within the organizational framework.

D. Smart Assistants Background

The use of smart assistants (sometimes termed as *bots*) in developing computing systems has become increasingly prominent, primarily due to advancements in generative artificial intelligence (AI) and machine learning (ML). Smart assistants in the context of software engineering are tools and platforms that leverage AI and ML to aid developers in various aspects of software development, maintenance, and management. These AI-driven tools enhance productivity, improve code quality, and streamline development processes.

The integration of smart assistants into software engineering is transforming the way developers write code, test software, manage projects, and interact with development environments. As AI technologies continue to evolve, these tools are expected to become even more sophisticated, further enhancing their capabilities to support the system development lifecycle.

The penetration of smart assistants has been observed in all major areas of the system development process to assist with different tasks. As summarized by several systematic literature reviews [9]–[11] the assistants can span tasks from *development automation* (requirements processing, code generation [12], debugging, testing [13], documentation generation), *real-time collaboration and support* (coding assistance via code completion [14], error detection and correction [15], [16], code reviews [17]–[19]) to *project management* (tracking progress, predicting timelines, and identifying bottlenecks) just to enummerate the main ones.

The evolution of large language models (LLMs) has enabled many of the previously enumerated activities. However, recent studies [20] have shown that their benefits are limited by a set of open problems, such as hallucinations.Nevertheless, the authors of [21] provide a survey of how LLM-based agents can support the planning activities of complex processes, which can also be applied to IT systems.

Following similar techniques and practices as described above, smart assistants can also be used to assist with cybersecurity-related activities throughout the development process. Threat detection and analysis, automated response, vulnerability management, compliance and reporting, security training and awareness, and forensics are just a few that can benefit from the capabilities of smart assistants. However, as emphasized in [22], there is still a need for automated intelligent tools to assist cybersecurity-related tasks.

III. SMART ASSISTANTS FOR CONTINUOUS HOLISTIC SECURITY VERIFICATION

Utilizing smart assistants within the VeriDevOps framework represents a cutting-edge innovation in system development. Such assistants facilitate various aspects of the security lifecycle, from requirements specification to vulnerability analysis and remediation. By harnessing the power of AI, VeriDevOps

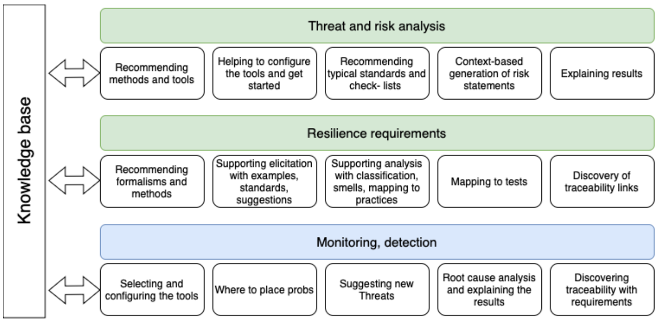

Knowledge base
Threat and risk analysis
Recommending methods and tools
Helping to configure the tools and get started
Recommending typical standards and check- lists
Context-based generation of risk statements
Explaining results
Resilience requirements
Recommending formalisms and methods
Supporting elicitation with examples, standards, suggestions
Supporting analysis with classification, smells, mapping to practices
Mapping to tests
Discovery of traceability links
Monitoring, detection
Selecting and configuring the tools
Where to place probs
Suggesting new Threats
Root cause analysis and explaining the results
Discovering traceability with requirements

Fig. 2. Tasks examples supported by Smart Assistants

empowers development teams to make informed security decisions and streamline security-related tasks, thus improving the efficiency and effectiveness of the development process. Integrating smart assistants within the VeriDevOps framework offers a multifaceted approach to addressing cybersecurity challenges across the development lifecycle. We outline key concepts and provide illustrative scenarios to motivate the application of smart assistants in real-world development.

To address the needs outlined above, we propose the comprehensive integration of specialized smart assistants across all lifecycle phases, including threat and risk analysis, verification, and monitoring, as in Figure 1. These smart assistants will collaborate and assume responsibilities such as selecting and configuring task-specific tools, generating relevant artefacts, and evaluating the outcomes of these processes. For instance, a smart assistant designated for verification tasks might select and configure appropriate tools for verification, such as static analysis and fuzzing tools, execute the analysis and testing procedures, and assess the results, as depicted in Figure 2.

To ensure a comprehensive approach towards the development lifecycle, the scope of the smart assistants extends beyond individual development phases and activities. These assistants will also facilitate the collaboration among various smart assistants responsible for different tasks. This is achieved through the sharing of results, thereby creating a continuous feedback loop among the smart assistants. As demonstrated in Figure 2, this systematic exchange of information fosters a holistic understanding of the cybersecurity posture of a product. Other smart assistants can leverage this aggregated information to enhance their outputs, e.g., by incorporating identified threats and risks into the verification phase.

Additionally, this collaborative framework benefits the operational dynamics among smart assistants and provides manufacturers with a comprehensive overview of the system through the storage and further processing of artefacts within a knowledge base. This is then utilized to generate summaries and overviews of various activities, thus offering a consolidated view of system security and performance.

A. *Smart Assistants for Structured Resilience Requirements*

Security is a dynamic challenge, as recent supply chain attacks on the Linux kernel and new cybersecurity regulation such as he Cyber Security Act (CSA) and the Cyber Resilience Act (CRA) emphasize the need for refined development methodologies. These methodologies should not only enable engineers to create and deliver products with elevated security standards and thorough verification and validation but also assist them in comprehending and integrating regulatory demands and requirements into the development process to demonstrate compliance with pertinent regulations and standards. Furthermore, emerging threats and risks associated with the supply chain necessitate a deeper understanding to address and document these concerns effectively.

Central to our proposal are AI-based smart assistants designed to optimize cyber resilience through enhanced threat modelling and analysis. These smart assistants utilize established standards like IEC 62443, regulations such as the CSA, CRA, and NIST guidelines to identify, analyze, and understand the implications of threats. Acting as a pivotal interface between system engineers and the development processes, these smart assistants ingest threat data from diverse sources, including vulnerability databases such as Common Vulnerabilities and Exposures, Common Weakness Enumerations, Security Technical Implementation Guides, Threat Intelligence and Management Platforms, and 0-day vulnerabilities. By referencing industry standards and regulations, the smart assistants evaluate the relevance of these threats to the system under development, ensuring compliance and pinpointing specific threats that need mitigation. This facilitates a comprehensive threat modelling process, aligning analyzed threat data with the system's design specifications, thereby enabling system engineers to proactively address vulnerabilities and incorporate security measures from the project's inception.

Due to the increasing complexity of security threats, effective yet flexible specification methods that support rigorous analysis of software security requirements are needed. Security requirements specifications that consider thematic roles and domain knowledge to enable deep semantic analysis are desirable. We aim to develop specific assistants similar to those for code generation (e.g., GitHub Copilot), which will empower engineers to generate consistent and testable specifications by interpreting natural language security requirements. The assistant will be based on our work on the semantic analysis framework of *ReSA*, a structured, pattern-based language and ontology for specifying embedded systems requirements [23]–[25], as well as on our previous work on EARS-based test generation for PLC programs [26].

1) *Example: Automated Threat Modelling and Analysis:*

Scenario: A manufacturing company is upgrading its industrial control systems (ICS) to enhance security and comply with international standards and regulations. The company aims to ensure that its systems are resilient against cyber threats and comply with relevant regulatory requirements.

Flow: To utilize smart assistants for conducting resilience threats analysis and suggesting mandatory and recommended requirements from EU regulatory frameworks (such as the CRA), IEC 62443, NIST guidelines, and the EUCC.

1) *Resilience Properties Modelling:* The company's cy-

bersecurity team initiates the resilience threats analysis process for the new ICS upgrade project. They model system properties with the smart assistant, including the system's architecture, intended operational environment, and potential threat vectors.
2) *Threat Identification:* The smart assistant analyzes the properties and identifies potential resilience threats specific to the ICS, such as supply chain attacks, insider threats, and vulnerabilities in communication protocols.
3) *Regulatory and Standards Compliance:* The smart assistant aligns identified threats with regulatory standards, suggesting mandatory requirements, e.g., from the EU CRA, guidelines from IEC 62443, best practices from NIST, and security assurance methods from the EUCC, covering risk management, incident reporting, access control, monitoring, and evaluation.
4) *Resilience Requirements Suggestions:* The smart assistant provides a categorized list of mandatory and recommended requirements, such as risk management (CRA), Security Level 3 compliance (IEC 62443), incident response planning (NIST), and EAL4+ certification (EUCC), along with recommendations like continuous monitoring, security audits, and penetration testing.
5) *Resilience Planning, Implementation and Verification:* The cybersecurity team reviews the smart assistant's suggestions and develops an action plan, assigning responsibilities, setting timelines, and allocating resources for each activity. The team implements the mandatory and recommended requirements, using the smart assistant for guidance. The assistant continuously monitors compliance status and alerts the team to deviations or emerging threats. After implementation, the smart assistant helps to verify compliance through automated checks and generates detailed reports. These reports are used for internal review and are submitted to regulatory bodies for compliance verification.

B. Smart Assistants for Security Testing

Cybersecurity testing is a critical phase in the development lifecycle, comprising various activities including planning, requirements analysis, test design, execution, evaluation, and comprehensive reporting. There exists a plethora of methodologies, strategies, and tools designed to facilitate these processes. Notwithstanding, activities such as planning frequently remain manual tasks. Smart assistants can be invaluable in these areas, particularly in translating test strategies, cybersecurity requirements, and risk assessments into detailed test plans. Additionally, they can streamline tasks related to the preparation of testing processes, such as selecting and configurating tools, which can be an arduous activity. Interpreting the outcomes of security testing requires extensive technical knowledge concerning the system, operating platforms, programming languages, and weaknesses to accurately evaluate the implications of vulnerabilities. Moreover, techniques like static analysis and dynamic testing exhibit unique strengths and limitations, necessitating further analysis and additional testing to refine and verify results. Smart assistants can play a pivotal role by deploying advanced tools to convert intermediate data into final outcomes, effectively distinguishing between true and false positives, thereby improving the accuracy of cybersecurity testing. Ultimately, smart assistants contribute to the development of effective security patches through automated program repair techniques, e.g., using Code LLMs.

1) Example: Security Test Generation: **Scenario**: A software development team is tasked with building a new web application with stringent security requirements. They leverage automated test generation and vulnerability localization tools within the VeriDevOps framework to ensure the application's resilience against potential cyber threats.

Flow:
1) *Security Requirements Analysis*: The development team is using smart assistants to identify comprehensive security test requirements for the web application based on the knowledge base populated by the smart assistant.
2) *Security Test Planning & Control*: Smart assistant will support the planning of the security testing, in particular, propose complementary security testing tools and test exit criteria based on the security test requirements, existing licenses, and used technology.
3) *Security Test Generation & Execution*: Smart assistants propose security testing tools from the planning phase of which the development team selects those that fit best to their experience. Smart assistants configure these tools using knowledge from previous projects and from the community. Developers check the configuration and modify it where necessary based on their own expertise. Smart assistants learn from these modifications for future processes. Using security testing tools configured by smart assistants and the development in a collaborative manner, a test suite is generated and executed, aimed at evaluating the application's security posture. The test suite encompasses both positive and negative scenarios, covering various attack vectors and vulnerabilities.
4) *Security Test Evaluation* The test cases produce a large number of results. Smart assistants help the development team to get an overview, e.g., by analyzing the test results with respect to relevance and severity. Further tests can be generated and executed by smart assistants to obtain more information on findings. Finally, smart assistants can automatically populate bug-tracking systems with consolidated, prioritized information from the test evaluation based on results from testing and threat analysis and alert the development teams if required.
5) *Reporting*: Smart assistants generate draft test reports based on the performed activities. The development team completes reviews and finalizes these reports. Smart assistants summarize these reports for the management.
6) *Remediation and Patching*: The development team addresses the identified vulnerabilities. Smart assistants propose potential security patches and assess their appropriateness using patch validation techniques. The development team selects promising patch candidates

and improves them. Smart assistants perform again patch validation and regression techniques to assess the patch until an effective one has been identified.

This flow demonstrates how development teams can benefit from smart assistants to prepare, perform, evaluate and report security tests in a consolidated and efficient manner, where the development team and smart assistants collaborate to identify vulnerabilities with increased efficiency.

C. Smart Assistants for Resilience Monitoring

Specific assistants are needed to help monitor, detect, and respond to security threats more efficiently. By enhancing and automating many aspects of cybersecurity operations, smart assistants are becoming an essential component of modern security strategies, helping to mitigate the increasing complexity and frequency of cyber threats. Smart assistants will enable Automated Threat Detection by continuously monitoring network traffic, system logs, and other data sources for suspicious activity and detecting anomalies that might indicate a breach or an attempted attack. Upon detecting a potential threat, these AI-driven systems can generate real-time alerts. This immediate notification allows security teams to act swiftly, potentially stopping attackers before they can cause significant damage. As such, smart assistants will provide automated response capabilities, deciding and executing predefined actions when certain types of threats are detected. This might include isolating affected systems, blocking IP addresses, or initiating patches to vulnerable software. In addition, by analyzing historical data and identifying trends, smart assistants can predict and identify potential future threats. This predictive capability helps in proactive threat management, allowing organizations to strengthen defences before an attack occurs.

1) Example: Automated Threat Detection and Response: **Scenario**: An organization is deploying a new software application that handles sensitive user data. To ensure the security of this application, the organization integrates automated threat detection and response mechanisms by applying the VeriDevOps pipeline with the help of specific assistants.

Flow:

1) *Security Requirements Specification*: The organization generate structured security requirements based on natural language descriptions of potential threats and vulnerabilities associated with the application.
2) *Continuous Monitoring Setup*: Smart assistants configure tools to continuously monitor network traffic, system logs, and application behaviour for suspicious activity.
3) *Real-time Threat Detection*: As the application is deployed and operational, the monitoring tools detect anomalous patterns in user access patterns and data usage, signalling potential security threats.
4) *Automated Alert Generation*: Upon detecting suspicious activity, the monitoring tools automatically generate real-time alerts, notifying the security team of the potential security breach.
5) *Automated Response*: Smart assistants trigger predefined actions, such as isolating affected systems, blocking IP addresses associated with suspicious activity, and initiating patches to mitigate vulnerabilities.
6) *Incident Analysis and Resolution*: The security team analyzes the alerts and response actions to identify the root cause of an incident and implement further measures to prevent similar incidents in the future.

This flow illustrates how automated threat detection and response mechanisms, integrated within the VeriDevOps framework, enable organizations to identify and mitigate security threats in real-time proactively.

D. Smart Assistants Collaboration

Our proposed framework leverages a hierarchical structure of smart assistant agents to address various aspects of the DevSecOps lifecycle. High-level agents oversee the entire process, strategically delegating tasks to lower-level agents specializing in specific areas. These lower-level agents can automate tasks such as identifying vulnerabilities based on established security protocols, performing code reviews in order to detect potential security concerns within the codebase, generating test cases, and even proposing patches for vulnerabilities. This collaborative approach fosters a more efficient workflow, where specialized agents handle routine tasks while high-level agents maintain an overarching view and ensure progress towards overall security objectives.

To further enhance the reliability of the framework, the agents can employ self-assessment techniques. These techniques involve cross-referencing findings with established security protocols and identifying potential biases or errors within the generated outputs, such as "hallucinations" in the context of LLMs. When necessary, the agents can seek clarification from human experts, ensuring the accuracy and effectiveness of their work. This self-assessment helps to minimize errors and ensures that the agents operate within the bounds of established security best practices.

In summary, the proposed framework utilizing a hierarchy of smart assistants represents a significant advancement towards a more automated, efficient, and secure software development lifecycle. By automating and streamlining various tasks, developers and operations teams can focus their efforts on more complex and critical aspects of the development lifecycle. Furthermore, we plan to have proper safeguards, validation mechanisms, and human oversight within the framework to ensure the reliability and security of the outputs generated by these agents. Additionally, continuous training and refinement of AI-based assistants would be necessary to keep pace with the ever-evolving landscape of software development practices and security threats.

IV. CONCLUSIONS

In this paper, we have presented a proposal for a holistic approach to enhancing system security and resilience through the integration of smart assistants within the software development lifecycle. By combining innovative methodologies, automated processes, and AI-driven assistance, our approach offers a comprehensive framework for building secure and resilient system systems. We introduced the concept of smart assistants tailored to various stages of the system development lifecycle, from requirements specification to vulnerability analysis and remediation. These assistants leverage AI technologies to empower development teams, streamline security-related tasks, and make informed security decisions. Through their integration, organizations can proactively identify and mitigate security threats throughout the development process. By automating tasks such as security requirements analysis, code review, and vulnerability testing, smart assistants enable development teams to focus on building high-quality software while ensuring security best practices are followed.

However, the adoption of smart assistants is not without challenges. Integration complexities, ensuring clarity of security requirements, and addressing automation limitations are among the hurdles that organizations encounter. Overcoming these challenges will require collaboration across the organization and ongoing research and development efforts in the field of software security. Despite these challenges, the benefits of smart assistants within the development lifecycle are significant. By leveraging AI-driven assistance, organizations can build more secure and resilient systems, ultimately mitigating the impact of cyber threats and protecting critical digital assets.

The integration of smart assistants represents a promising approach to enhancing software security in today's digital landscape. By embracing innovation, addressing challenges, and fostering collaboration, organizations can leverage smart assistants to build a more secure and resilient digital future.

ACKNOWLEDGMENT

The proposed work is the fruit of the authors' extensive collaboration in many research projects and mainly AIDOaRt [27], which has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007350. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Sweden, Austria, Czech Republic, Finland, France, Italy, and Spain.

REFERENCES

[1] M. Cankar et al. Security in devsecops: Applying tools and machine learning to verification and monitoring steps. In M. Vieira et al., editors, *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering, ICPE 2023, Coimbra, Portugal, April 15-19, 2023*, pp. 201–205. ACM, 2023. 10.1145/3578245.3584943.

[2] Y. He et al. Towards security threats of deep learning systems: A survey. *IEEE Trans. Software Eng.*, 48(5):1743–1770, 2022.

[3] D. Bassi and H. Singh. A systematic literature review on software vulnerability prediction models. *IEEE Access*, 11:110289–110311, 2023.

[4] A. Freund. backdoor in upstream xz/liblzma leading to ssh server compromise. post on mailing list oss-security@openwall. https://openwall.com/lists/oss-security/2024/03/29/4, 2024. Accessed: 2024-04-25.

[5] A. Sadovykh et al. VeriDevOps: Automated Protection and Prevention to Meet Security Requirements in DevOps. In *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1330–1333, February 2021.

[6] J. Jia et al. Software approaches for resilience of high performance computing systems: a survey. *Frontiers Comput. Sci.*, 17(4):174105, 2023.

[7] S. M. Alhidaifi et al. A survey on cyber resilience: Key strategies, research challenges, and future directions. *ACM Comput. Surv.*, 56(8):196:1–196:48, 2024.

[8] E. P. Enoiu et al. VeriDevOps Software Methodology: Security Verification and Validation for DevOps Practices. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, ARES '23, pp. 1–9, New York, NY, USA, August 2023. Association for Computing Machinery.

[9] M. Savary-Leblanc et al. Software assistants in software engineering: A systematic mapping study. *Software: Practice and Experience*, 53(3):856–892, 2023.

[10] S. Santhanam et al. Bots in software engineering: a systematic mapping study. *PeerJ Comput Sci*, 8:e866, February 2022.

[11] R. Moguel-Sánchez et al. Bots and their uses in software development: A systematic mapping study. In *2022 10th International Conference in Software Engineering Research and Innovation (CONISOFT)*, pp. 140–149, 2022.

[12] A. Svyatkovskiy et al. Intellicode compose: code generation using transformer. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pp. 1433–1443, New York, NY, USA, 2020. Association for Computing Machinery.

[13] A. Fontes and G. Gay. The integration of machine learning into automated test generation: A systematic mapping study. *Software Testing, Verification and Reliability*, 33(4):e1845, 2023.

[14] M. Ciniselli et al. An empirical study on the usage of transformer models for code completion. *IEEE Transactions on Software Engineering*, 48(12):4818–4837, 2022.

[15] D. Drain et al. Generating bug-fixes using pretrained transformers. In *Proceedings of the 5th ACM SIGPLAN International Symposium on Machine Programming*, MAPS 2021, pp. 1–8, New York, NY, USA, 2021. Association for Computing Machinery.

[16] B. Berabi et al. Tfix: Learning to fix coding errors with a text-to-text transformer. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 780–791. PMLR, 18–24 Jul 2021.

[17] Z. Li et al. Automating code review activities by large-scale pre-training. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, pp. 1035–1047, New York, NY, USA, 2022. Association for Computing Machinery.

[18] P. Thongtanunam et al. Autotransform: Automated code transformation to support modern code review process. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, pp. 237–248, 2022.

[19] R. Tufano et al. Using pre-trained models to boost code review automation. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, pp. 2291–2302, New York, NY, USA, 2022. Association for Computing Machinery.

[20] A. Fan et al. Large language models for software engineering: Survey and open problems. In *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*, pp. 31–53, Los Alamitos, CA, USA, may 2023. IEEE Computer Society.

[21] X. Huang et al. Understanding the planning of llm agents: A survey, 2024.

[22] R. Kaur et al. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97:101804, 2023.

[23] N. Mahmud et al. Resa: An ontology-based requirement specification language tailored to automotive systems. In *10th IEEE International Symposium on Industrial Embedded Systems (SIES)*, pp. 1–10, 2015.

[24] N. Mahmud et al. Resa tool: Structured requirements specification and sat-based consistency-checking. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1737–1746, 2016.

[25] N. Mahmud et al. Specification and semantic analysis of embedded systems requirements: From description logic to temporal logic. In *Software Engineering and Formal Methods (SEFM)*, pp. 332–348, Cham, 2017. Springer.

[26] M. E. Salari et al. An experiment in requirements engineering and testing using EARS notation for PLC systems. In *IEEE International Conference on Software Testing, Verification and Validation, ICST 2023 - Workshops, Dublin, Ireland, April 16-20, 2023*, pp. 10–17. IEEE, 2023.

[27] H. Bruneliere et al. AIDOaRt: AI-augmented Automation for DevOps, a model-based framework for continuous development in Cyber–Physical Systems. *Microprocessors and Microsystems*, 94:104672, October 2022.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 159–168 DOI: 10.15439/2024F7168 ISSN 2300-5963 ACSIS, Vol. 41

Psychological Needs as Credible Song Signals: Testing Large Language Models to Annotate Lyrics

E. Cho Smith, Yinxuan Wang, and Eric T. Matson
0009-0008-7817-690X
0009-0003-8740-4912
0000-0001-9200-4903
Computer and Information Technology,
Purdue University,
West Lafayette, Indiana 47906-3603, USA
Email: {wilso287, wang3910, ematson}@purdue.edu

Abstract—**Our preliminary study presents a new perspective in music information retrieval by investigating how contemporary song-making and listening emulate our innate responses, similar to the primal vocalizations of primates, drawing from music's origins as credible signaling. The diversity of musical expressions within a single culture suggests that it arises from group dynamics reflecting individuals' psychological fitness. Derived from the temporal need-threat framework of ostracism—an evolutionarily stable strategy to influence individuals in a group, we argue that individual differences in song-making and listening can be reduced to songs' lyrical expressions in terms of four basic psychological needs: self-esteem, self-control, seeking to belong, and seeking recognition. We propose a four-binary-decision model to classify English song lyrics for hierarchically organizing the variations of musical expressions. Annotating 260 English song lyrics using ChatGPT-4s with human validation and fine-tuning GPT-3.5-turbo to develop an automated classifier have identified some limitations in current large language models.**

Index Terms—**psychological needs recognition, music information retrieval, large language models**

I. INTRODUCTION

AMID a lively school atmosphere, a solitary individual with earbuds sits apart, symbolizing an invisible barrier separating them from others, with only the sound of music accompanying their solitude. This hypothetical person finds refuge in a musical asylum [1] from ostracism, which, as defined by Oxford Languages, means the temporary banishment from a city-state decided by popular vote in ancient Greece. A scenario often portrayed in media mirrors the real-life tendency of young individuals to use music as a coping mechanism in socially exclusive contexts [2], [3], [4]. Unlike explicit rejection, ostracism involves being ignored by groups and dyads[5], [6] through actions such as avoiding eye contact, using the silent treatment, and withholding information [7]. While ostracism does not result in immediate physical harm, its prolonged effects can profoundly impact our psychological well-being, leading to feelings of alienation, depression, helplessness, and unworthiness, which are often referred to as "social death" [5]. Although research on the relationship between ostracism and music listening is sparse, one study exemplifies how music can counteract ostracism's adverse effects [3]. Specifically, among "metal-heads," music promotes a shared sense of identity and cohesion, shielding them from the despair associated with ostracism [3]. Music's psychological function can become more salient during emotionally challenging circumstances like ostracism, one of the most pervasive forms of social exclusion among young individuals in their critical phases of identity formation and social adaptation. The hypersensitivity to ostracism [8], particularly, in this age group makes them susceptible to its negative impacts, necessitating interventions and earlier detection.

Listening to "song lyrics" resonates with listeners, serving multiple functions such as regulating emotions [9], [10], [11], evoking nostalgic memories [12], reflecting one's identity [13], [3], fostering social bonds [14], functioning as cultural markers [15] and many others. Research with 834 participants identifies 129 unique music functions, projecting to three main dimensions: arousal regulation, self-awareness, and social connections [16]. Yet, internal functions are far more relevant than external, challenging music's origin as social bonding [16]. Another study emphasizes music's psychological functions among adolescents in fulfilling needs including self-regulation, sense of agency, and belonging regardless of an individual's conscious recognition of music's influence [13].

To date, despite the noteworthy role of songs in expressing psychological needs when coping with social exclusion [2], [3], [4], current Music Information Retrieval (MIR) research has not fully appreciated the lyrical content of songs that listeners immerse themselves in and sing along to. This oversight limits the understanding of how contemporary songs function as a medium for the vicarious vocalization of listeners' psychological needs. Switching the vantage point from recognizing musical emotions, our study proposes examining psychological needs closely related to musical emotions as contextual and semantic motivations. The proposed shift provides a more objective representation of listeners' states, beyond subjective emotional experiences.

To develop a proof of concept for our approach to automating Psychological Needs Recognition (PNR) expressed in songs, our preliminary study explores English song lyrics to annotate the binary states of four basic psychological needs: self-esteem, self-control, seeking to belong, and seeking

©2024, PTI

recognition. These needs are derived from the temporal need-threat framework of ostracism [17], a quintessential group dynamic prevalent in all modern societies [18]. Following the framework, we suggest that threats to these fundamental needs can activate music's primeval functions as non-random signals, motivating individuals to create or listen to songs (vicarious vocalization) that express the critical states of their psychological needs. We further elaborate on theoretical and empirical grounds for our conceptual model in the Related Work section to elucidate our proposed new perspective in MIR, which can organize popular songs hierarchically, from the root node to fundamental psychological needs, taking precedence over musical emotions and styles.

II. RELATED WORK

Our study builds on the premise that "listening to songs," differs from hearing music as ambient sounds, is an intentional communication, deeper and more intrinsic than the tangible outcomes we may recall afterward. Following Mehr et al., we speculate music's evolutionary origins in primeval functions such as "territorial advertisements and contact calls, two types of vocal signals that are widespread in primates and other species" [19, p. 140]. The credible signaling hypothesis explains that natural selection refines human abilities to create and appreciate music's rhythms for inter-group alliance/intra-group cohesion and melodies for caregiving and seeking [20]. The notion supports our premise that music's proximate functions stem from its evolutionary origins to communicate "credible signals" of individuals' fitness about group dynamics. We suggest our song-making and listening in today's music consumption signals the resonance of individuals seeking fundamental psychological needs in flux.

A. Rethinking Musical Emotions as Signaling and Signals

According to signaling theory in evolutionary biology, signaling involves one party (the sender) transmitting information to another party (the receiver) through specific, non-random actions or symbols [21]. These signals act as credible and honest indicators aligning with the mutual interests of both senders and receivers [21]. In song-making and listening, musical emotions, constructed as the power of music features that evoke intended emotion among listeners [22], act as both signaling and signals. Musicians communicate their internal states through compositions and performances, while the audience perceives these musical emotions as signals, evoking specific emotional responses and influencing their internal states. Effective musical communication occurs when the internal states of both parties align, often described as "emotional resonance." Although perfect alignment is rare due to unintentional noise or missing contexts in the signals, we argue, musical emotions can intentionally be aligned by storytelling to create a shared emotional resonance between creators and listeners of song lyrics.

Emotions are not ends but means in internal processes, serving as internal signals (not verifiable) for individuals to change or maintain their states at any given time. Labeling musical emotions risks us falling into the complexity of emotions [23], which has muddled the representation of musical emotions for consensus. Musical emotions are modeled as discrete [24], [25], two-dimensional [26], [27], or three-dimensional [28] constructs. Additionally, a recent study suggests that emotions are higher dimensional, semantic spaces that are neither discrete nor simplified to two or three dimensions [29]. From the 1970s into the 21st century, theorists like Paul Ekman and Carroll Izard assert the existence of universal, primary, distinct emotions fulfilling adaptive roles as basic emotions [24]. Izard's differential emotions theory extends the basic emotions concept, advocating for a collection of biologically embedded emotions, such as fear, anger, joy, sadness, disgust, and surprise [25] for they are evolutionarily honed for our survival [24]. While disputing the recognition of musical emotions in MIR, we have overlooked the central psychological facet of emotions, "feelings" and warrants attention for connecting cognition and emotion [30].

When we rethink musical emotions as signaling and signals, it makes sense to move beyond conventional labels since emotional experiences are subjective and internal, making them difficult and costly to verify in affective computing [31]. Instead of labeling musical emotions, we should focus on internal signals that hold meaning for both the creators and listeners of popular songs in their contexts, addressing the daily psychological needs of living humans. Our proposed perspective emphasizes personal and unique connections within individuals and variations within a single society, arising when people create and listen to popular songs. This approach acknowledges the potentially shareable but individualized nature of musical experiences. By proposing psychological needs as credible song content, we move away from labeling and toward understanding musical emotions that value internal meaning over naming. Moreover, the limited contextual understanding of musical emotions as signals by listeners has led to more costly approaches in affective computing, such as analyzing listeners' physiological signals [32] and ambient noises in their environments, raising concerns about data privacy and protection [33].

B. Compiling Credible Signals, Mirror Neurons and Empathy

Music's evolutionary roots can be traced back to basic functions such as "territorial advertisements and contact calls," which are common vocal signals among primates and other species [19]. The credible signaling hypothesis suggests that natural selection has refined human abilities to produce and appreciate musical rhythms, which help form inter-group alliances and enhance intra-group cohesion and melodies, which are important for infant caregiving and seeking [19]. As the universality of proto-human signals has diversified over time, the value of credible signaling persists. Contemporary humans continue to create music to exchange musical signals for communication. We assert that the modern functions of popular songs, particularly those using human voices, reflect their ancestral functions. Human songs serve as a means to communicate "credible, honest signals" about individuals' fitness within group dynamics, the universal norms of any human civilization. While there exists a degree of intentionality that separates music from language [34], research suggests that music, as a form of intentional communication within group dynamics, can assist listeners in achieving self-awareness [35], self-regulation [13], a sense of agency [13], and belonging [13], [36], thereby contributing to psychological well-being [37].

Mirror neurons enable individuals to grasp the meaning and intentions behind actions by internally mirroring those actions in their own brains [38]. Found in specific regions of the monkey brain, these neurons fire both when an action is performed and when it is observed, aiding in the prediction and understanding of others' intentions [39]. In humans, the mirror neuron system is activated by visual and auditory signals, for comprehending others' actions and emotions [38]. This neural mechanism is thought to support higher cognitive functions like empathy [40] and the shared understanding necessary for speech perception [41]. For example, individuals with autism are impaired in emotional expressions and social communication. Research suggests their impaired functions are linked to the human mirror neuron system and engaging their impaired regions through music-making activities enables many children with autism to enhance their understanding of others' facial expressions of emotions to improve their communication and social skills [42], [43].

In summary, we support that the origin of music lies in its universal function as credible, non-random signals that engage our ancestral brains, activating our mirror neurons to share intentions. These intentions are linked to fundamental psychological needs, especially in today's material affluence. Thus, contemporary song-making and listening should systematically aim to decipher psychological needs signals as musical intentions. We address the need for computational models that can automate the detection of psychological needs expressed in song signals. By doing so, we can help individuals cope with the increasingly automated (no human in the loop) music platforms of today.

C. Reasoning Basic Psychological Needs as Song Signals

Feelings are visceral sensations tied to emotions, linking to motivation and aiding in psychological processes such as individualization [30]. Feelings organize cognitive processes and guide adaptive actions [30]. We argue that the link between feelings and adaptive actions forms individuals' strategies to meet psychological needs. However, experiencing emotions does not guarantee their conscious recognition or expression, which depends on their intensity and our language skills [30]. We argue, therefore, impacting our capability to have musical expressions to convey our emotional states, as we often struggle to articulate emotions while feeling in them. Aside from the disagreement in representations of emotions' dimensions, Plutchik's psychoevolutionary theory of emotions asserts that certain emotions are primary and fundamental, as primary colors, with more complex emotions evolving from the basic primary emotions such as fear, anger, and joy [44]. The primary emotions are evident across different evolutionary stages and form the core of adaptive survival responses [44].

Although the direct applications of the adaptive paths are not yet fully validated, the role of musical emotions as adaptive signals provides a theoretical framework for modeling musical expressions as individualized, adaptive responses that vary within a society, the fundamental unit of our evolution as social beings. A book chapter on self-esteem speculates its construction as "a means of interpreting mood, which encourages and inhibits conduct in various situations. Mood is a response to positive and negative experiences; self-esteem is a construction of mood fitted to a culture and its themes" [45, p. 310]. From the earliest recordings of Greek history to today's digital era, ostracism, the deliberate or indiscriminate disregard of individuals by others in a group activates our threat responses elicited by painful feelings to fortify basic psychological needs[5].

Williams' temporal need-threat framework [17] outlines ostracism experiences into immediate-reflexive reactions, intermediate-reflective copings, and prolonged-chronic stages. The immediate-reflexive stage involves the abrupt, universal experience of painful threats to four basic psychological needs: self-esteem, self-control, seeking to belong, and seeking recognition [17]. The painful sensations trigger a range of negative emotions, including pain, anger, and sadness [18]. The intermediate-reflective stage is characterized by contextually influenced, individualized strategies to restore the threatened needs, such as seeking belongingness, others' recognition [17], or solitude [46]. Adaptive responses in the intermediate stage include tend-and-befriend [47], fight, freeze, or flight reactions [5], which parallel essential neurophysiological states described as core affect (positivity, negativity, energy, and fatigue) that influence our reflexes, perception, and cognition [48]. Prolonged-chronic ostracism gradually depletes the resources of those affected, leading to feelings of depression, alienation, and unworthiness [17]. At the immediate-reflexive stage of experiencing ostracism, we highlight the paths of painful feelings as visceral responses before individuals are emotionally experienced by labeling the sensations as anger or sadness in self-reports. Moving into the intermediate-reflective coping stage, emotional responses such as sadness, anger, or hurt feelings, along with cognitive appraisals of ostracism, work to mitigate its adverse effects within individuals [49], which is where, we argue, surface individuals' variations of musical expressions chosen to make or listen to songs.

The temporal need-threat framework is invaluable for establishing psychological needs as credible indicators in songs, reflecting adaptive responses of individuals within normative group dynamics. Studies have shown that music can help individuals cope with social exclusion [3], [4] and regulate negative emotions during the COVID-19 lockdown [50], [51] when most industrialized countries required social distancing measures. The framework explains group dynamics and supports the concept of adaptive individualization by addressing the self-regulation of psychological needs as motivations. While many emotion theories implicitly suggest that emotions act as motivators, Williams' framework explicitly details how ostracism activates inherent motivations. These psychological needs drive individuals not only to alleviate negative emotions but also to fortify their impaired psychological needs, which can be replenished by interacting with a group they feel belong to. By understanding these motivations, today's music platforms can identify musical emotions related to psychological needs as credible signals to assist listeners in discovering therapeutic songs.

D. Transfer Learning by Large Language Models

Transfer learning is a technique where a computational model trained for a particular task is reused as the starting point for another task, applying the knowledge (e.g., pre-trained weights and biases of deep neural networks) gained from tasks to the following tasks that are deemed similar. Presented in "Attention is All You Need" [52], the Transformer architecture, the foundational structure of Large Language Models (LLMs), marks a departure from sequence-based deep learning architectures. Unlike its predecessors, Transformers utilize attention mechanisms that simultaneously compute the entire input sequence [52]. The attention mechanism enables the model to focus on various parts of the input sequence when predicting an output, greatly enhancing its contextual understanding [52]. The introduction of the Transformer architecture has led to the development of GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and other models. These models have set new standards in language translation, question-answering, and text generation [53]. By this "somewhat generalized intelligence," transformers are used in emotion detection and analysis from texts, with their superior understanding of contexts[54], [55], [56]. In short, Transformers, the basis of popular LLMs such as GPT and BERT, pave the way for transfer learning in machines' natural language processing (NLP).

III. METHODOLOGY

A. Psychological Needs as Credible Song Signals Model

As shown in Fig. 1, we formalize the variations in terms of four basic psychological needs using binary decision points for simplicity. There are sixteen classes of lyrical expressions, based on whether the subjects of the lyrics (narrators or protagonists) express the states of High Self-Esteem (HSE), Low Self-Esteem (LSE), High Self-Control (HSC), Low Self-Control (LSC), Positive Seeking to Belong (PSB), Negative Seeking to Belong (NSB), Positive Seeking Recognition (PSR), and Negative Seeking Recognition (NSR). The classes are enumerated as 0:[HSE, HSC, PSB, NSR], 1:[HSE, HSC, PSB, PSR], 2:[HSE, HSC, NSB, PSR], 3:[HSE, HSC, NSB, NSR], 4:[HSE, LSC, PSB, NSR], 5:[HSE, LSC, PSB, PSR], 6:[HSE, LSC, NSB, PSR], 7: [HSE, LSC, NSB, NSR], 8:[LSE, LSC, PSB, NSR], 9:[LSE, LSC, PSB, PSR], 10:[LSE, LSC, NSB, PSR], 11:[LSE, LSC, NSB, NSR], 12:[LSE, HSC, PSB, NSR], 13:[LSE, HSC, PSB, PSR], 14:[LSE, HSC, NSB, PSR], and 15:[LSE, HSC, NSB, NSR].

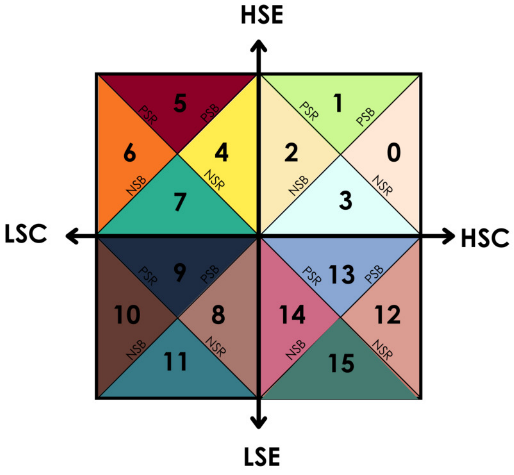


Fig. 1. The sixteen classes are based on four binary states of psychological needs: High Self-Esteem (HSE), Low Self-Esteem (LSE), High Self-Control (HSC), Low Self-Control (LSC), Positive Seeking to Belong (PSB), Negative Seeking to Belong (NSB), Positive Seeking Recognition (PSR), and Negative Seeking Recognition (NSR). The classes are enumerated using the first letter of each sub-dimension as 0:HHPN, 1:HHPP, 2:HHNP, 3: HHNN, 4:HLPN, 5:HLPP, 6:HLNP, 7: HLNN, 8:LLPN, 9:LLPP, 10:LLNP, 11:LLNN, 12:LHPN, 13:LHPP, 14:LHNP, and 15:LHNN.

B. Leveraging Model Distillation to Annotate English Lyrics

We utilize GPT-4 and GPT-4o, the fourth-generation GPT models by OpenAI. The latter, known as "GPT-4-turbo", is an optimized version of GPT-4 designed for greater efficiency and performance. To produce labeled data quickly, our methodology involves model distillation, where larger, pre-trained models such as GPT-4 and GPT-4o are used to generate labels for psychological needs expressed in English songs. Traditionally, labels in Music Emotion Recognition (MER) come from manual annotation processes, where experts or hired individuals assign labels based on predefined emotional categories. These labels require extensive human intelligence, involving subjective interpretations that can be inconsistent and prone to individual biases and lapses in attention. Due to its annotation challenges [57], [58], social tags online are often used to increase efficiency. However, in our case, there are no existing social tags for psychological needs expressed in song lyrics to utilize through crawling techniques. By leveraging generative LLMs in NLP, we streamline the rapid labeling process while ensuring the accuracy of machines by human-in-the-loop validation for testing our proposed computational model. Based on the demonstrated feasibility of using ChatGPT for annotating English text data in various topics [59], [60], [61] and increase the human and computer interaction more naturally, we employ ChatGPT 4.0 and 4o to automate the process of annotating psychological needs expressed in English song lyrics. We test whether the complexity of English lyrics can be generalized to the binary states of four psychological needs: self-esteem, self-control, seeking to belong, and seeking recognition.

C. Prompt Engineering and Fine-Tuning of LLMs

To expedite the automation of annotation, we turn to prompt engineering. This involves crafting precise instructions along with English lyrics as inputs for ChatGPT-4 and 4o, aiming to generate the most accurate labels that categorize English lyrics into sixteen classes as shown in Fig. 1. The quality of outputs produced by generative LLMs is highly dependent on the quality of inputs, including data with instructions and labels. Given that existing strategies are quite general, we address the novel challenge of annotating psychological needs in English song lyrics using ChatGPT-4 and 4o, testing our instructions for generating labels based on the proposed model Fig. 1.

A new paradigm called "prompt-based learning" modifies prompts—inputs given to generative AI to guide their responses [62]. Prompts can be questions, statements, or commands provided by users to generate the AI's outputs, utilizing zero-shot and few-shot learning [62]. Zero-shot learning enables an AI model to classify objects or concepts it has not seen or learned before and it has been shown to produce sentiment time series closely matching those from fine-tuned models, though they slightly underestimate negative examples [63]. The need for LLMs to quickly adapt to various semantic categories with minimal training has led to the development of n-shot learning, including few-shot learning and one-shot learning [64]. Few-shot learning frequently utilizes transfer learning and meta-learning strategies to train pre-trained models to identify new classes using only a few labeled data, or just one labeled example in one-shot learning [64].

According to the OpenAI API documentation, the recommended approach is to begin with prompt engineering, then proceed to fine-tuning if necessary, and refine the fine-tuning process to minimize training and validation losses. Hence, we start by employing prompt engineering, using zero-shot (no-example) prompts with ChatGPT-4 and 4o to leverage their state-of-the-art performance in labeling 16 psychological needs classes in English lyrics. The high performance of ChatGPT-4 and 4o in annotating texts is demonstrated and compared to human annotators for accuracy in a study [61]. Next, we use few-shot (some examples) prompts, providing more detailed instructions and annotated lyrics. Finally, we fine-tune GPT-3.5 as the last step in our model distillation process.

IV. EXPERIMENTS AND RESULTS

A. Curated Dataset of 260 English Song Lyrics

We built our lyrics dataset from the perspective of a hypothetical user. Imagine a user experiencing negative emotions due to ostracism and listening to music to regulate their mood and restore their sense of psychological well-being while cognitively appraising their situation. Some of the songs in our curated dataset were collected from Reddit conversations where users shared song titles and artists while experiencing episodes of ostracism at school, work, or within their families. Based on this scenario, we defined ostracism proxy experiences, such as the loss of close relationships and significant life events that trigger innate threat responses, including breakups, bereavement, and social distancing during the COVID-19 lockdown. We collected lyrics based on keyword searches, such as songs listened to after breakups, during the pandemic, or while grieving. Additionally, about 20% of songs in the dataset were chosen to reflect experiences from the authors' volatile youth periods, identified by asking for songs listened to during our younger years without explicitly priming for ostracism experiences. To maintain the variance of musical expressions in the dataset, 100 English songs were randomly selected from the Music4All dataset [65], ensuring similarity to the within-class base rate.

B. ChatGPT-4 Zero-shot Annotation with Human in the Loop

For approximately six months, we have annotated lyrics using ChatGPT-4 and its recently optimized version, 4o. Due to human attention issues, we tested only a handful of lyrics per day to train ChatGPT-4s to perform as lyrics annotators using the following instructions:

> As a lyrics annotator, your task is to categorize the subjects' (the narrators or protagonists) states expressed or closely assumed in lyrics into one of the sixteen combinations based on four binary sub-dimensions and they are:
>
> 1) The subjects can either have High Self-Esteem (HSE) or Low Self-Esteem (LSE).
> 2) The subjects can either have High Self-Control (HSC) or Low Self-Control (LSC).
> 3) The subjects can either have Seeking to Belong (PSB) or Not Seeking to Belong (NSB).
> 4) The subjects can either have Seeking Recognition (PSR) or Not Seeking Recognition (NSR).
>
> Therefore, when you classify each lyrics, you must determine the either-or sub-dimensions in four dimensions: self-esteem, self-control, seeking to belong, and seeking recognition.

A song titled "Creep" by Radiohead on Genius, narrates perhaps one of the lowest levels of self-esteem, describing oneself as a creep. The following excerpt is how ChatGPT-4 annotates the song in four dimensions. In the Self-Esteem dimension: The singer repeatedly labels themselves as a "creep" and a "weirdo" indicating feelings of self-deprecation classified as Low Self-Esteem (LSE). In the Self-Control dimension: There's a longing for control and a perfect self that suggests current dissatisfaction and lack of control over one's life and self-perception indicating Low Self-Control (LSC). In Seeking to Belong dimension: Although the singer expresses feeling out of place "I don't belong here," there isn't an active search to belong; rather, there's a resignation to their perceived outsider status supporting Not Seeking to Belong (NSB). In the Seeking Recognition dimension: The focus is not on seeking validation or attention from others but rather on an internal struggle with their own identity and worth suggesting Not Seeking Recognition (NSR). ChatGPT-4 determines the classification as LSE, LSC, NSB, and NSR, which corresponds to Class 11.

Another song titled, "So What" by P!nk on Genius, displays high self-esteem describing oneself as a rock star. The following excerpt is the analysis of ChatGPT-4 in four dimensions. In the Self-Esteem dimension: The refrain "I am a rock star" suggests High Self-Esteem (HSE), implying confidence and self-worth. In the Self-Control dimension: The lines "I wanna start a fight" and "I'm gonna drink my money" could be seen as impulsive, pointing to Low Self-Control (LSC). In the Seeking to Belong dimension: The repeated statement "I don't need you" and the celebratory tone about being single indicate Not Seeking to Belong (NSB). In the Seeking Recognition dimension: The strong assertions of identity, a rock star, and existence in the face of adversities imply a demand for recognition of their independence and resistance suggesting seeking recognition (PSR). Therefore, ChatGPT-4 determines the classification as HSE, LSC, NSB, and PSR, which corresponds to Class 6.

TABLE I
DISTRIBUTION OF SIXTEEN PSYCHOLOGICAL NEEDS CLASSES

Annotated by ChatGPT 4.0 & 4o with Human Feedback																
class	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
count	24	7	7	36	7	18	12	14	28	33	18	18	9	9	8	12

As a result, the distribution of 260 lyrics across the 16 classes, before splitting into training and testing datasets, is shown in Table 1. To increase the accuracy of annotations, we implement three rounds of refining prompts and annotating each set of lyrics with ChatGPT-4 and 4o using human validation and feedback (human in the loop). This process enables us to settle on the most feasible classification. Class 3 and Class 9 are the two most common classes, with 36 and 33 instances, respectively. When internal resources (such as self-esteem and self-control) are higher in Class 3 [HSE, HSC, NSB, NSR], the subjects of the lyrics do not express a need to belong or to be validated by others. Conversely, when internal resources are lower in Class 9 [LSE, LSC, PSB, PSR], the subjects of the lyrics signal a need to belong and seek validation. Considering 160 songs listened to during ostracism proxy experiences (such as breakups, grief, and social distancing during the pandemic) or when feeling ostracized (songs discussed in Reddit conversations), it is feasible that individuals feeling threatened by these experiences might listen to songs that either express their internal psychological needs or their external psychological needs. These songs signal the subjects' psychological fitness (expressed in lyrics) and resonate with listeners by figuratively allowing themselves to step into the signalers' shoes, thereby activating their mirror neurons at full throttle [38].

C. GPT-4o Assistant and Gemini 1.5 Few-shot Annotation

Using human feedback in the loop, ChatGPT-4 and 4o have shown a tendency to forget previous instructions over prolonged use, requiring us to start over with prompting. Additionally, we have encountered inconsistent answer formats and irrelevant analyses. To improve the effectiveness of generating large datasets with generative AI, we enter the next phase of model distillation to train more efficient models to annotate lyrics. We use two different instructions and datasets:

- **Trial 1:** Instructions given to ChatGPT-4s with 80 lyrics.
- **Trial 2:** Detailed Instructions with 80 lyrics that are sliced into [verse] + [chorus] for data augmentation.
- **Trial 3:** Detailed Instructions with 80 lyrics.

We utilize Gemini 1.5 by Google and GPT-4o Assistant by OpenAI, setting the randomness of outputs (temperature) to 0.2 and the probability mass (top p) to 0.1 to produce more deterministic results. These hyperparameters control the randomness of the model's outputs, with lower values (0.2 and 0.1 respectively) focusing token selection on the most probable options, limiting it to the top 10%.

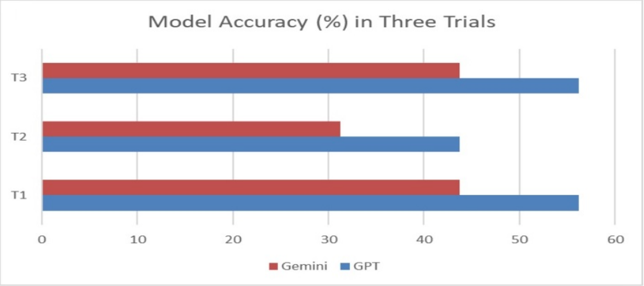


Fig. 2. Performance Comparison: GPT-4o and Gemini 1.5

As shown in Fig. 2, we use the simple average percentage for accuracy to evaluate the performance of two few-shot learning models. GPT-4o's accuracy was higher than Gemini 1.5's in all trials, with scores of 56.25%, 43.75%, and 56.75% compared to 43.75%, 31.25%, and 43.76%. Both Gemini 1.5 and GPT-4o performed worse in Trial 2 when we used detailed instructions and data augmentation techniques with split lyrics. It makes sense for lyrics to be categorized as one class with four dimensions, requiring all the lyrics (not splits). The highest score for GPT-4o in Trials 1 and 3 was 56.75%. Given that we have 16 classes, the accuracy indicates that while the model performs much better than random guessing (which would have an accuracy of 6.25% for 16 classes), there is certainly substantial room for improvement.

TABLE II
ACCURACY IN FOUR PSYCHOLOGICAL NEEDS DIMENSIONS

	GPT-4o	**Gemini 1.5**
Self-Esteem (HSE/LSE)	93.75%	93.75%
Self-Control (HSC/LSC)	87.50%	81.75%
Seeking to Belong (PSB/NSB)	81.25%	81.25%
Seeking Recognition (PSR/NSR)	56.25%	50.00%
Overall Accuracy (%)	79.69%	76.56%

As shown in Table 2, based on the analysis of four-dimensional accuracy in Trial 3, both GPT-4o and Gemini 1.5 achieve an impressive 93.75% accuracy in the Self-Esteem

(HSE/LSE) dimension. However, GPT-4o surpasses Gemini 1.5 in the Self-Control (HSC/LSC) dimension, with accuracy of 87.50% and 81.75%, respectively. The models perform equally moderately in the Seeking to Belong (PSB/NSB) dimension, each scoring 81.25%. However, in the Seeking Recognition (PSR/NSR) dimension, GPT-4o demonstrates a higher accuracy of 56.25% compared to Gemini 1.5's 50.00%, marking closer and exactly to random guessing. Overall, GPT-4o has a marginally higher overall accuracy at 79.69% compared to Gemini 1.5's 76.56%. The declining pattern in accuracy suggests that both models may be overfitting, as learning appears to occur in the order of Self-Esteem, Self-Control, Seeking to Belong, and Seeking Recognition dimension.

The declining accuracy pattern in both models can have potential reasons. Firstly, the dimensions might be ordered by increasing complexity, with Self-Esteem (HSE/LSE) being the simplest to classify and Seeking Recognition (PSR/NSR) the most challenging. Secondly, the models might have limitations in understanding the nuances required for the later dimensions, such as the subtle cues needed to recognize Seeking Recognition (PSR/NSR) due to working memory limitations. Thirdly, because the models were trained on a limited dataset, they might overfit familiar patterns like Self-Esteem but struggle with less common patterns or highly varying patterns in Seeking Recognition. Lastly, because the models were generated and learned reasons sequentially, earlier tasks might receive more optimization, causing progressively lower performance on later tasks.

To improve accuracy, after Trial 1, we developed conceptually more detailed instructions for Trials 2 and 3. We defined the four dimensions (self-esteem, self-control, seeking to belong, and seeking recognition) and provided behavioral inferences to anchor these definitions. For example, we added 33 detailed behavioral anchors to the High Self-Esteem subdimension, with similar numbers of anchors for all other subdimensions. Here are abbreviated versions of the High Self-Esteem (HSE) sub-dimension behavioral anchors included in Trial 2 and 3 prompts (some of them are shortened for brevity):

- Understand one's self-worth and the value of one's life.
- Express a confident, self-assured positive outlook.
- Express one's contentment and a sense of fulfillment.
- Show one's confidence in vocalizing their feelings.
- Recognize their own flaws but not let the world bring them down.
- Express a strong sense of self and a willingness to support another person.
- Recognize their own needs and desires, and the refusal to settle for less.
- Express an inherent belief in love and its transformative power.

Surprisingly, these detailed anchors did not affect the performance of both GPT-4o and Gemini 1.5. Therefore, we decided to abandon the detailed instructions and focus on increasing the example size per class. Due to the imbalanced dataset of 260 lyrics, we tested few-shot learning with 80 lyrics, balanced with 5 lyrics in each class. At the time of trials, increasing the number of examples was limited due to token limits. With this predicament, we venture the data augmentation technique to increase the number of examples with splitting lyrics. Not surprisingly, as mentioned earlier, splitting lyrics into one verse and a chorus (repeated in many splits depending on the number of verses in the original) did not improve model performance. This intuitively makes sense because the repeated annotations determined by four dimensions might not be relevant among splits, potentially increasing noise.

D. GPT-3.5-Turbo-0125 (Recommended) Fine-tuning

With the results of three few-shot trials, we move to the next step: fine-tuning a smaller version of GPT to finalize our model distillation. Although the model sizes are not publicly available, we start with GPT-4 and 4o, the updated version of GPT-3 consisting of over 175 billion parameters. We test the performance of these larger GPT-4 models in generating labels for our sixteen psychological needs classes. After the authors validate the labels, we harvest 260 lyrics with 16 class labels. Out of the total of 260 imbalanced datasets (refer to the distribution in Table 1 in subsection B), we use sets of 80 lyrics (a perfectly balanced set) and 109 lyrics (an approximately balanced set with 7 lyrics per class, except for classes 1, 2, and 3, which have 6 lyrics each). The following descriptions are based on reports from the GPT-3.5-turbo-0125 (recommended) version after fine-tuning it in two trials:

- **Trial 4 (80 lyrics):** 3,251,361 tokens, Epochs=3, Batch size=1, LR multiplier=2, Train loss=0.4324, Full validation loss=0.9579.
- **Trial 5 (109 lyrics):** 4,968,234 tokens, Epochs=3, Batch size=1, LR multiplier=2, Training loss=0.7395, Full validation loss=0.6827.

In Trial 4, we fine-tune the GPT-3.5-turbo-0125 model using a dataset consisting of 80 lyrics, processing a total of 3,251,361 tokens. The training process spans 3 epochs with a batch size of 1, allowing for individual processing and updating of each lyric. We employ a learning rate multiplier of 2 to enhance the base learning rate and expedite convergence. The training loss achieved is 0.4324, reflecting a relatively low error on the training data. However, the validation loss is 0.9579, indicating that although the model performs well on the training set, it shows signs of overfitting, as evidenced by the higher error on the validation set. This discrepancy underscores the need for further tuning to improve generalization to new data. For the subsequent trial, we increase the dataset size to 109 lyrics to maintain dataset balance and potentially enhance the model's performance.

In Trial 5, we fine-tune the GPT-3.5-turbo-0125 model using a dataset of 109 lyrics, resulting in the processing of 4,968,234 tokens. As in Trial 4, the training is conducted over 3 epochs with a batch size of 1 and a learning rate multiplier of 2. The training loss is 0.7395, higher than in Trial 4. However, the validation loss is notably lower at 0.6827, signifying improved generalization and performance on unseen data compared to Trial 4. The improvement suggests that the larger dataset in Trial 5 contributes to a more robust model. Despite the improvement, significant overfitting remains, alarming that the fine-tuned model is still far from optimal.

Consequently, we use a test set consisting of 16 hand-picked lyrics, chosen for their clear quality and distinct class differences, across five trials, including three few-shot trials, to test the accuracy of GPT-4o and Gemini 1.5. Somewhat unexpectedly, the accuracy of GPT-3.5-0125 in Trial 4 is 12.5%, nearly equivalent to a random response, even considering the smaller size of GPT-3.5 compared to GPT-4. In Trial 5, the accuracy is 31.25%, noticeably higher than in Trial 4. However, this is the same as the lowest accuracy shown in the second trial by Gemini 1.5, where we used splits of lyrics resulting in increased noise. Intuitively, the increase in the number of examples, from 80 to 109 lyrics, has improved accuracy. We are pleased to see that the difference in accuracy between using 80 lyrics and 109 lyrics is 18.75%, a hopeful increase achieved by adding two more lyrics per class, except for Classes 1, 2, and 3, which each had one additional lyrics. However, having closer to 7 examples per class falls short of the recommended minimum of 10 to 50 examples per class, according to some experimenters' anecdotal estimation commented in the OpenAI Forum (click here). In essence, the minimum number of examples depends on the specific tasks and domains. Until we test it with varying sizes of examples, we cannot estimate this with certainty. What is certain, though, is that at this time, the size of the examples matters in increasing the accuracy of the smaller model, considering the accuracy of 12.5% for the GPT-3.5-turbo-0125 model and 56.75% for the GPT-4o Assistant, both trained with 80 lyrics and slightly different interface.

The budget for all five trials cost us approximately $575.00, which is significantly less than the time and resources we would have spent training the models from scratch. Moreover, quickly obtaining the proof of concept is invaluable. However, because GPT-3.5-turbo-0125 is proprietary, fine-tuning it comes with limitations such as restricted access to closed information and less control over hyperparameters. In this preliminary study, facilitating OpenAI's fine-tuning interface to gain the proof of concept is sensible due to its user-friendliness and quick results. Nevertheless, it is advisable to switch to a more efficient pre-trained transformer model for production-level work, especially in mobile platforms.

V. DISCUSSION

In today's digital era, the internet continues to expand rapidly, reflecting the dynamics of the modern world. This digital realm encompasses various types of unfiltered big data, contributing to the digital representation of human emotional and psychological expression. Just as in the physical world, the digital world requires methods to understand and monitor public emotional and psychological health. Understanding psychological needs is crucial for analyzing the digital footprints of human emotional and psychological conditions.

The development of the proposed model to recognize psychological needs expressed in song lyrics began after I listened to a story about the life-altering moments of a teenage boy who chose to live instead of fading away like dew in the early morning. This story was shared under the title "Power of One" in Prof. Williams' ostracism course at Purdue, inspiring our project to start with the concept of a mobile agent, PO2. After taking a fatal dose of psychotropic medication, the teenage boy saw the sole, smiling face of a kind girl from his school before falling into the abyss. Her smile prompted him to call 911 and save himself. After his recovery, he left a note explaining how her smile had helped him break through his darkest hour and thanking her for smiling at him, everyone's invisible person at school. Our vision with the PO2 project is to help identify and utilize popular songs to assist young people in coping with ongoing offline and online ostracism [66]. Our framework and computational methods are designed to recognize four fundamental psychological needs, marking the first step toward materializing the mission of our PO2 mobile agent. This agent aims to alleviate the harmful effects of ostracism through the therapeutic use of popular song listening. By promoting healthier personal and social adjustment facilitated by the lyrics of recommended songs, the PO2 agent, much like a smiling face, seeks to detect and mitigate the perpetuation of ostracism's detrimental impacts.

In addition to its therapeutic potential, identifying psychological needs as song content can address data privacy and protection issues inherent in current efforts to integrate physical, physiological, and contextual signals of listeners' emotions. For example, Spotify has disclosed a patented speech emotion recognition [67] and ambient noise detection [33]. Despite Spotify's stance on upholding user privacy, this technology has raised concerns among users, artists, and activist groups regarding potential invasions of privacy and discrimination against gender-transitioning individuals [33].

Moreover, in automating music search and retrieval, identifying psychological needs in song lyrics can significantly enhance music recommendation systems by better deciphering listeners' emotional and psychological states. A key player in this domain is Music Emotion Recognition (MER) recommendation systems, which rely on accurately recognizing emotions by extracting the emotional content from music [68] and assessing the emotional states of listeners [69]. MER recommendation systems, however, face a critical drawback due to the disparity between musical emotions and listeners' emotional states. Aside from privacy concerns, these technologies aim to detect and respond to listeners' transient emotional states but fail to address their more enduring psychological needs linked to their emotional states. These latent psychological needs, which are fundamental to listeners' motivation and closely tied to their emotional experiences, align more closely with their goal-setting behaviors and are better suited as metrics for tracking listeners' intentions.

Due to its novelty, our preliminary study presents not-yet-validated speculations that are further challenged by interdisciplinary dialogues and lack substantial corroborating evidence from prior research. Moving forward from the preliminary study, we are optimistic about leveraging generative AI to develop a classifier for psychological needs as credible song signals. Our next step is to continue annotating English song lyrics using an improved and faster version of ChatGPT-4o, incorporating human-in-the-loop processes, to achieve a balanced dataset with 50 to 100 lyrics per class and to expand our final test set to a minimum of 20% of the total dataset. Using model distillation, moving forward, we aim to fine-tune more efficient open-source LLMs with this extensive dataset.

ACKNOWLEDGMENT

Thank you, Dr. Kipling Williams, for your invaluable encouragement in developing my thesis. Your passion and knowledge in ostracism research were transformative to our vision.

REFERENCES

[1] T. DeNora, *Music asylums: Wellbeing through music in everyday life*. Routledge, 2016.

[2] W. M. Randall and N. S. Rickard, “Reasons for personal music listening: A mobile experience sampling study of emotional outcomes,” *Psychology of Music*, vol. 45, no. 4, pp. 479–495, 2017. http://dx.doi.org/10.1177/0305735616666939

[3] P. Rowe and B. Guerin, “Contextualizing the mental health of metal youth: A community for social protection, identity, and musical empowerment,” *Journal of Community Psychology*, vol. 46, no. 4, pp. 429–441, 2018. http://dx.doi.org/10.1002/jcop.21949

[4] T. ter Bogt, N. Canale, M. Lenzi, A. Vieno, and R. van den Eijnden, “Sad music depresses sad adolescents: A listener’s profile,” *Psychology of Music*, vol. 49, no. 2, pp. 257–272, 2021. http://dx.doi.org/10.1177/0305735619849622

[5] K. D. Williams, “Ostracism,” *Annu. Rev. Psychol.*, vol. 58, pp. 425–452, 2007. http://dx.doi.org/10.1146/annurev.psych.58.110405.085641

[6] ——, *Ostracism: The power of silence*. Guilford Press, 2002.

[7] E. D. Wesselmann, M. R. Grzybowski, D. M. Steakley-Freeman, E. R. DeSouza, J. B. Nezlek, and K. D. Williams, “Social exclusion in everyday life,” *Social exclusion: Psychological approaches to understanding and reducing its impact*, pp. 3–23, 2016. http://dx.doi.org/10.1007/978-3-319-33033-4_1

[8] C. M. Büttner, M. Jauch, M. Marinucci, K. D. Williams, R. Greifeneder, P. Riva, and S. C. Rudert, “It will (never) stop hurting: Do repeated or chronic experiences of exclusion lead to hyper-or hyposensitive psychological responses?” *Group Processes & Intergroup Relations*, p. 13684302221140002, 2023. http://dx.doi.org/10.1177/13684302221140002

[9] G. A. Dingle, J. Hodges, and A. Kunde, “Tuned in emotion regulation program using music listening: Effectiveness for adolescents in educational settings,” *Frontiers in psychology*, vol. 7, p. 859, 2016. http://dx.doi.org/10.3389/fpsyg.2016.00859

[10] S. Uhlig, E. Jansen, and E. Scherder, ““being a bully isn’t very cool…”: Rap & sing music therapy for enhanced emotional self-regulation in an adolescent school setting–a randomized controlled trial,” *Psychology of music*, vol. 46, no. 4, pp. 568–587, 2018. http://dx.doi.org/10.1177/0305735617719154

[11] S. Garrido, M. Du Toit, and T. Meade, “Music listening and emotion regulation: Young people’s perspectives on strategies, outcomes, and intervening factors.” *Psychomusicology: Music, Mind, and Brain*, 2022. http://dx.doi.org/10.1037/pmu0000285

[12] E. Michels-Ratliff and M. Ennis, “This is your song: Using participants’ music selections to evoke nostalgia and autobiographical memories efficiently.” *Psychomusicology: Music, Mind, and Brain*, vol. 26, no. 4, p. 379, 2016. http://dx.doi.org/10.1037/pmu0000167

[13] S. Laiho, “The psychological functions of music in adolescence,” *Nordic Journal of music therapy*, vol. 13, no. 1, pp. 47–63, 2004. http://dx.doi.org/10.1080/08098130409478097

[14] G. Kreutz, “Does singing facilitate social bonding?” *Music and Medicine*, vol. 6, no. 2, pp. 51–60, 2014. http://dx.doi.org/10.47513/mmd.v6i2.180

[15] C. N. DeWall, R. S. Pond Jr, W. K. Campbell, and J. M. Twenge, “Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular us song lyrics.” *Psychology of Aesthetics, Creativity, and the Arts*, vol. 5, no. 3, p. 200, 2011. http://dx.doi.org/10.1037/a0023195

[16] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, “The psychological functions of music listening,” *Frontiers in psychology*, vol. 4, p. 511, 2013. http://dx.doi.org/10.3389/fpsyg.2013.00511

[17] K. D. Williams, “Ostracism: A temporal need-threat model,” *Advances in experimental social psychology*, vol. 41, pp. 275–314, 2009. http://dx.doi.org/10.1016/S0065-2601(08)00406-1

[18] C. H. Hartgerink, I. Van Beest, J. M. Wicherts, and K. D. Williams, “The ordinal effects of ostracism: A meta-analysis of 120 cyberball studies,” *PloS one*, vol. 10, no. 5, p. e0127002, 2015. http://dx.doi.org/10.1371/journal.pone.0127002

[19] S. A. Mehr, M. M. Krasnow, G. A. Bryant, and E. H. Hagen, “Toward a productive evolutionary understanding of music,” *The Behavioral and brain sciences*, vol. 44, p. e122, 2021. http://dx.doi.org/10.1017/s0140525x21000030

[20] ——, “Origins of music in credible signaling,” *Behavioral and Brain Sciences*, vol. 44, p. e60, 2021. http://dx.doi.org/10.1017/S0140525X20000345

[21] W. contributors, “Signalling theory,” 2023, accessed: 2024-05-20. [Online]. Available: https://en.wikipedia.org/wiki/Signalling_theory

[22] P. N. Juslin and D. Västfjäll, “Emotional responses to music: The need to consider underlying mechanisms,” *Behavioral and brain sciences*, vol. 31, no. 5, pp. 559–575, 2008. http://dx.doi.org/10.1017/s0140525x08005293

[23] T. Eerola and J. K. Vuoskoski, “A review of music and emotion studies: Approaches, emotion models, and stimuli,” *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2012. http://dx.doi.org/10.1525/mp.2012.30.3.307

[24] P. Ekman, “An argument for basic emotions,” *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. http://dx.doi.org/10.1080/02699939208411068

[25] C. E. Izard and C. E. Izard, “Differential emotions theory,” *Human emotions*, pp. 43–66, 1977. http://dx.doi.org/10.1007/978-1-4899-2209-0_3

[26] R. E. Thayer, J. R. Newman, and T. M. McClain, “Self-regulation of mood: strategies for changing a bad mood, raising energy, and reducing tension.” *Journal of personality and social psychology*, vol. 67, no. 5, p. 910, 1994. http://dx.doi.org/10.1037//0022-3514.67.5.910

[27] J. A. Russell, “A circumplex model of affect.” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980. http://dx.doi.org/10.1037/h0077714

[28] R. Plutchik, *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

[29] D. Keltner, J. A. Brooks, and A. Cowen, “Semantic space theory: Data-driven insights into basic emotions,” *Current Directions in Psychological Science*, p. 09637214221150511, 2023. http://dx.doi.org/10.1177/09637214221150511

[30] C. E. Izard, “Emotion theory and research: Highlights, unanswered questions, and emerging issues,” *Annual review of psychology*, vol. 60, pp. 1–25, 2009. http://dx.doi.org/10.1146/annurev.psych.60.110707.163539

[31] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, “A systematic review on affective computing: Emotion models, databases, and recent advances,” *Information Fusion*, vol. 83, pp. 19–52, 2022. http://dx.doi.org/10.48550/arXiv.2203.06935

[32] V. Chaturvedi, A. B. Kaur, V. Varshney, A. Garg, G. S. Chhabra, and M. Kumar, “Music mood and human emotion recognition based on physiological signals: a systematic review,” *Multimedia Systems*, vol. 28, no. 1, pp. 21–44, 2022. http://dx.doi.org/10.1007/s00530-021-00786-6

[33] A. Name, “Spotify patented emotional recognition technology to recommend songs based on users’ emotions,” 2022, accessed: 2023-10-16. [Online]. Available: https://jolt.richmond.edu/2022/01/11/spotify-patented-emotional-recognition-technology-to-recommend-\protect\penalty\z@songs-based-on-users-emotions/

[34] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. D. Friederici, “Music, language and meaning: brain signatures of semantic processing,” *Nature neuroscience*, vol. 7, no. 3, pp. 302–307, 2004. http://dx.doi.org/https://doi.org/10.1038/nn1197

[35] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, “The psychological functions of music listening,” *Frontiers in psychology*, vol. 4, p. 511, 2013. http://dx.doi.org/10.3389/fpsyg.2013.00511

[36] K. Schäfer and T. Eerola, "How listening to music and engagement with other media provide a sense of belonging: An exploratory study of social surrogacy," *Psychology of Music*, vol. 48, no. 2, pp. 232–251, 2020. http://dx.doi.org/10.1177/0305735618795036

[37] R. Granot, D. H. Spitz, B. R. Cherki, P. Loui, R. Timmers, R. S. Schaefer, J. K. Vuoskoski, R.-N. Cárdenas-Soler, J. F. Soares-Quadros Jr, S. Li *et al.*, ""help! i need somebody": music as a global resource for obtaining wellbeing goals in times of crisis," *Frontiers in Psychology*, vol. 12, p. 648013, 2021. http://dx.doi.org/10.3389/fpsyg.2021.648013

[38] I. Molnar-Szakacs and K. Overy, "Music and mirror neurons: from motion to'e'motion," *Social cognitive and affective neuroscience*, vol. 1, no. 3, pp. 235–241, 2006. http://dx.doi.org/10.1093/scan/nsl029

[39] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, vol. 308, no. 5722, pp. 662–667, 2005. http://dx.doi.org/10.1126/science.1106138

[40] L. Carr, M. Iacoboni, M.-C. Dubeau, J. C. Mazziotta, and G. L. Lenzi, "Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas," *Proceedings of the national Academy of Sciences*, vol. 100, no. 9, pp. 5497–5502, 2003. http://dx.doi.org/10.1073/pnas.0935845100

[41] A. M. Liberman and D. H. Whalen, "On the relation of speech to language," *Trends in cognitive sciences*, vol. 4, no. 5, pp. 187–196, 2000. http://dx.doi.org/https://doi.org/10.1016/s1364-6613(00)01471-6

[42] C. Y. Wan, K. Demaine, L. Zipse, A. Norton, and G. Schlaug, "From music making to speaking: Engaging the mirror neuron system in autism," *Brain research bulletin*, vol. 82, no. 3-4, pp. 161–168, 2010. http://dx.doi.org/10.1016/j.brainresbull.2010.04.010

[43] G. L. Wagener, M. Berning, A. P. Costa, G. Steffgen, and A. Melzer, "Effects of emotional music on facial emotion recognition in children with autism spectrum disorder (asd)," *Journal of Autism and Developmental Disorders*, vol. 51, pp. 3256–3265, 2021. http://dx.doi.org/10.1007/s10803-020-04781-0

[44] R. Plutchik, "A psychoevolutionary theory of emotions," 1982.

[45] J. P. Hewitt, "22 the social construction of self-esteem," *The Oxford handbook of positive psychology*, p. 309, 2020. http://dx.doi.org/10.1093/oxfordhb/9780195187243.013.0020

[46] D. Ren, E. Wesselmann, and K. D. Williams, "Evidence for another response to ostracism: Solitude seeking," *Social Psychological and Personality Science*, vol. 7, no. 3, pp. 204–212, 2016. http://dx.doi.org/10.1177/1948550615616169

[47] S. E. Taylor, L. C. Klein, B. P. Lewis, T. L. Gruenewald, R. A. Gurung, and J. A. Updegraff, "Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight." *Psychological review*, vol. 107, no. 3, p. 411, 2000. http://dx.doi.org/10.1037/0033-295x.107.3.411

[48] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003. http://dx.doi.org/10.1037/0033-295x.110.1.145

[49] E. Yaakobi, "Immediate moderating effects of attribution on ostracism distress," *Australian Psychologist*, vol. 56, no. 3, pp. 193–203, 2021. http://dx.doi.org/10.1080/00050067.2020.1829452

[50] T. Y.-C. Yeung, "Did the covid-19 pandemic trigger nostalgia? evidence of music consumption on spotify," *Evidence of Music Consumption on Spotify (August 21, 2020)*, 2020.

[51] N. Henry, D. Kayser, and H. Egermann, "Music in mood regulation and coping orientations in response to covid-19 lockdown measures within the united kingdom," *Frontiers in psychology*, vol. 12, p. 647879, 2021. http://dx.doi.org/10.3389/fpsyg.2021.647879

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. http://dx.doi.org/10.48550/arXiv.1706.03762

[53] M. O. Topal, A. Bas, and I. van Heerden, "Exploring transformers in natural language generation: Gpt, bert, and xlnet," *arXiv preprint arXiv:2102.08036*, 2021. http://dx.doi.org/10.48550/arXiv.2102.08036

[54] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, pp. 1–41, 2021. http://dx.doi.org/10.1007/s10462-021-09958-2

[55] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020. http://dx.doi.org/10.1109/ICASSP40776.2020.9053762 pp. 3507–3511.

[56] D. Cortiz, "Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra," *arXiv preprint arXiv:2104.02041*, 2021. http://dx.doi.org/10.48550/arXiv.2104.02041

[57] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021. http://dx.doi.org/10.1109/MSP.2021.3106232

[58] Y. Wang, J. Tian, D. S. Ones, R. N. Landers *et al.*, "Using natural language processing and machine learning to replace human content coders." *Psychological Methods*, 2022. http://dx.doi.org/10.1037/met0000518

[59] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, J. D. Bermeo, M. Korobeynikova, and F. Gilardi, "Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks," *arXiv preprint arXiv:2307.02179*, 2023. http://dx.doi.org/10.48550/arXiv.2307.02179

[60] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, and D. Yang, "Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation," *arXiv preprint arXiv:2310.15638*, 2023. http://dx.doi.org/10.18653/v1/2023.emnlp-main.92

[61] M. Hasanain, F. Ahmed, and F. Alam, "Large language models for propaganda span annotation," *arXiv preprint arXiv:2311.09812*, 2023. http://dx.doi.org/10.48550/arXiv.2311.09812

[62] 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. http://dx.doi.org/10.48550/arXiv.2107.13586

[63] S. Mohammadi, M. Mohammadi, S. Sinaei, A. Balador, E. Nowroozi, F. Flammini, and M. Conti, "Balancing privacy and accuracy in federated learning for speech emotion recognition," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. IEEE, 2023. http://dx.doi.org/http://dx.doi.org/10.15439/2023F444 p. 191–199.

[64] IBM, "Zero-shot learning," 2024, accessed: 2024-07-20. [Online]. Available: https://www.ibm.com/topics/zero-shot-learning

[65] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020. http://dx.doi.org/10.1109/IWSSIP48289.2020.9145170 pp. 399–404.

[66] K. D. Williams, C. K. Cheung, and W. Choi, "Cyberostracism: effects of being ignored over the internet." *Journal of personality and social psychology*, vol. 79, no. 5, p. 748, 2000. http://dx.doi.org/https://psycnet.apa.org/doi/10.1037/0022-3514.79.5.748

[67] J. Borst, L. Wehrheim, A. Niekler, and M. Burghardt, "An evaluation of a zero-shot approach to aspect-based sentiment classification in historic german stock market reports," in *Communication Papers of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 37. PTI, 2023. http://dx.doi.org/http://dx.doi.org/10.15439/2023F3725 p. 51–60.

[68] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, no. 6, p. 166335, 2022. http://dx.doi.org/http://dx.doi.org/10.1007/s11704-021-0569-4

[69] M. de Witte, A. d. S. Pinho, G.-J. Stams, X. Moonen, A. E. Bos, and S. van Hooren, "Music therapy for stress reduction: a systematic review and meta-analysis," *Health Psychology Review*, vol. 16, no. 1, pp. 134–159, 2022. http://dx.doi.org/https://doi.org/10.1080/17437199.2020.1846580

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 169–176 ISSN 2300-5963 ACSIS, Vol. 41
DOI: 10.15439/2024F6229

Integrating Artificial Intelligence-based programs into Autism Therapy: Innovations for Personalized Rehabilitation

Fabrizio Stasolla
Università Giustino Fortunato
Benevento Italy
Email: f.stasolla@unifortunato.eu
orcid: https://orcid.org/0000-0003-1626-9664

Enza Curcio
Università Giustino Fortunato
Benevento Italy
Email: e.curcio@unifortunato.eu

Anna Passaro
Università Giustino Fortunato
Benevento Italy
Email: a.passaro@unifortunato.eu

Antonio Zullo
Universitas Mercatorum
Email: a.zullo@unifortunato.eu

Mariacarla Di Gioia
Universitas Mercatorum
Email: m.digioia@unifortunato.eu

***Abstract*—Autism Spectrum Disorder (ASD) is a challenging clinical condition that requires tailored therapies to boost cognitive and social skills in those affected. Lately, artificial intelligence (AI) has shown great potential in the field of autism assessment and rehabilitation. This article explores how AI plays a crucial role in improving autism clinical conditions. Thus, smart systems for early diagnosis, personalized treatment, and continuous progress tracking were adopted. The paper looks at the difficulties and possibilities of using AI in individuals with ASD. This included concerns like safeguarding data privacy, accurately understanding behavioral cues, and developing interactive, welcoming therapy settings. More specifically, the article explored how techniques from machine learning and artificial intelligence could be woven into rehabilitation methods to enhance learning and promote independence and social inclusion for individuals with autism. This examination provided a fresh and enlightening view on how clinical approaches are evolving, showing how AI could greatly improve the lives of individuals with autism. Implications for research and clinical practice were critically discussed.**

Keywords:
Autism Spectrum Disorders, Artificial Intelligence, Rehabilitation, Machine Learning

I. INTRODUCTION

ASD IS a neurodevelopmental condition that affects communication and social interaction, often involving repetitive behaviors and reduced interests [13]. Over the years, the number of people diagnosed with ASD has risen significantly. In 2000, the prevalence was 1 in 150 children, which increased to 1 in 36 by 2017. This rise has made ASD a leading cause of disability in children, posing significant management challenges and financial burdens. In the U.S., lifetime support for an individual with intellectual disabilities can cost around 2.4 million, and about 2.2 million in the U.K., heavily impacting families and society [2]. Persons with autism spectrum disorders (ASD) may have notable difficulties, requiring tailored therapies that can improve cognitive, emotional, and social skills. Over the past few years, there has been an increasing interest in using AI to enhance autism rehabilitation. This overview investigates how AI holds potential in improving therapy results for children diagnosed with ASD [31]. The literature review emphasizes how AI technologies play a role in improving outcomes through early diagnosis, personalized treatment strategies, and ongoing progress monitoring. Important features to be considered include but are not limited to safeguarding data privacy, accurately interpreting behavioral cues, and developing interactive and inclusive therapy environments. Addressing these challenges is essential to optimize the effectiveness and ethical implementation of AI in supporting individuals with autism [10]. The article further explores how machine learning and AI methods can be incorporated into rehabilitation plans to improve learning results and encourage independence and social inclusion for individuals with ASD [1]. Integrating AI can be considered promising in improving therapeutic methods and fostering increased independence and inclusion for individuals with ASD [7]. This analysis provides insights concerning how clinical practices are evolving in autism rehabilitation, showcasing how AI has the potential to greatly enhance the lives of individuals with ASD. Integrating AI can be considered promising in improving therapeutic methods and fostering increased independence and inclusion for individuals with ASD [20]. Recently, AI has made significant strides in helping treat ASD. For example, the humanoid robot Kaspar has improved daily living skills in children with ASD by mimicking human behavior [30]. Research showed that children who participated in robot-assisted learning programs were more engaged and learned better, suggesting that AI robots could be highly beneficial in classrooms. Additionally, augmented reality smart glasses have been effective in reducing irritability, lethargy, rigid behaviors, and speech difficulties in children with autism. Studies indicate that AI technology enhances cognitive and social skills and holds great potential for rehabilitating chil-

©2024, PTI

dren and adolescents with severe intellectual disabilities [4]. The aim of this paper was to summarize and compare the current literature pertaining to rehabilitation in children and adolescents diagnosed with ASD through AI to provide an overview of effective strategies and successful improvements in both clinical practice and research. The novelty features may include the newest empirical studies on this specific topic and targeting comprehensive rehabilitative strategies. This review addressed the existing gap in the literature regarding comprehensive and contemporary analyses of AI applications in ASD rehabilitation. It summarized recent empirical studies, offering novel insights into the efficacy of AI-based interventions. By critically evaluating diverse AI technologies and discussing ethical considerations, this review provided new knowledge on practical implementation strategies. The review intended to analyze the current rehabilitation tools that exploit AI and the related therapeutic methods, in order to highlight the best strategies to promote autonomy and social inclusion for people with ASD, and lightened the burden of operators and caregivers. Furthermore, this overview, by analyzing the results of empirical studies conducted on the topic, addressed perspectives for future research, especially regarding the possibility of combining the use of AI-based programs with other methodologies and rehabilitation strategies used to promote social skills in people with ASD.

II. UNDERSTANDING AUTISM SPECTRUM DISORDER: CAUSES, PREVALENCE, AND DIAGNOSTIC CHALLENGES

While the exact cause of ASD is currently unknown, research suggests it can be linked to genetics, brain structure differences, and environmental factors [6]. Diagnosing ASD can be challenging because there are no standard medical tests, like blood tests, for it. The process usually begins with general practitioners (GPs) screening for possible autistic-like traits. If there are symptoms of ASD, the GP will refer the child to specialized psychologists or psychiatrists for a more thorough behavioral and cognitive assessment. Although the diagnostic process can start with toddlers as young as 18 months, a final diagnosis might not be made until late [3]. The diagnosis involves clinical experts evaluating the child's developmental age across various categories, such as behavior, communication, self-care, and social skills. This approach, known as clinical judgment, is widely accepted. Common diagnostic tools include the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R), where professionals assess the child through a series of questions and activities. Early diagnosis is crucial because it takes advantage of neuroplasticity, which is greater when children are younger[8]. Recent studies suggest that AI technology not only boosts the cognitive and social skills of children with ASD but also has great potential for rehabilitating children and adolescents with severe intellectual disabilities [17]. Additionally, it's important to highlight the often-neglected negative impact on the quality of life for those with ASD and their families. ASD can significantly increase the burden on caregivers and families, underscoring the need for effective AI-driven interventions to alleviate these challenges. The impact of ASDs on quality of life and families is profound and multifaceted, as highlighted by the study on health-related quality of life (HRQoL) of parents with children diagnosed with ASDs [11]. The research, utilizing both quantitative and qualitative methods, underscores several critical areas of concern. Quantitatively, parents of children with ASDs reported HRQoL scores that were notably lower than those of the general population, particularly in areas related to stress and mental health. The average HRQoL score from SF-6D was 0.74, a clinically significant deviation from the normative U.S. population, indicating substantial stress and mental health challenges. Furthermore, 40% of parents reported experiencing clinical depression symptoms, with unmarried parents showing higher depression levels than their married counterparts. This statistic alone is alarming, highlighting the severe emotional toll on parents. Additionally, families with three or more CSHCN experienced even lower HRQoL and higher caregiving burdens. This finding emphasizes how the cumulative demands of caring for multiple children with special needs can exacerbate stress and reduce overall family well-being. The qualitative data collected through focus groups further corroborates these findings, revealing the deep-seated challenges parents face in managing their children's conditions. This evidence clearly shows that ASDs significantly strain families, affecting both the mental health and overall quality of life of parents. Such impacts are often neglected in broader discussions about autism, which tend to focus primarily on the affected children. However, the well-being of caregivers is crucial; their mental health and quality of life directly influence their ability to provide effective care and support for their children[12]. Therefore, it is imperative that interventions for children with ASDs also include components aimed at supporting parents and reducing their caregiving burden. A recent study [17] suggests that AI technology not only boosts the cognitive and social skills of children with ASD but also has great potential for rehabilitating children and adolescents with severe intellectual disabilities. In light of the above, a selective overview of the newest empirical studies on the use of AI-based programs for recovery ASD individuals was the main objective of this paper. Strengths and weaknesses of the reviewed studies were emphasized, and the implications of the findings were critically discussed.

III. METHOD

To analyze the fundamental characteristics of AI-based programs for the promotion of social skills in children and young adolescents with ASD, a search for empirical studies on the topic was conducted on Scopus. The standard guidelines adopted in this review were in line with PRISMA statement [19] as also demonstrated in Figure 1. The inclusion criteria were:

Keywords: "autism spectrum disorders", artificial intelligence" and "rehabilitation";

Studies published from 2013 to 2024;

Empirical studies;

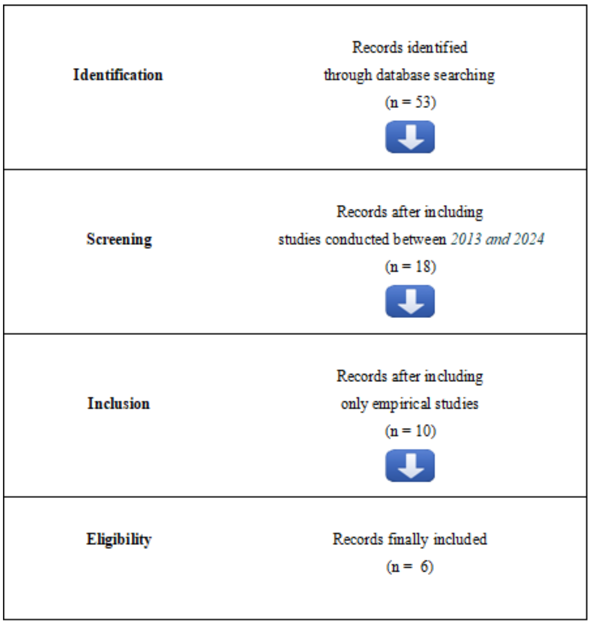


Fig. 1. Flowchart of the studies selection process

Language: English;
Setting: school;
Pertinence to the research question (AI-based programs for promoting social skills in young adolescents with ASD);
Participants: young adolescents (aged between 5 and 11 years).
The exclusion criteria were:
assessment, home or medical center setting;
reviews and conference papers;
age of participants: over 11 years.

An initial search was conducted on Scopus, entering the search keywords “artificial intelligence”, “autism” and “rehabilitation”; the search produced fifty-three results. Including all studies conducted between 2013 and 2024, 18 documents were identified. The search led to eighteen documents. Of these studies, reviews were excluded, and only empirical studies were considered, for a total of ten studies. The eligibility criterion adopted is relevance to the topic of the use of AI-based programs for the promotion of social skills in children and young adolescents with ASD. The research therefore led to six results.

IV. LITERATURE OVERVIEW

An overview was conducted to identify existing research on ASD, AI applications in healthcare, and rehabilitation strategies as also shown in Table I. Documents published between 2013 and 2024 were considered to ensure the inclusion of the most recent and relevant studies. Li et al.[16] showed the idea for the Modular Learning Augmentative and Alternative Communication (MLAAC) system was born from observing a participant with ASD struggling to communicate using existing tools like Picture Exchange Communication System (PECS). Despite these challenges, the participant was able to use an iPad to watch videos and play games. A virtual interface has been developed based on the principle of augmentative and alternative communication (AAC), compatible with devices such as the iPhone. This initial project evolved into MLAAC, a sophisticated communication tool that used modular network technology and required minimal synchronization. Unlike traditional vector graphics, MLAAC leverages advanced web programming languages to offer customizable features such as adjustable sizes, group settings, high-contrast modes, and responsive feedback. These features could be tailored by the user or caregiver, ensuring the tool met individual needs. Overall, MLAAC was aimed to address the diverse communication needs of AAC users through its adaptable, user-friendly design, providing a practical and accessible solution for individuals with communication challenges. Marrauwi et al.[18] showed that using the Internet of Things (IoT) and AI technologies to analyze behavior was rapidly becoming an integral part of their daily lives. As ASD became more prevalent, there was an increasing need to leverage IoT to train and rehabilitate individuals with autism according to their unique abilities and needs. One critical skill for people with autism was transitioning from one place or activity to another. However, specialists often struggled to track progress and determine the effectiveness of different aids and responses to these aids. This challenge was compounded by external factors in school, social, or health settings. Their study focused on developing a sensor-based software tool to analyze the behavior of individuals with autism as they learned to transition from their classroom to the specialist’s room for treatment or rehabilitation. The tool provided indicators of progress and the effectiveness of personalized training programs. By using spatial sensors with Ultra-Wide Band technology, they could accurately track the subject’s movements, dispersion, and the time it takes to transition, while accounting for external influences. The tool successfully processed data from both the entered variables and the sensors, offering trainers and supervisors daily and continuous feedback on the training process. This feedback allowed for quick adjustments to the training plan, thereby reducing the time required to master this transition skill. The result was an efficient and effective training process tailored to the needs of individuals with autism. Safi et al. [23] showed that young adolescents with ASD often struggled with communication and social interaction, primarily due to difficulties in language acquisition. While AI has shown promise in aiding young adolescents with ASD, the effectiveness of Virtual Voice Assistants (VVAs) in this context has not been thoroughly explored. This study was aimed to evaluate the impact of VVAs on enhancing speech and social interaction skills in young adolescents with ASD. The findings revealed notable improvements in the participants’ expressive verbal vocabulary, ability to produce short phrases, and social interactions during the intervention phases compared to traditional methods. Young adolescents interacted effectively with the VVA platform, showing progress in expressive verbal output and social engagement. They pronounced more words correctly with fewer attempts and increased interactions

TABLE I
SYNOPTIC TABLE OF THE REVIEWED STUDIES

Ref	Objectives	Participants	Results
[16]	To evaluate machine learning augmentative and alternative communication (MLAAC) to enhance communicative abilities	Three persons with ASD who showed impaired verbal contact (age not specified)	MLAAC improved communicative abilities of participants
[18]	To evaluate a sensor-based software tool to analyze a person's behavior with ASD	Seven children with ASD (age not specified)	The tool successfully processed the data and provided the trainer and supervisor with daily and ongoing feedback on the training. This resulted in an automatic and rapid examination of the outputs of the individual training plan to make any necessary adjustments, resulting in a reduction in the time spent training this skill.
[23]	To evaluate Virtual Voice Assistant (VVAs) to improve speech skills and social interaction skills	Three young adolescents with ASD (4–11 years old)	Participants showed increases in social interactions. VVAs had positive effects on the speech and social interaction skills of young adolescents with ASD
[29]	To evaluate efficacy of an automated imitation assessment system	Twenty children with ASD (mean age: 4.95 years), and twenty typically developing children s (mean age: 5.30 years)L	The proposed system is able to continuously assess the quality of actions
[32]	To improve the reliability of data acquisition and maximize the potential of data using a multimodal data-driven rehabilitation strategy auxiliary feedback	Twenty children with ASD aged three to seven participated in the study	Multimodal data-driven rehabilitation strategy auxiliary feedback method can provide effective feedback for individuals or groups
[33]	To evaluate a smart rehabilitation product service system based on virtual scenarios for upgrading the rehabilitation service system	Twelve children with ASD (age not specified)	The service system based on the proposed methods can construct an optimal virtual driving system and a rehabilitation program based on the evaluation of patients

with their siblings. Mothers reported satisfaction with the program, indicating that VVAs could be a valuable home-based intervention for young adolescents with speech and social difficulties. Similar findings were reported by Sahin et al. [24], who used digital augmented reality in interventions for social communication, motivation, and cognition in individuals with ASD in a school setting. This study evidenced a technology readily available at home, reinforcing the idea that VVAs' "humanlike" conversational skills could support speech and social development at home. Studies by Parsons et al. [22] also showed improvements in social interactions using avatars and virtual reality applications in simulated environments. This study aligned with those findings, demonstrating the effectiveness of VVAs in enhancing speech and social skills. Mothers noted multiple benefits, including increased sibling interaction and young adolescents' heightened interest in activities involving Siri. Some young adolescents began imitating new words they heard from Siri, which indicated that VVAs could indirectly benefit from expressive language and social interaction. While the study showed positive effects, it remained unclear if these improvements were solely due to the VVAs or also influenced by the mothers' adherence to the intervention. Nonetheless, the children's interest in VVAs over traditional methods suggested that VVAs played a significant role. A single-subject design, used in this study, allowed for rich, in-depth data collection and individual progress tracking, making it advantageous for ASD interventions. Given the heterogeneity of ASD, customized experimental interventions tailored to each participant's needs and goals were beneficial, as each child's pre-study performance served as their baseline. The work of [29] showed that one of the main symptoms of ASD in children was difficulty with imitation skills. In their study, they examined the body gesture imitation performance of 20 children with autism (ASD group) and compared it to 20 typically developing children (TD group) through a series of imitation tasks involving both robots and humans. Manual scoring by two specialists revealed that the TD group significantly outperformed the ASD group in these tasks. Both groups performed better in human-child interactions compared to robot-child interactions within their experimental setup. To advance an automated system for assessing imitation skills, they tested various mathematical models using State-Image based algorithms such as Acceptable Bound, Mahalanobis Distance, and Signals' Cross-Correlations, along with Hidden

Markov Models based on time-dependent kinematic data of participants' joints. Among these models, the "State-Image Acceptable Bound method with position, velocity, and acceleration features" emerged as the most effective, showing a mean Pearson correlation of around 45 per cent, which was comparable to related studies outside the autism field in assessing dynamic action quality. For therapeutic purposes, they proposed using AI algorithms for an automated and unbiased evaluation of children's behaviors. Their suggested approach involved a reciprocal gross imitation human-robot interaction platform, which holded potential for aiding the cognitive rehabilitation of children with autism. Zhao et al. [32] was inspired by observing children with ASD struggling to communicate, researchers developed a multimodal data-driven rehabilitation strategy called MLAAC. This approach combined various technologies to provide effective auxiliary feedback for rehabilitation. The rapid development of Industry 4.0 technologies, such as sensors and deep learning, has revolutionized the medical industry. However, relying on a single technology to meet data requirements is often insufficient. To address this issue, researchers proposed a multidimensional analysis approach that matched technology to the specific needs of doctors and patients. By integrating various data collection methods, a reliable tool could better support the patient's rehabilitation process. Using devices like Microsoft Kinect for ethology (movement tracking) and fNIRS for brain function, researchers collected stable, high-quality data. These tools helped track patients' movements and monitor brain activity, providing comprehensive insights into their condition. Data analysis involved motor representation and statistical methods to correlate movement and brain function changes, creating a robust foundation for auxiliary diagnosis. This layer focused on providing feedback through both quantitative and qualitative analyses. The system collected multidimensional patient data and used correlation and difference analysis to establish relationships between ethology and brain function. These results supported intervention treatment plans, making the rehabilitation process more targeted and effective. The study examined the use of this approach in rhythm rehabilitation training for children with ASD. By observing ethological differences and brain function activation in children with ASD, researchers found that these factors might be age-related. This insight could guide the age-specific design of interventional therapies. In [33] authors showed that the SRP (Smart Rehabilitation Platform) service system leverages virtual scenarios to prioritize user needs by collecting data on user interactions, physiological responses, and behaviors. This comprehensive data collection helped establish reliable evaluation standards for innovative rehabilitation services. The system fostered collaboration between physicians and manufacturers throughout the rehabilitation process, enhancing service quality and promoting user rehabilitation. A case study on a virtual driving system demonstrated the system's effectiveness in creating tailored rehabilitation programs and guiding subsequent rehabilitation stages. Compared to traditional methods, the SRP system allowed for the creation of knowledge graph-based rehabilitation plans, optimizing virtual scenario-based product designs based on patient rehabilitation evaluations. This approach encouraged cooperation between medical institutions and manufacturers, optimizing medical services and products, and improving rehabilitation outcomes. Industry 4.0 has enabled the use of intelligent terminals and sensors to collect extensive data, making smart rehabilitation services more objective, quantitative, and efficient than traditional methods. This study included a design method for the SRP service system, focusing on virtual scenarios, particularly virtual driving for young adolescents with ASD. The system's main contributions included: 1. Overall Design Architecture: Integrating scene building, data collection, data analysis, and innovative services to enhance SRP service modes and promote intelligent development. 2. Quantitative Evaluation Method: Improving the accuracy of rehabilitation efficacy evaluation through comprehensive multimodal data analysis. 3. Collaborative Rehabilitation Services: Helping physicians and manufacturers create knowledge graph-based rehabilitation plans and optimization schemes, guiding subsequent rehabilitation stages. 4. Case Study on Virtual Driving for Young Adolescents with ASD: Analyzing interaction, physiological, and behavioral data to evaluate rehabilitation effects and verify the proposed service system framework. By focusing on these areas, the SRP service system aimed to provide more effective and personalized rehabilitation services, leveraging the latest technological advancements to meet diverse health needs.

V. DISCUSSION

The reviewed studies demonstrated how AI-based programs significantly enhance the assessment and rehabilitation of individuals with ASD. By leveraging advanced technologies, these programs provided tailored and precise interventions that addressed the unique needs of each individual. For instance, Li et al. [16] developed the MLAAC system, which utilizes a customizable virtual interface to support communication, demonstrating that AI could create adaptable tools that cater to specific user requirements. This flexibility ensured that individuals with ASD receive the most appropriate and effective support. AI-based programs have also been instrumental in fostering inclusion in daily settings. Marrauwi et al. [18] highlighted how IoT and AI technologies could track and analyze behavior, particularly in transitions between activities, a common challenge for individuals with ASD. Their sensor-based software tool provided continuous feedback, allowing for real-time adjustments to training programs. This approach not only enhanced the individual's ability to navigate daily environments but also promoted greater independence. The use of safe AI to support independence and quality of life emerged in several studies [5]. Safi et al. [23] showed that Virtual Voice Assistants (VVAs) improved speech and social interaction skills, making home-based interventions more effective. The ability of young adolescents to interact with VVAs and show progress in verbal and social skills highlighted the potential for AI to create supportive and engaging environments that promoted independence and self-determination [23]. Constructive engagement through AI-based interventions was another significant benefit. Taheri et al. [29] demonstrated how a human-robot interaction platform could aid cognitive rehabilitation by assessing imitation skills. The use of AI algorithms for automated evaluations ensured objective and unbiased assessments, facilitating personalized and effective therapy. The burden on caregivers and families is also alleviated through AI interventions. Safi et al. [23] reported that mothers of young adolescents using VVAs noted improvements in their young adolescents' interactions and expressed satisfaction with the program. This suggested that AI tools could provide meaningful support to families, reducing the stress and effort involved in managing ASD-related challenges. However, there were notable weaknesses in the reviewed studies. Many of them relied on small sample sizes, which limited the generalizability of the findings. For example, Taheri et al. [29] included only 20 children in each group, making it difficult to draw broad conclusions. Additionally, there was often a lack of long-term follow-up, which was crucial to understanding the sustained impact of AI-based interventions. The studies also varied in their methodologies and outcome measures, making direct comparisons challenging. In conclusion, AI-based programs hold significant promise for enhancing the assessment, rehabilitation, and quality of life for individuals with ASD. These technologies offered personalized, adaptable, and engaging interventions that could support independence and reduce caregiver burden. However, future research should address the limitations of current studies by including larger sample sizes, ensuring long-term follow-up, and standardizing methodologies to validate and expand upon these initial findings.

VI. CONCLUSION

This overview addressed a critical gap in the literature by providing a comprehensive and up-to-date analysis of AI-based programs for enhancing social skills in young adolescents with ASD. It synthesized recent empirical studies, offering novel insights into the efficacy of various AI-driven interventions. By evaluating the effectiveness of these technologies, discussing ethical considerations, and providing practical implementation strategies, this review significantly advanced our understanding of AI's role in ASD rehabilitation. The findings underscored the potential of AI to create personalized, adaptable, and engaging therapeutic environments that foster independence and improve quality of life for individuals with ASD. AI tools such as the MLAAC system, sensor-based behavior analysis software, Virtual Voice Assistants, and human-robot interaction platforms demonstrated considerable promise in addressing communication challenges, promoting social inclusion, and reducing the burden on caregivers. The reviewed studies collectively highlighted the transformative potential of AI-driven tools in addressing core challenges faced by individuals with ASD. Specifically:

Enhanced Communication: Tools like the Modular Learning Augmentative and Alternative Communication (MLAAC) system provided customizable and user-friendly solutions that cater to individual communication needs. These advancements suggested a move towards more adaptable and accessible communication aids that could be tailored by caregivers and users alike.

Behavioral Monitoring and Training: The use of IoT and AI technologies for real-time behavioral tracking, as demonstrated by Marrauwi et al., enabled precise and continuous feedback. This facilitated more effective and personalized training programs, allowing for quicker adjustments and improvements in skill acquisition.

Virtual Voice Assistants (VVAs): Safi et al.'s [23] findings on VVAs underscored their potential as home-based interventions that could significantly enhance verbal expression and social interactions. The accessibility and convenience of VVAs made them a practical addition to traditional therapeutic approaches, offering supplementary support that could be seamlessly integrated into daily routines.

Imitation and Interaction: The exploration of human-robot interactions by Taheri et al. [29] pointed to the potential of AI in assessing and improving imitation skills. This approach could lead to more objective and automated evaluation methods, reducing the reliance on manual scoring and potentially increasing the accuracy and efficiency of assessments.

Multimodal Rehabilitation Strategies: The studies by Zhao et al. [32] illustrated the benefits of integrating multiple technologies to provide comprehensive rehabilitation support. The Smart Rehabilitation Platform (SRP) and similar multimodal approaches offered a holistic view of the patient's progress, enhancing the personalization and effectiveness of intervention plans.

VII. LIMITATIONS AND FUTURE RESEARCH PERSPECTIVES

Despite the potential practical applications of using AI for the development of rehabilitation strategies aimed at young adolescents with ASD, some limitations emerged from the reviewed studies, which need to be addressed in future research on the topic.

Li W. et al [16] showed that the core functionality of MLAAC is nearly complete, but there are still a few details that need refinement. One key area needing further assessment is the deep learning capability. It's crucial to tailor intelligent suggestions to be genuinely useful rather than irritating or misleading. Researchers have enabled the option to turn these advanced features on or off based on user needs, aiming to understand better how people interact with and utilize these suggestions to perfect them. Beyond this detailed assessment, deep learning techniques could be further developed to offer more precise recommendations. For instance, the device could suggest additional cards based on the user's current location. Additionally, expanding audio processing to include different voice and speech options, as well as refining audio clips, would enhance the device's functionality. Adding comprehensive photo identification and saving features would also be beneficial. Ultimately, the identity design process could incorporate more user data to offer customizable platform configurations, improving the overall user experience.

The study of [23] had several limitations. First, the single-subject design and low number of participants limit the generalizability of the findings. Second, the use of the A-B-A design due to time constraints could be improved with an A-B-A-B design, which would offer a second evaluation phase to better assess the intervention's impact on target skills or behaviors. Third, the limited number of sessions per phase may have restricted the accuracy in determining the intervention's effectiveness. Increasing the number of sessions could provide a clearer picture of the VVA intervention's impact. To build on these findings, future research should address these limitations. Expanding the target population to include not only young adolescents with ASD but also those with other speech and social difficulties would be beneficial. Additionally, comparing the impact of mothers' interventions with regular therapy sessions conducted by therapists or special education teachers using validated assessment tools would provide valuable insights. Despite its limitations, this study contributes to the understanding of how VVAs can assist young adolescents with speech and social difficulties.

While the study of Zhao in [32] contributed to developing a rehabilitation strategy-assisted diagnostic feedback system, it does have some limitations. One major issue is the small sample size, which, although sufficient for verifying the feasibility of the research method, may not provide the robust data needed for comprehensive feedback results. Additionally, linking brain science data with behavioral data remains a challenge. To improve the system, future studies should increase the sample size and conduct more experiments to validate the accuracy of the feedback results. Building a comprehensive ethology and brain science database for ASD across different age groups will also be beneficial. By incorporating machine learning, they can use behavioral and brain function features as data inputs to achieve automatic identification and feedback for ASD through specific tasks. This approach will enhance the reliability and effectiveness of the rehabilitation strategy-assisted diagnostic feedback system. Zhao in [33] showed that despite its benefits, the data collection and analysis process face challenges, such as limited data due to small sample sizes and difficulty establishing correlations between brain science and behavioral data. To improve rehabilitation services, the system should expand to special user groups, integrate into daily life, and enhance behavior monitoring systems. Advanced multimodal data analysis methods can provide high-quality health services and design better rehabilitation products by deeply understanding user needs.

Looking ahead to future research, there are several promising avenues to explore, building upon the limitations identified in current studies. One key aspect is the need to involve more participants in research studies to ensure broader representation and more robust findings. By increasing the participant pool, researchers can gather diverse perspectives and better understand the varied experiences within the target population. Furthermore, it's essential to delve deeper into the implications for privacy and data protection in research involving sensitive information. As technology continues to advance, ensuring the security and confidentiality of participant data becomes increasingly critical. Future research should prioritize exploring innovative approaches to safeguarding privacy while still facilitating meaningful data collection and analysis [21].

Long-term follow-up studies are also necessary to assess the sustained impact of AI-based interventions on individuals with ASD. Standardizing methodologies and outcome measures across studies will facilitate more direct comparisons and robust conclusions. Exploring the integration of multiple AI technologies within a single intervention framework could yield more comprehensive and effective rehabilitation strategies. There's potential for the development of personalized tools tailored to the specific needs of individuals with ASD. By leveraging AI and other cutting-edge technologies, such as augmented reality, researchers can create customized interventions and support systems that address the unique challenges faced by everyone [25]. Integrating AI into interventions might enhance the effectiveness and efficiency of support services, providing personalized recommendations and adaptive learning experiences. Augmented reality offers exciting possibilities for creating immersive and interactive tools that engage individuals with ASD in therapeutic activities and social skills training. Future research in the field of ASD should focus on expanding participant involvement, addressing privacy concerns, developing personalized tools, and leveraging emerging technologies like AI and augmented reality to improve outcomes and support for individuals with ASD. By embracing innovation and collaboration, researchers can continue to advance the understanding and treatment of ASD, ultimately enhancing the quality of life for affected individuals and their families [9]. Exploring new applications of AI, such as virtual reality (VR) and augmented reality (AR), further therapeutic practices can be enhanced. These technologies may offer immersive and engaging ways to support social interaction, communication, and behavior training [28], [26]. Future research should also explore integrating AI with assistive technology interventions and reinforcement learning principles. This combination could lead to highly personalized and tailored solutions, optimizing the learning process, encouraging active participation in training and alleviating the burden of healthcare workers and families [15], [26], [14], [27].

Additionally, Investigating the ethical implications and ensuring data privacy and security will be crucial as AI becomes more integrated into therapeutic settings. Finally, fostering collaborations between researchers, clinicians, caregivers, and technology developers will be vital in creating innovative, user-centered AI solutions that meet the diverse needs of individuals with ASD. In conclusion, the integration of AI into ASD rehabilitation practices presents a significant advancement in the field. By addressing communication challenges, enhancing behavioral training, and providing comprehensive, personalized interventions, AI-driven tools hold the potential to greatly improve the quality of life for individuals with ASD. Continued research, ethical considerations, and interdisciplinary collaboration will be key to realizing the full potential of these innovative technologies.

REFERENCES

[1] Prabal Datta Barua, Jahmunah Vicnesh, Raj Gururajan, Shu Lih Oh, Elizabeth Palmer, Muhammad Mokhzaini Azizan, Nahrizul Adib Kadri, and U Rajendra Acharya. Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. *International Journal of Environmental Research and Public Health*, 2022.

[2] Amanda J Baxter, TS Brugha, Holly E Erskine, Roman W Scheurer, Theo Vos, and James G Scott. The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, 2015.

[3] Nadire Cavus, Abdulmalik A Lawan, Zurki Ibrahim, Abdullahi Dahiru, Sadiya Tahir, Usama Ishaq Abdulrazak, and Adamu Hussaini. A systematic literature review on the application of machine-learning models in behavioral assessment of autism spectrum disorder. *Journal of Personalized Medicine*, 2021.

[4] Antonio Coronato and Muddasar Naeem. A reinforcement learning based intelligent system for the healthcare treatment assistance of patients with disabilities. In *International Symposium on Pervasive Systems, Algorithms and Networks*, 2019.

[5] Mario Fiorino, Muddasar Naeem, Mario Ciampi, and Antonio Coronato. Defining a metric-driven approach for learning hazardous situations. *Technologies*.

[6] Dennis R Grayson and Alessandro Guidotti. Merging data from genetic and epigenetic approaches to better understand autistic spectrum disorder. *Epigenomics*, 2016.

[7] Sajjad Haghshenas, Reza Rostami, et al. Rehabilitation in autism spectrum disorder (asd): a mixture of neurofeedback training and auditory integration training (ait). *Procedia-Social and Behavioral Sciences*, 2011.

[8] Inge Kamp-Becker, Johannes Tauscher, Nicole Wolff, Charlotte Küpper, Luise Poustka, Stefan Roepke, Veit Roessner, Dominik Heider, and Sanna Stroth. Is the combination of ados and adi-r necessary to classify asd? rethinking the “gold standard” in diagnosing asd. *Frontiers in Psychiatry*, 2021.

[9] Umamah bint Khalid, Muddasar Naeem, Fabrizio Stasolla, Madiha Haider Syed, Musarat Abbas, and Antonio Coronato. Impact of ai-powered solutions in rehabilitation process: Recent improvements and future trends. *International Journal of General Medicine*, 2024.

[10] Jana Christina Koehler and Christine M Falter-Wagner. Digitally assisted diagnostics of autism spectrum disorder. *Frontiers in Psychiatry*, 2023.

[11] Karen Kuhlthau, Nalin Payakachat, Jennifer Delahaye, Jill Hurson, Jeff Pyne, Erica Kovacs, and J. Tilford. Quality of life for parents of children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 2014.

[12] Karen A. Kuhlthau, Sheila R. Bloom, Jeanne Van Cleave, Alixandra A. Knapp, Diane Romm, Kirsten Klatka, Charles J. Homer, Paul W. Newacheck, and James M. Perrin. Evidence for family-centered care for children with special health care needs: a systematic review. *Academic pediatrics*, 2011.

[13] Christine Ladd-Acosta, Kasper D Hansen, Eirikur Briem, M Daniele Fallin, Walter E Kaufmann, and Andrew P Feinberg. Common dna methylation alterations in multiple brain regions in autism. *Molecular psychiatry*, 2014.

[14] Giulio E Lancioni, Marta Olivetti Belardinelli, Fabrizio Stasolla, Nirbhay N Singh, Mark F O’Reilly, Jeff Sigafoos, and Maria Teresa Angelillo. Promoting engagement, requests and choice by a man with post-coma pervasive motor impairment and minimally conscious state through a technology-based program. *Journal of Developmental and Physical Disabilities*, 2008.

[15] Giulio E Lancioni, Jorge Navarro, Nirbhay N Singh, Mark F O’Reilly, Jeff Sigafoos, Antonella Mellino, Pietro Arcuri, Gloria Alberti, and Valeria Chiariello. People with neuro-motor impairment, lack of speech, and general passivity can engage in basic forms of activity and communication with technology support. *Advances in Neurodevelopmental Disorders*, 2024.

[16] Wei Li, Xiaoli Qiu, Yang Li, Jing Ji, Xinxin Liu, and Shuanzhu Li. Towards a novel machine learning approach to support augmentative and alternative communication (aac). *International Journal of Speech Technology*, 2022.

[17] Runpeng Liu, Joseph P Salisbury, Arshya Vahabzadeh, and Ned T Sahin. Feasibility of an autism-focused augmented reality smartglasses system for social communication and behavioral coaching. *Frontiers in pediatrics*, 2017.

[18] Mohammed Marrauwi, Aznida Abu Bakar Sajak, and Eiad Yafi. Behavior analysis of people with autism using internet of things—a conceptual study.

[19] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and t PRISMA Group*. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 2009.

[20] Muddasar Naeem and Antonio Coronato. An ai-empowered home-infrastructure to minimize medication errors. *Journal of Sensor and Actuator Networks*, 11(1):13, 2022.

[21] Muddasar Naeem, Antonio Coronato, and Giovanni Paragliola. Adaptive treatment assisting system for patients using machine learning. In *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, 2019.

[22] Sarah Parsons, Anne Lancaster, and Peter Mitchell. Virtual environments for social skills training: Comments from two adolescents with autistic spectrum disorder. *Computers & Education*, 2006.

[23] Mohammed F Safi, Badriya Al Sadrani, and Ashraf Mustafa. Virtual voice assistant applications improved expressive verbal abilities and social interactions in children with autism spectrum disorder: a single-subject experimental study. *International Journal of Developmental Disabilities*, 2023.

[24] Ned Sahin, Rafiq Abdus-Sabur, Neha Keshav, Runpeng Liu, Joey Salisbury, and Arshya Vahabzadeh. Case study of a digital augmented reality intervention for autism in school classrooms: Associated with improved social communication, cognition, and motivation via educator and parent assessment. *Frontiers in Education*, 2018.

[25] Syed Ihtesham Hussain Shah, Antonio Coronato, Muddasar Naeem, and Giuseppe De Pietro. Learning and assessing optimal dynamic treatment regimes through cooperative imitation learning. *IEEE Access*, 2022.

[26] Fabrizio Stasolla, Khalida Akbar, Anna Passaro, Mirella Dragone, Mariacarla Di Gioia, and Antonio Zullo. Integrating reinforcement learning and serious games to support people with rare genetic diseases and neurodevelopmental disorders: outcomes on parents and caregivers. *Frontiers in Psychology*, 2024.

[27] Fabrizio Stasolla, Alessandro O Caffò, Viviana Perilli, Adele Boccasini, Rita Damiani, and Fiora D’Amico. Assistive technology for promoting adaptive skills of children with cerebral palsy: ten cases evaluation. *Disability and rehabilitation: Assistive technology*, 2019.

[28] Fabrizio Stasolla, Anna Passaro, Mariacarla Di Gioia, Enza Curcio, and Antonio Zullo. Combined extended reality and reinforcement learning to promote healthcare and reduce social anxiety in fragile x syndrome: a new assessment tool and a rehabilitative strategy. *Frontiers in Psychology*, 2023.

[29] Alireza Taheri, Ali Meghdari, and Mohammad H Mahoor. A close look at the imitation performance of children with autism and typically developing children using a robotic system. *International Journal of Social Robotics*, 2021.

[30] Joshua Wainer, Kerstin Dautenhahn, Ben Robins, and Farshid Amirabdollahian. Collaborating with kaspar: Using an autonomous humanoid robot to foster cooperative dyadic play among children with autism. In *2010 10th IEEE-RAS International Conference on Humanoid Robots*, 2010.

[31] Shouyao Zhang, Shuang Wang, Ruilu Liu, Hang Dong, Xinghe Zhang, and Xiantao Tai. A bibliometric analysis of research trends of artificial intelligence in the treatment of autistic spectrum disorders. *Frontiers in Psychiatry*, 2022.

[32] Lei Zhao, Haoran Sun, Fei Yang, Zilin Wang, Yufei Zhao, Weizhong Tang, and Lingguo Bu. A multimodal data driven rehabilitation strategy auxiliary feedback method: A case study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[33] Lei Zhao, Yufei Zhao, Lingguo Bu, Haoran Sun, Wanzhi Tang, Kun Li, Wei Zhang, Weizhong Tang, and Yu Zhang. Design method of a smart rehabilitation product service system based on virtual scenarios: A case study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 177–184 DOI: 10.15439/2024F2570 ISSN 2300-5963 ACSIS, Vol. 41

Prediction of Temperature and Precipitation Changes for Serbia Using Time Series Models with Machine Learning

Miroslav Tomić, Slavica Kordić,
Vladimir Dimitrieski
0000-0002-8987-6695
0000-0002-3992-0096
0000-0003-3234-6543
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6,
21000 Novi Sad, Serbia
Email: {tmiroslav, slavica, dimitrieski}@uns.ac.rs

Marija Đukić, Ivan Luković
0000-0002-1136-4278
0000-0003-1319-488X
University of Belgrade, Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: {marija.djukic, ivan.lukovic}@fon.bg.ac.rs

***Abstract*—In the past few decades, there has been an evident change in climatic conditions worldwide as well as on the territory of Serbia. Extremely high temperatures, heavy floods, and sudden changes in the weather are increasingly frequent occurrences that bring great social and material damage. Climate change affects many economic sectors, like tourism and agriculture, which are potentially at risk. In Serbia, one of the vital economic sectors is agriculture. In order to act preventive, the main goal of this research was to predict the mean monthly temperature and precipitation for Serbia for periods 2021-2050 and 2071-2100. We collected a dataset titled ERA5 monthly averaged data on single levels from 1940 to present from the Climate Data Store. The dataset was analyzed and prepared to be used with SARIMA(X) and ARIMA(X) methods, which are utilized for prediction. The results that we identified are presented in this paper.**

***Index Terms*—prediction, temperature, precipitation, time series, machine learning.**

I. Introduction

CLIMATE changes and weather conditions have a great impact on human life. They can bring great material and financial damage, and even human lives can be at risk. Also, many economic sectors, such as tourism, hospitality industry, and construction, are affected. Besides this, climate change negatively affects agriculture in a financial sense. With climate change, there is a severe risk of crop economic losses as well as food security [1]. In Serbia, agriculture proved to be the most vital economic sector [2]. It participated in the creation of GDP with around 8% in 2022 [3], while before, this number was even greater around 20% [2]. The results of a study created by Đuričin et al. [4] indicate that there is a high degree of impact of climate conditions and reduction in crop production in Serbia. Poor crop yields can be a consequence of insect pests that can be affected by changing climate conditions in several ways [1], [5], [6]. They can result in things like an increased number of generations, expansion of their geographic distribution, and increased survival during overwintering. Some attempts were made to detect plant pathogen infestation using AI [7] to increase corn income. Extending such an approach with the prediction of the climate conditions could bring even more benefits. Climate change in Serbia will directly cause drier and warmer regions where vineyards are located, which will have a negative impact on grape growth [8]. Due to climate change, crop yields in Serbia may be reduced, but this can be avoided with preventive actions [9]. One of the possible preventive actions is to predict how much weather conditions will be changed. Additionally, the impact of climate change is important in forest ecosystems. Orlović et al. [10] discussed this impact in their study, which is made for forest ecosystems in Serbia. Also, Stojanović et al. [11] discussed the impact of weather extremes on the forests in Serbia.

Climate change is a consequence of global warming, which is a serious problem that the world is facing today. Global warming has reached record-breaking levels, visible as an increase in atmospheric temperature and sea level [12]. The Earth could experience global warming of 1.4 to 5.8 °C over the next century, which is based on many global climate models and development scenarios [13].

The climate of Serbia can be described as moderate-continental with more or less local characteristics when taking the standard normal period of climatology 1961-1990 [9]. Countries in Southeastern Europe, including Serbia, are facing significant impacts from climate change according to results from all general circulation models. Lalić et al. [14] observed that there is a positive trend in the extreme winter and summer temperatures in Serbia. As shown in a study created by the Ministry of Environmental Protection [15], from 2000 to 2015, the total material damage in Serbia caused by extreme climatic and weather conditions was more than five billion euros. Among the direct consequences of climate change is an increase in temperature, followed by heavy rainfall, which further causes floods and landslides. The cat-

---

This research has been supported by the Ministry of Science, Technological Development and Innovation, Faculty of Technical Sciences, University of Novi Sad and Faculty of Organizational Sciences, University of Belgrade

©2024, PTI

astrophic floods that occurred in 2014 in Serbia were a direct consequence of heavy rainfall, and it was estimated that 1.35 billion euros would be needed to recover from those floods. This directly indicates that it is necessary to take climate change into account when planning sectoral development and infrastructure. According to this study, Serbia is more affected by rising temperatures than most places on Earth, which is a consequence of climate change. Besides this, national defense issues are also affected by climate change. It took a long time for the discussion regarding climate change to expand from the framework of environmental protection to the functioning of national defense [16].

In order to reduce damage, it is possible to act preventive. One of the possible ways is to predict climate change. Quality predictions require the creation of quality predictive models. With this in place, the preparation for upcoming disasters would be much better, and the damage would be significantly reduced. The benefit of such a system is primarily social but also economic. Climate prediction at the state level (certainly at the global level) would help a state to determine its strategic goals. A concrete example of this is the prediction of how many dry years will occur in the coming period. If such predictions guided a country, it could set its strategic goals as early as possible and start with earlier investments in irrigation systems, which would certainly be more favorable from the financial side at this stage and would bring greater profit in later years. A concrete example of this is given by Ruml et al. [8], who stated that in the future, in a certain scenario, the vineyard yield would decrease if there were no irrigation.

Guided by such knowledge, we researched weather conditions and possible climate changes in Serbia. We aim to identify if there will be changes in temperature and precipitation in Serbia using mean monthly temperature and precipitation. The main goal of our research is to predict the mean monthly temperatures and precipitation of Serbia for the periods 2021-2050 and 2071-2100 using machine learning (ML) methods with weather variables that are expected to have the most impact on the prediction. These two periods of time have been choose because other authors also used these periods for prediction, and thus, we can compare our results with their. The obstacle to achieving this prediction with meteorological models is their complexity, which requires a deep understanding of mathematical equations. Also, the number of variables that could affect temperature and precipitation is huge, and it is almost impossible to include all of them. We expect to identify that there will be an increase in temperature and a possible change in the amount of precipitation. These results can help the country to identify earlier which preventive measures can be taken to reduce the damage that can be caused by climate change. In addition, the results could help civil services to better prepare for upcoming disasters, and farmers would be able to make long-term plans for growing crops.

This paper is organized into seven sections. Besides the introduction section, related work is given in the second section. A description of the materials that are used, as well as applied methods, can be found in the third section. The results are presented in the fourth section. The fifth section contains a discussion of the results. Threats to validity are given in the sixth section, and in the seventh section, a conclusion is given.

II. Related Work

In the literature review, we encountered many papers dealing with climate change and the prediction of weather conditions. We classified papers into two categories. The first category contains papers describing statistical and ML methods used to predict temperature and precipitation. The papers from the meteorological domain that focus on Serbia and the impact of climate change on agriculture are in the second category. We use them to compare their results with our findings.

Papacharalampous et al. [17] considered each continent on Earth as a whole, but they did not consider geographical differences within the continent, such as the proximity of the sea mountains. In this paper, the authors used different methods to predict monthly temperature and the monthly amount of precipitation. They used methods: naïve, random walk, AutoRegressive Fractionally Integrated Moving Average (ARFIMA), Box-Cox transformation, ARMA errors, Trend and Seasonal Components (BATS), Simple exponential smoothing, Theta, and Prophet. Data contains a sample of 985 40-year-long monthly temperatures and 1552 40-year-long monthly precipitation time series. For model evaluation, they used RMSE and NSE. In our research, we also have time series data, and we also consider the country as a whole, and we use different statistical and ML methods to achieve better results.

Camelo et al. [18] used Autoregressive Integrated Moving Average with eXogenous Variable (ARIMAX) and the Holt-Winters (HW), both combined with Artificial Neural Networks (ANN) to predict wind speed. As an input to their model, they used time series data, which contains weather-related variables. In our case, we also plan to use a similar combination of models and data. The authors used MAE, MAPE, and RMSE to evaluate models.

According to the literature [18], [19] that we have found, it is possible to apply different types of predictions with ARIMA models, such as long-term and short-term. In both cases, it is possible to predict multiple values or to predict only one value (one-step ahead). Short-term prediction is executed on data where data closer in time has a greater influence on the target variable. It is possible to use short-term and long-term predictions on climate data, which depends also on the amount of available data and the goal of prediction.

To the best of our knowledge, several studies have presented analyses of climate change in Serbia. From a meteorological perspective, the best way to simulate future climate change is by using global oceanic-atmosphere coupled models [20]. Since regional climate has its own specificity, the best way is dynamic downscaling to provide fine-scale information [21]. Because of this, many regional climate model systems have been developed, such as EBU-POM [22]. Kržić et al. [20] predicted changes in climate indices for Serbia using EBU-POM for SRES-A1B and SRES-A2 scenarios. Their results show an overall increase in the surface air temperature of about 2 and 4 °C and a decrease in seasonal precipitation. The number of days with absolute maximum temperature > 30 °C (tropical days) will increase, while the total number of days with absolute minimum temperature < 0 °C (frost days) will decrease in the future.

Mihailović et al. [9] analyzed climate change effects on crop yields in Serbia. They used different meteorological models to predict temperature and precipitation. In meteorology, a 30-year period is used as a reference period, and then an additional period of 30 years should be used for the evaluation of models that are created based on the reference period. This was applied by Mihailović et al. [9]. They used the period 1961-1990 as a reference period, and they gave predictions for the periods 2001-2030 and 2071-2100. In their work, they also included CO2 emissions besides weather variables. Vukovic et al. [23] analyzed the effect of global warming on climate change in Serbia. They analyzed the period 1961-2100 and presented concrete changes in temperature and precipitation. Both papers conclude that Serbia will be affected by climate change due to global warming and that there will be an increase in temperature and an increased amount of precipitation. We use these two papers and paper by Kržić et al. [20] to compare concrete values of predicted monthly temperatures and precipitation with the results that we get.

Gocic and Trajkovic [24] analyzed data for twelve stations in Serbia during 1980-2010 with non-parametric Mann-Kendall and Sen's methods. Their results indicate an increasing trend in temperature. One of the most recent studies on climate conditions in Serbia by Burić et al. [25] shows a tendency towards an arid climate with a significant increase in temperature and changes in precipitation patterns in Serbia.

III. Material and Methods

In this section we first describe the dataset, data preprocessing, data analysis and software tools that we used. Then, we present the method being used in our research.

A. Dataset

The dataset was downloaded from the Climate Data Store (CDS), which provides information on Earth's climate. In our research, we used a dataset titled ERA5 monthly averaged data on single levels from 1940 to present [5]. This dataset is created and maintained by the European Center for Medium-Range Weather Forecasts (ECMWF). The abbreviation ERA5 stands for ECMWF Reanalysis 5th Generation. ERA5 is the climate reanalysis that offers data on an hourly or monthly temporal resolution about atmospheric, land, and ocean parameters and uncertainty estimates. Climate reanalysis combines past observations with meteorological models to generate a consistent time series of multiple climate variables. The data is gridded to a regular latitude-longitude grid of 0.25 degrees and contains 261 variables. After conversation with domain experts, we decided to use five variables in the first phase of our research. We aim to determine if it is possible to create smaller models that don't need many variables but can still make accurate predictions. Variables are described in Table I.

Table I.
Description of variables from the source dataset

Name	Units	Description
10m wind speed	ms^{-1}	The horizontal speed of the wind at a height of 10 meters above the surface of the Earth.
2m temperature	K	The temperature of the air at 2 meters above the surface of the land, sea or inland waters.
2m dewpoint temperature	K	The temperature to which the air, at 2 meters above the surface of the Earth, would have to be cooled for saturation to occur. It can be used to calculate relative humidity.
total precipitation	m	The accumulated liquid and frozen water, comprising rain and snow, falls to the Earth's surface.
evaporation	m of water equivalent	The accumulated amount of water that has evaporated from the Earth's surface.

To analyze data and use it for statistical and ML models, we had to preprocess data, which we did in three steps. In the first step, we downloaded data via CDS Application Programming Interface (API) in Gridded Binary of General Reqularly-distributed Information in Binary form (GRIB). Before downloading data, we had to define the area for the data to be downloaded. We set latitude boundaries between 41 and 47 degrees and longitude boundaries between 18 and 23 degrees which approximately corresponds to Serbia's borders. Also, we defined desired variables, years, and months. In the second step, we went sequentially through the downloaded data and extracted the desired values of variables from GRIB format. Then, we transformed data as a part of the third step and created the dataset that is described in Table II.

Variables 2m temperature and 2m dewpoint temperature were transformed from Kelvin degrees to Celsius degrees. Total precipitation and evaporation were transformed from meters to millimeters. Also, the dataset was appended with additional variable humidity, as the domain expert advised. According to found literature humidity can be calculated as humidity = 100*es(2m dewpoint temperature)/es(2m temperature) where es(t)=610.94*e^(17.625*t/(243.04+t)). Since we used the variable 2m dewpoint temperature only to calculate humidity, we dropped this variable from the dataset. After the data preprocessing we got the dataset which have 6 variables where each variable has 720 values.

TABLE II.
DESCRIPTION OF VARIABLES FROM THE PREPROCESSED DATASET

Name	Units	Description
date	YYYY-MM (format)	Year and month
wind_speed	ms-1	Average monthly wind speed
temp_mean	°C	Average monthly air temperature
prec	mm	Average monthly precipitation
evap	mm of water equivalent	Average monthly evaporation

Since the dataset does not contain any missing values, we analyzed the correlation between data to see if there are dependent variables. Fig 1. shows that the highest correlation is between temperature and humidity. Since a domain expert advised including this variable, we decided to preserve it. Also, there is a correlation between temperature and evaporation. A correlation between temperature and wind speed exists, but according to Wooten , it is only near the Earth's surface, and because of this, we can't say that temperature and wind speed are correlated in general. Precipitation is not correlated with any other variable in the dataset.

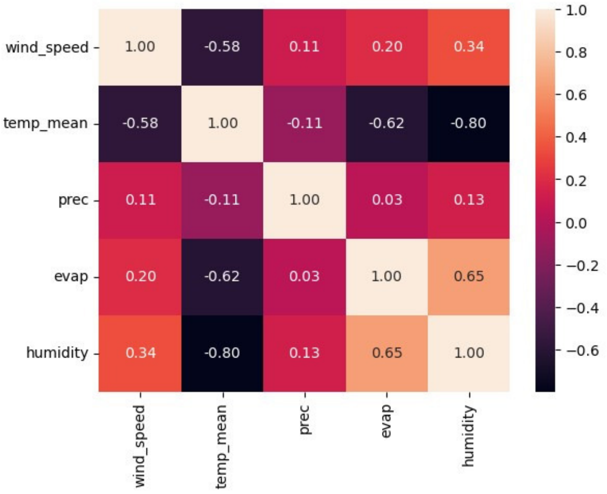


Fig 1. Correlation matrix

Next, we analyzed outlier values with interquartile range. For wind speed, temperature, and evaporation there were no outlier values. For humidity, there are 3 outlier values that are below the lower whisker. In the case of precipitation, 3 outlier values are above the upper whisker. From this analysis, it is obvious that there is a very small number of outliers and we decided to preserve all values.

Since time series data can have seasonality, it was necessary to check if this data has such a component. We applied AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) to check seasonality. The outcome of this check is important for choosing parameters for statistical models. ACF and PACF were applied to the subset of data to be able to visualize it in Fig 2. and Fig 3. These diagrams can be seen for temperature and precipitation, respectively. In the case of temperature, we can notice a seasonality since data repeat every 12 months. This is expected because Serbia has a moderate continental climate condition. On the other hand, for precipitation, seasonality can't be observed, which is typical for precipitations in Serbia, according to the Republic Hydrometeorological Service of Serbia .

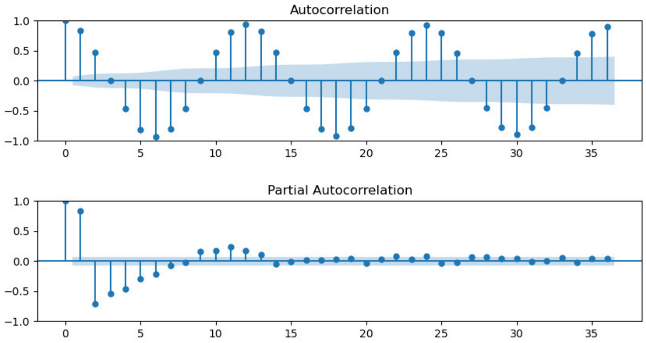


Fig 2. ACF and PACF for temperature

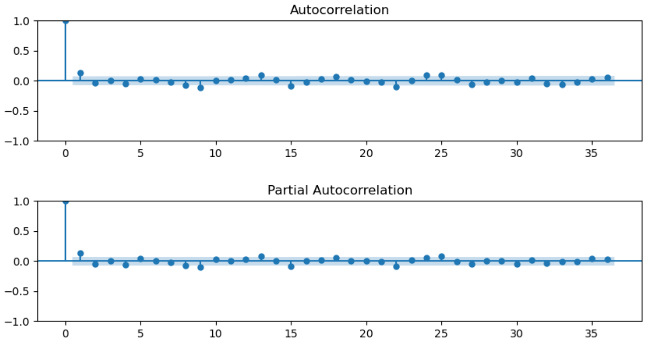


Fig 3. ACF and PACF for precipitation

The statistical models that we used are sensitive to data that is not stationary, which was checked with the Augmented Dickey-Fuller test (ADF). With this test, we get the p-value, and if the p-value value is greater than 0.05, the data is not stationary, and if the p-value is lower or equal to 0.05, the data is stationary. After running this test, we found that all the data from the dataset is stationary. Also, we found that there is no significant trend in the data. After data analysis, we defined methods that will be used for prediction, which is described in the following subsection.

To implement the previously described preprocessing of the dataset, we used the Python programming language with the libraries OS, pandas, matplotlib, seaborn, plotly, and statsmodel. We also used the Python programming language with the libraries OS, itertools, numpy, pandas, math, time, sklearn, and statsmodels to implement methods that will be described in the following subsection.

B. Methods

To achieve the defined goal of this research we decided to use the following models: ARIMAX and Seasonal ARIMA(X) (SARIMAX) with and without exogenous variables. The ARIMA method is widely used in statistical technique for time series analysis and forecasting. It combines three key components: autoregression (AR), differencing (I), and moving average (MA), to model a variety of temporal data. The autoregressive part leverages the dependency between an observation and several lagged observations, while differencing involves subtracting an observation from a previous observation to make the time series stationary. The moving average part models the relationship between an observation and a residual error from a moving average model applied to lagged observations. These components make ARIMA a robust tool for handling non-stationarity data by converting it into a stationary form through differencing. The model parameters (p, d, q) are optimized to minimize forecast errors, where p denotes the number of lag observations, d is the number of times the data needs to be differenced to achieve stationarity, and q is the size of the moving average window. The SARIMA method extends the ARIMA model to handle seasonal variations in time series data. By incorporating seasonal components along with non-seasonal ones, SARIMA can effectively model and forecast data exhibiting periodic patterns. The model is extended by a seasonal part of parameters (P, D, Q, s) where s represent the length of the seasonal cycle. SARIMA offers a framework well-suited for applications such as weather forecasting and other fields where seasonality is a significant factor. Also, we have data that cover a long period of time and since we aim to predict temperature and precipitation for a long period of time in the future we decided to use long-term predictions from SARIMAX model.

IV. RESULTS

For this research, it is important to find appropriate values for all (S)ARIMA(X) parameters to increase the quality of the model. In order to find these values, we used a generic approach called grid search, where all parameter combinations from a limited set of parameter values are exhaustively considered. For parameters related to SARIMAX components we created a set of possible values and for exogenous variables, we defined partitive set of available variables. Each iteration of the grid search created model is evaluated on the test data, and its evaluation is saved to determine the best parameter combination. We use Root Mean Square Error (RMSE) as a metric for evaluation. At the same time, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are calculated to be used potentially by other researchers to compare their results.

Since we aimed to predict the mean monthly temperature and precipitation, we divided the models into these two categories. Thus, for the temperature model, we will use long-term SARIMA with and without exogenous variables, and we will name these models *temp_model_1* and *temp_model_2*, respectively. For the precipitation model, we will use long-term ARIMA because there is no seasonality in precipitation data. This model will also be used with and without exogenous variables, and the names of these models will be *prec_model_1* and *prec_model_2*. For all of these models, for the training set, we used data from the reference period 1961-1990, and for the test set, we used data from the period 1991-2020. For the future predictions we chose the periods 2021-2050 and 2071-2100.

After applying the models *temp_model_1*, *temp_model_2*, *prec_model_1*, and *prec_model_2*, we got the results that indicate that there will be an increase in the mean month temperature. Also, results for precipitation indicate that there will be no change in mean month precipitation. Concrete values are depicted in this section, while details about the results are discussed in the next section.

In the case of *temp_model_1* and *prec_model_1*, we set exogenous variables, and for *temp_model_2* and *prec_model_2*, we didn't use exogenous variables. Since precipitation data does not show seasonality, we didn't use seasonal parameter in hyper-parameter optimization for the precipitation model. After hyper-parameter optimization, we got results which are shown in Table III.

TABLE III.
HYPER-PARAMETERS FOR TEMPERATURE AND PRECIPITATION MODELS

Model	**(p, d, q)**	**(P, D, Q, s)**	**trend**	**exog. variables**
temp_model_1	(1, 1, 0)	(1, 1, 1, 12)	/	evap and humidity
temp_model_2	(1, 1, 0)	(0, 1, 1, 12)	/	/
prec_model_1	(1, 0, 0)	/	constant	evap and humidity
prec_model_2	(0, 0, 0)	/	constant	/

The values obtained in hyper-parameter optimization align with data analysis outcomes. We have evaluated these models over the period 1991-2020. and obtained values that are shown in Table IV. From these values, we can conclude that a temperature model with exogenous parameters gives better results than a model without exogenous variables.

TABLE IV.
EVALUATION OF TEMPERATURE AND PRECIPITATION MODELS FOR THE PERIOD 1991-2020

Model	**RMSE**	**MAE**	**MAPE**
temp_model_1	1.5195	1.1244	0.3857
temp_model_2	1.7007	1.3064	0.3369
prec_model_1	1.0710	0.8893	0.5578
prec_model_2	1.0803	0.8967	0.5773

The other part of our goal was to investigate whether temperature and precipitation increased or decreased when we compare the mean temperature of the reference period and the period 1991-2020. Real values of mean temperature for these periods will be shown in Table V. as well as predictions from temperature models. For precipitation, real values with predictions from precipitation models are shown in Table VI. Also, we created predictions for the mean monthly temperature and precipitation for periods 2021-2050 and 2071-2100, which are shown respectively in Table VII and Table VIII. For this prediction, since exogenous variables are not known for the future period, we used *temp_model_2* and *prec_model_2*, which don't require exogenous variables.

TABLE V.
REAL VS PREDICTED MEAN TEMPERATURE FOR THE PERIOD 1991-2020

Model	**Variable**	**Value (°C)**
/	Temp_true_1961_1990	10.6090
	Temp_true_1991_2020	11.4543
	ΔTemp_true_1991_1961	0.8453
temp_model_1	Temp_pred_1991_2020	11.6192
	ΔTemp_pred_1991_1961	0.1649
temp_model_2	Temp_pred_1991_2020	11.1005
	ΔTemp_pred_1991_1961	-0.3537

TABLE VI.
REAL VS PREDICTED MEAN PRECIPITATION FOR THE PERIOD 1991-2020

Model	**Variable**	**Value (mm)**
/	Prec_true_1961_1990	2.5624
	Prec_true_1991_2020	2.5178
	ΔPrec_true_1991_1961	-0.0446
prec_model_1	Prec_pred_1991_2020	2.5198
	ΔPrec_pred_1991_1961	0.0020
prec_model_2	Prec_pred_1991_2020	2.5657
	ΔPrec_pred_1991_1961	0.0480

TABLE VII.
THE PREDICTED MEAN TEMPERATURE FOR THE PERIODS 2021-2050 AND 2071-2100

Model	**Variable**	**Value (°C)**
/	Temp_true_1961_1990	10.6090
	Temp_true_1991_2020	11.4543
temp_model_2	Temp_pred_2021_2050	11.5712
	Temp_pred_2071_2100	12.3568
	ΔTemp_pred_2021_1961	0.9623
	ΔTemp_pred_2021_1991	0.1170
	ΔTemp_pred_2071_1961	1.7478
	ΔTemp_pred_2071_1991	0.9025

TABLE VIII.
THE PREDICTED MEAN PRECIPITATION FOR THE PERIOD 2021-2050 AND 2071-2100

Model	**Variable**	**Value (mm)**
/	Prec_true_1961_1990	2.5624
	Prec_true_1991_2020	2.5178
prec_model_2	Prec_pred_2021_2050	2.5657
	Prec_pred_2071_2100	2.5657
	ΔPrec_pred_2021_1961	0.0034
	ΔPrec_pred_2021_1991	0.0480
	ΔPrec_pred_2071_1961	0.0034
	ΔPrec_pred_2071_1991	0.0480

V. DISCUSSION

In this section, we discuss our achieved results and compare them with the results found in the literature review.

All models that we created achieved good performances when we consider the RMSE metric, as it is shown in Table IV. As we expected, models with exogenous variables gave better results in the case of temperature prediction, while in the case of precipitation, we got identical results. There is one downside of models with exogenous variables they can't be easily used for future predictions. To use these models, exogenous variables values must be known, which means that the values of these variables must be predicted with another model.

A comparison between the period 1991-2020 and the reference period 1961-1990 for both temperature and precipitation is presented in Table V. and Table VI, respectively. In the case of temperature, we can see that the *temp_model_1* predicts an increase in temperature while *temp_model_2* predicts a small decrease. The true value indicates that there was a small increase in temperature. Concrete values from *temp_model_1* are close to values from papers [9], [23]. For precipitation, both models predicted that there would be a small increase in precipitation, while the true value indicates that there was a small decrease.

Since Vuković et al. [23] and Mihailović et al. [9] also predicted mean temperature for the period 2071-2100, we decided to use *temp_model_2* to predict mean temperature for that period and to compare results. The *temp_model_2* is used since we don't have exogenous variables for this period. Besides this period, we predicted a mean temperature for the period 2021-2050 in order to see if the trend of temperature increase will continue. The values that we got from our prediction indicate that there will be an increase in temperature. These results are shown in Table VII, where we can see that models predicted an increase in the temperature. When we compare this prediction with the prediction by Vuković et al. [23] and Mihailović et al. [9], we conclude the same that there will be an increase in temperature over the years. Kržić et al. [20] estimated that the overall temperature will increase by about 2 and 4 °C for the period 2071-2100 compared to the reference period 1961-1990. We also identified an increase in temperature, but our results show that the overall increase in temperature when we compare the period 2071-2100 with the reference period 1961-1990 will be around 1.75 °C. Our results in Table VIII. show that precipitation will be almost constant, which is different from the results in the papers [9], [20], [23], where it is said that precipitation will change. Precisely, Kržić et al. [20] identified a decrease in precipitation of about 13 and 6 mm while our results show that there will be no change in precipitation.

VI. Threats to Validity

In this section, we express threats to the validity of the proposed work:

- Due to limited hardware resources, we couldn't create a bigger set of parameter combinations that could be used in order to find the best combination with which models could be trained and evaluated. An example is the order of integration where we used range 0-2 which might not be enough. Because of this limitation, we can't be certain if these models could show better performances for temperature and precipitation prediction. To overcome this threat, we could find appropriate hardware to run these models.
- Since the used models are methods that utilize maximum likelihood, we noticed that after multiple runs with the same data and parameters, results can differ. Such a difference is small and appears at the fifth or sixth decimal place in the result. This was determined empirically after multiple tries. In every try, we didn't get results that changed a trend in the temperature and precipitation data that we presented in this paper.

VII. Conclusion

In this paper, we presented an approach to show that it is possible to utilize SARIMA(X) models for climate prediction. During the literature review that we conducted, we didn't find any study that utilizes ML methods to predict climate changes in Serbia. Even though there are many studies that describe the application of ML methods for climate prediction for different areas, there still could be differences in application due to area specificity. The concrete code that is used for this research should be adapted and packaged in an application for use by any user. Since this approach serves more for future predictions in this form, it can only be used for strategic planning. For this approach to be used for operative purposes, a data collection module and a similar model that could be used for short-term predictions are required. Finally, domain experts should validate the results that are obtained from our approach in order to use this approach.

The models we created are simple to implement and utilize on any hardware, which is one benefit. Even if these models give predictions with a small error, they can still help create future plans. Also, because of low hardware requirements, they can be used on farms, where a farmer can use their predictions to create more accurate plans for growing crops. Since these models give accurate predictions for a temperature trend in the future, they can be used for government strategic planning. As an example, agriculture authorities could better allocate funding for irrigation systems. During creation of migrational and socio-economic politics authorities should consider climate change. Additionally, this prediction could influence plans for infrastructure projects as well as companies that want to start a business in Serbia.

It is observed by analyzing models that the best results for predicting mean monthly temperature give a long-term SARIMA model with exogenous variables. The downside of this model is that it can only be used with exogenous variables for future prediction. Even with such a shortage, a temperature model predicted an increase in temperature over the years, which aligns with the outcomes we found during the literature review. In the case of precipitation, our models didn't predict a significant change, which is not in line with outcomes from the literature that we found.

Our plan for future work is to adapt these models for short-term prediction, which could be useful for operative purposes. Since these models could be used in different geographical locations in Serbia, models created for geographical locations with similar climate conditions could show better performances than general models for the whole country described in this paper. It could be helpful to create a few different models for such purposes. Besides this, we would like to increase the number of weather variables used in prediction models, which could contribute to a more accurate prediction. Additionally, using a walk-forward validation technique could give better results. Also, more advanced methods than SARIMA(X), like Recurrent Neural Networks (RNNs), could be used with time-series data. We plan to implement this method and compare the results with the one we presented here. Also, we plan to provide a study on using exogenous variables. Last but not least, we would like to create a different kind of prediction where we will predict the distribution of temperature and precipitation. This is of great importance to avoid disasters caused by sudden changes in climatic conditions.

Acknowledgment

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad” (No. 01-3394/1); and by Faculty of Organizational Sciences, University of Belgrade.

REFERENCES

[1] S. Skendžić, M. Zovko, I. P. Živković, V. Lešić, and D. Lemić, “The Impact of Climate Change on Agricultural Insect Pests,” *Insects*, vol. 12, no. 5, Art. no. 440, 2021, doi: http://dx.doi.org/10.3390/insects12050440.

[2] P. Radovan, “Razvojni problemi i prioriteti poljoprivrede Republike Srbije,” in *Performanse i razvojne mogućnosti agrarne industrije u AP Vojvodini : prilog regionalnoj strategiji unapređenja agrarnog biznisa*, vol. 15, Subotica, Serbia: Univerzitet u Novom Sadu, Subotica: Ekonomski fakultet, 2014, pp. 29–66. doi: http://dx.doi.org/10.2298/PRMAIV2013029P.

[3] Statistical Office of the Republic of Serbia, “Statistical release | Gross Domestic Product, 2022.” Accessed: May 25, 2024. [Online]. Available: https://www.stat.gov.rs/en-US/vesti/statisticalrelease/?p=13820

[4] S. Đuričin, I. Beraha, and M. Đukić, “The impact of climate extremes on agricultural output in the Republic of Serbia,” *Ecol. Nauka Privreda Iskustva*, vol. 20, no. 72, Art. no. 72, 2013.

[5] V. Trkulja, A. Tomić, T. Popović, and R. Iličić, “The Impact of Climate Change on the Occurrence of Diseases and Pests of Agricultural Plants and Forest Trees,” in *Natural Resources Management in a Changing Climate*, vol. 8, 2023. doi: http://dx.doi.org/10.7251/EORU2308477T.

[6] R. Jevtić, S. Maširević, and D. Vajgand, “The impact of climate change on diseases and pests of small grains and sunflower in the Vojvodina Region (Serbia). In Essays on Fundamental and Applied Environmental Topics,” in *Essays on Fundamental and Applied Environmental Topics*, New York, NY: Nova Science Pub Inc, 2012, pp. 277–306.

[7] J. Stypułkowska, “The Use of AI to Determine the Condition of Corn in a Field Robot that Meets the Requirements of Precision Farming,” in *Annals of Computer Science and Information Systems*, 2023, pp. 319–327. doi: http://dx.doi.org/10.15439/2023F3649.

[8] M. Ruml *et al.*, “On the use of regional climate models: Implications of climate change for viticulture in Serbia,” *Agric. For. Meteorol.*, vol. 158–159, pp. 53–62, 2012, doi: http://dx.doi.org/10.1016/j.agrformet.2012.02.004.

[9] D. T. Mihailović, B. Lalić, N. Drešković, G. Mimić, V. Djurdjević, and M. Jančić, “Climate change effects on crop yields in Serbia and related shifts of Köppen climate zones under the SRES-A1B and SRES-A2,” *Int. J. Climatol.*, vol. 35, no. 11, pp. 3320–3334, 2015, doi: http://dx.doi.org/10.1002/joc.4209.

[10] S. Orlović, Z. Galić, S. Stojnić, and B. Klašnja, “Monitoring of forest ecosystems in Serbia,” in *Essays on Fundamental and Applied Environmental Topics*, New York, NY: Nova Science Pub Inc, 2012, pp. 253–276.

[11] D. B. Stojanović *et al.*, “Climate change within Serbian forests: Current state and future perspectives,” *Topola*, no. 208, pp. 39–56, 2021, doi: http://dx.doi.org/10.5937/topola2108039S.

[12] “Climate change 2014: impacts, adaptation, and vulnerability: Working Group II contribution to the fifth assessment report of the Intergovernmental Panel on Climate Change,” Cambridge University Press, New York, NY, 2014.

[13] “Climate change 2007: the physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,” Cambridge University Press, Cambridge ; New York, 2007.

[14] B. Lalic, J. Eitzinger, D. T. Mihailovic, S. Thaler, and M. Jancic, “Climate change impacts on winter wheat yield change – which climatic parameters are crucial in Pannonian lowland?,” *J. Agric. Sci.*, vol. 151, no. 6, pp. 757–774, 2013, doi: http://dx.doi.org/10.1017/S0021859612000640.

[15] D. Bozanic and D. Mitrovic, “Study on the Socio-economic Aspects of Climate Change in the Republic of Serbia,” United Nations Development Programme, 2019.

[16] V. M. Cvetković, S. Z. Vučić, and J. Gačić, “Climate change and national defense,” *Vojno Delo*, vol. 67, no. 5, pp. 181–203, 2015, doi: http://dx.doi.org/10.5937/vojdelo1505181C.

[17] G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, “Predictability of monthly temperature and precipitation using automatic time series forecasting methods,” *Acta Geophys.*, vol. 66, no. 4, pp. 807–831, 2018, doi: http://dx.doi.org/10.1007/s11600-018-0120-7.

[18] H. do N. Camelo, P. S. Lucio, J. B. V. Leal Junior, P. C. M. de Carvalho, and D. von G. dos Santos, “Innovative hybrid models for forecasting time series applied in wind generation based on the combination of time series models with artificial neural networks,” *Energy*, vol. 151, pp. 347–357, 2018, doi: http://dx.doi.org/10.1016/j.energy.2018.03.077.

[19] G. K. Venayagamoorthy, K. Rohrig, and I. Erlich, “One Step Ahead: Short-Term Wind Power Forecasting and Intelligent Predictive Control Based on Data Analytics,” *IEEE Power Energy Mag.*, vol. 10, no. 5, pp. 70–78, 2012, doi: http://dx.doi.org/10.1109/MPE.2012.2205322.

[20] A. Kržič, I. Tošić, V. Djurdjević, K. Veljović, and B. Rajković, “Changes in climate indices for Serbia according to the SRES-A1B and SRES-A2 scenarios,” *Clim. Res.*, vol. 49, pp. 73–86, 2011, doi: http://dx.doi.org/10.3354/cr01008.

[21] F. Giorgi and L. O. Mearns, “Introduction to special section: Regional Climate Modeling Revisited,” *J. Geophys. Res. Atmospheres*, vol. 104, no. D6, pp. 6335–6352, 1999, doi: http://dx.doi.org/10.1029/98JD02072.

[22] V. Djurdjević and B. Rajković, “Development of the EBU-POM coupled regional climate model and results from climate change experiments,” in *Advances in Environmental Modeling and Measurements*, UK ed. edition., New York, NY: Nova Science Pub Inc, 2010, pp. 23–32.

[23] A. Vukovic *et al.*, “Global warming impact on climate change in Serbia for the period 1961-2100,” *Therm. Sci.*, vol. 22, no. 6 Part A, pp. 2267–2280, 2018, doi: http://dx.doi.org/10.2298/TSCI180411168V.

[24] M. Gocic and S. Trajkovic, “Analysis of changes in meteorological variables using Mann-Kendall and Sen’s slope estimator statistical tests in Serbia,” *Glob. Planet. Change*, vol. 100, pp. 172–182, 2013, doi: http://dx.doi.org/10.1016/j.gloplacha.2012.10.014.

[25] D. Burić, J. Mihajlović, V. Ducić, M. Milenković, and G. Anđelković, “Contribution to the study of climate change in Serbia using continentality, oceanity, and aridity indices,” *Időjárás*, vol. 127, no. 3, pp. 379–399, 2023, doi: http://dx.doi.org/10.28974/idojaras.2023.3.6.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 185–190 DOI: 10.15439/2024F7856 ISSN 2300-5963 ACSIS, Vol. 41

FedCSIS 2024 Data Science Challenge: Predicting Stock Trends by a Multi-Dimensional Approach

Quang Hieu Vu
GREENFEED, Vietnam
hieu.vq@greenfeed.com.vn

Dymitr Ruta, Ling Cen, Ming Liu
EBTIC, Khalifa University, UAE
{dymitr.ruta,cen.ling,liu.ming}@ku.ac.ae

***Abstract*—Predicting stock market trends is a challenge that is extremely difficult to solve, yet keeps captivating financial analysts, economists and small investors alike. Stock prices are very volatile, trends follow complex trajectories while the whole financial markets are marred by uncertainty and efficiency principles claiming its unpredictability comes from the fact that any useful evidence in the market is immediately discounted and priced in, such that the price actions of stocks resemble random walk. This very challenge has been proposed as an objective of the FedCSIS'2024 Competition concerned with prediction of optimal equity trade actions based on the established fundamental analysis indicators derived from financial statements and published reports. The dataset comprising thousands of such statements for 300 S&P 500-listed companies from 11 different sectors spanning a period of a decade has been made available along with the optimal trade action labels attached for the training part based on the future return. To address this challenge we have proposed a robust multidimensional model that leverages multiple supervised ML mechanisms to achieve alternative and diverse predictors that are eventually combined in an efficient ensemble to reach the final predictions. Our pragmatic approach vetted with the strict validation set complexity control achieved a very good generalization abilities and won the 2nd place in the competition surpassing in the final evaluation very many competitive models that turned out to be massively overfitted.**

***Index Terms*—Stock trend prediction, Multi-dimensional approach, Classification, Regression, Ensemble, Stacking.**

I. INTRODUCTION

PREDICTING stock market trends is sometimes considered a task to predict the unpredictable. Efficient markets immediately exploit any emerging shred of useful evidence and leave the price actions to follow trajectories that resemble random walks. Inherent volatility, complexity and countless of subtle possible impacts make this task even more hopeless. However, there are certainly some market participants that seem to achieve a consistent risk-adjusted positive returns over long period of time. When considering how this could be possible we might take a look at various time resolutions and predictive horizons and it seems that indeed at the low frequency trading resolution of days and beyond there seems to be a lot of useful merit-based fundamental evidence that offers much more that random guess in relation to the company immediate future return. Several approaches have been typically adopted to address the stock prediction challenge, of which the most popular are:

- Technical Analysis: is an approach that predicts future stock movements by using statistical based methods such as Moving Average (MA), Relative Strength Index (RSI), Bollinger Bands (BB) etc, through the analysis of the stocks' historical price and volume that is typically extrapolated into the future.
- Fundamental Analysis: is an approach that relies on the analysis of the company's financial statements, earnings reports, and economic indicators to determine the stock value from which to predict the stock price direction.
- Machine Learning Models: is an advanced approach that employs supervised machine learning algorithms from a wide range of the traditional models such as decision trees and random forests to deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) to predict future stock prices with models trained on historical data.
- Sentiment Analysis: is a different approach that assumes public sentiment have considerable impacts, and hence correlation with stock prices. As a result, the solution involves monitoring the sentiments from news and social media to identify likely stock price movements.

The 2024 FedCSIS Data Science Challenge[1], dedicated to Predicting Stock Trends, commemorates the 10th event hosted by the FedCSIS Conference on Computer Science and Intelligence Systems[2]. This special anniversary edition focuses on financial data, challenging participants to forecast the performance of selected stocks across different industry sectors. The competition is sponsored by Yettel.Bank (formerly known as Mobi Banka)[3] and the FedCSIS Conference itself.

In the FedCSIS 2024 competition [1], this challenge is revisited. The objective of the competition is to predict stock trends of companies across 11 different industrial sectors, based on the provided dataset containing 58 key financial indicators with annual changes of the 300 pre-selected S&P 500-listed companies from the last 10 years. The indicators are the classic measures used in the fundamental analysis to comprehensively capture the company's state, financial health and growth prospect and are derived from its financial statements and published financial reports. This set is matched with its annual change figures to capture the company's dynamics. This paper presents our solution that won the 2nd place in the

---

[1] https://knowledgepit.ml/fedcsis-2024-challenge/
[2] https://fedcsis.org/
[3] https://www.yettelbank.rs/en/

©2024, PTI

competition based on final model evaluation on the unseen testing set. The prevailing idea of our solution is a multi-dimensional approach consisting of different machine learning models, each of which focuses on a separate dimension of the forecast, from which the results are combined to make a final decision. Specifically, our multi-dimensional approach consists of the following machine learning models:

- As the outcome of the challenge is to predict whether a stock should be bought, sold, or held, our first model is a classification model that classifies the stock into these three classes accordingly.
- As the classification model often mis-classifies examples along the inter-class boundary, we came up with a second model that represents the task as a regression problem and fine-tuned the thresholds to make a better separation between sell and hold as well as hold and buy.
- Our third model also aims to improve separation among classes of trading actions. This model is a combination of a binary classification buy-or-not model, which determines whether a stock should be bought or not, and a binary classification sell-or-not model, which tries to separate sell class from the rest.
- The final model incorporates the continuous return into consideration. The purpose of this model is to provide a better look at the performance dimension of stocks to make a final decision.

Details of each model along with the engineered features will be presented in the later parts of this paper. The rest of the paper is organized as follows. Section II presents Related Work. Section III introduces features engineered for the models. Section IV discusses the details of our approach, and its implementation into a final solution. Section V discusses parametric optimisation and fine-tuning carried out to maximize model performance. Finally, Section VI makes some concluding remarks.

II. RELATED WORK

The use of computational techniques in stock market prediction has been explored extensively over the decades. Initial efforts primarily focused on statistical models like Auto Regressive Integrated Moving Average (ARIMA) [2], which were well-regarded for their predictive accuracy in linear data series. However, as financial market data complexity increased, these models became less sufficient.

With the rise of machine learning, researchers shifted towards models that could capture non-linear relationships in data. For instance, research in [3] demonstrated that Support Vector Machines (SVM) outperform traditional ARIMA models in forecasting stock prices. More comprehensive studies by [4] showed that Random Forests could effectively predict stock direction, providing better accuracy when combined with feature engineering techniques.

The advent of deep learning has introduced more sophisticated AI models like Long Short-Term Memory networks (LSTM) and Convolutional Neural Networks (CNNs), which are particularly adept at handling sequences and spatial structures in data. A pivotal study in [5] illustrated the superiority of LSTMs over traditional machine learning models in predicting stock market trends due to their ability to remember long-term dependencies.

Recent research has also explored the effectiveness of hybrid models that combine multiple AI techniques to enhance prediction accuracy. For example, in [6] developed a genetic algorithm assisted LSTM-CNN hybrid model integrating the LSTM with a complex event processing system to predict stock prices in real-time, showing an remarkable increase in prediction accuracy over using LSTM alone.

Despite these advancements, AI models for stock market prediction face significant challenges, primarily due to the noisy and non-stationary nature of financial data explored in [7]. Moreover, the problem of overfitting and the lack of transparency in deep learning models pose significant hurdles in their practical implementation.

The review of the literature underscores the transition from statistical to more complex AI models in stock market prediction, each offering improvements over previous methods. However, the field continues to face challenges such as model overfitting, data quality issues, and the need for model interpret-ability.

As a team we had actively been participating in the data science competitions [8] - [25] on the KnowledgePit platform[4], predominantly using GBDT-based algorithms for classification, regression, and other tasks and achieving high rankings and insightful experiences.

III. FEATURE ENGINEERING

A. Dataset Description

The objective is to build an accurate method for predicting optimal trading actions (buy, sell, hold). The provided training data consists of 8,000 instances of fundamental financial data in a tabular CSV format. Each instance represents a financial statement announcement for one of the 300 pre-selected companies and includes information about the company's sector, values for 58 key financial indicators, 1-year (absolute) changes for each of these indicators, target class information (in the 'Class' column), and the return performance for a period following the announcement (in the 'Perform' column).

B. Target and Evaluation

The test data, consisting of 2,000 instances, is also provided in the same tabular CSV-formatted file and follows the same naming scheme as the training data but does not include the 'Class' and 'Perform' columns.

Participants were expected to submit their solutions to the online evaluation system as a text file containing exactly 2,000 lines with predictions for the test instances. Each line in the submission must contain a single number from the set 1, 0, -1, representing the predicted trading action for the event: buy,

---

[4] https://knowledgepit.ai/

hold, sell, and exactly matching the order of examples in the testing set.

Submissions are be assessed based on the cost-weighted average error, using the cost matrix as shown in Table I.

Table I
MIS-CLASSIFICATION COST MATRIX

actual \predicted	**sell(-1)**	**hold(0)**	**buy(1)**
sell(-1)	0	1	2
hold(0)	1	0	1
buy(1)	2	1	0

The final model evaluation error (err) used throughout the competition is calculated using the following formula:

$$\mathrm{err} = \frac{\mathrm{confusion\_matrix(preds, gt)} \cdot \mathrm{cost\_matrix}}{\mathrm{length(gt)}}$$

C. Feature Grouping

Feature engineering over unknown dataset and without financial market expertise poses a real challenge. As a result, instead of leveraging domain knowledge, we attempted to randomly generate new features by blind brute-force combinations and aggregations of the original features (indicators), followed with evaluation of the importance of the new features during training to determine its value for the predictive task in hand. Specifically, for each group of financial indicators that we picked, we generated new features using basic statistical aggregation operators within groups as follows:

- Total number of financial indicators with positive values, total number of financial indicators with negative values and the difference between these two results.
- Sum of all financial indicators with positive values, sum of all financial indicators with negative values and sum of values from all financial indicators.
- Std of all financial indicators with positive values, std of all financial indicators with negative values and std of values from all indicators.
- Median, minimum, and maximum values from all financial indicators, as well as total number of financial indicators having NULL or N/A values.

Even though we had planned to perform brute-force search for good features, we quickly realized that the exponentially exploding number of possible combinations even among the basic 58 indicators makes it impossible to complete the task. As a result, to limit the complexity of the search, we tried to group financial indicators based on their perceived semantic similarity (e.g., those having similar keywords in the name or similar meaning). This way, we significantly reduced the number of group-items that we needed to test and hence, unlike for the individual features, it was possible to search though all the combinations of the feature groups. Preliminary baseline experiments evaluating predictive value of the new features revealed that statistics computed over the following groups of original features achieved promising results:

- Group of the 58 key financial indicators and group of the 1-year (absolute) change for each of the 58 indicators.
- Groups the following combinations: ($I1$, $I2$), ($I3$, $I4$), ($I8$, $I9$, 10), ($I11$, $I12$, $I13$, $I14$, $I15$), ($I17$, $I18$, $I19$, $I20$), ($I13$, $I21$, $I23$, $I36$), ($I30$, $I31$, $I32$), ($I39$, $I40$, $I41$, $I42$, $I43$, $I44$), and ($I45$, $I46$, $I54$, $I55$)
- Group of the similar combinations of the 1-year (absolute) change for financial indicators in the groups listed in the point above. For example, there exist group ($dI1$, $dI2$), which is similar to the group ($I1$, $I2$) and group ($dI3$, $dI4$), which is similar to the group ($I3$, $I4$).

With the above approach, we ended up with almost 500 different features for our selection. By using both K-Best and Recursive Feature Elimination (RFE) methods, we received the final set of approximately 270 features coming from all the three groups listed above. There are several interesting observations noted during our feature selection process that were listed below:

- There is no group of features that significantly outperforms other group of features. Feature importance analysis revealed that the important values of features do not vary significantly.
- Even though we used approximately 270 features in our final model, we could have achieved a similar performance with fewer than 100 features.
- While we could achieve similar performance with a much smaller number of features, the performance with reduced set of features is much less stable with respect to changes (e.g., changes of training parameters, number of folds, etc), compared to the performance obtained from a large number of features. Consequently, we chose to keep a large number of features for our final model.

IV. A MULTI-DIMENSIONAL APPROACH

As discussed in the previous section, feature engineering was not our strength in this competition and following preliminary testing we did not expect significant performance breakthroughs in this domain. Instead, we believe that our diversified multi-dimensional approach to model constriction was the key that led us to the very good second position in the final evaluation. The main idea behind our multi-dimensional approach was to look at the predictive problem at hand from many different points of view (or different dimensions) and try to derive the alternative and diverse predictive solutions that could be effectively combined in the final stage. In the following subsections, we present different approaches (dimensions) that we have implemented for this competition.

A. A classification model

Our first approach is a classification model. This approach follows from the original problem statement of: classifying a stock into the three trade action classes: buy, hold or sell. The challenge encountered in this approach, however, is not only the necessity to deal with the 3-class classification problem but also classes imbalance with only 14.21% of hold class and much larger sell (38.68%) and even larger buy (47.11%) class as visually depicted in Figure 1.

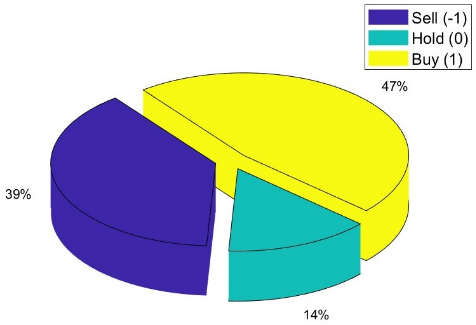


Figure 1. Distribution of trading action labels in the training dataset (8000 examples)

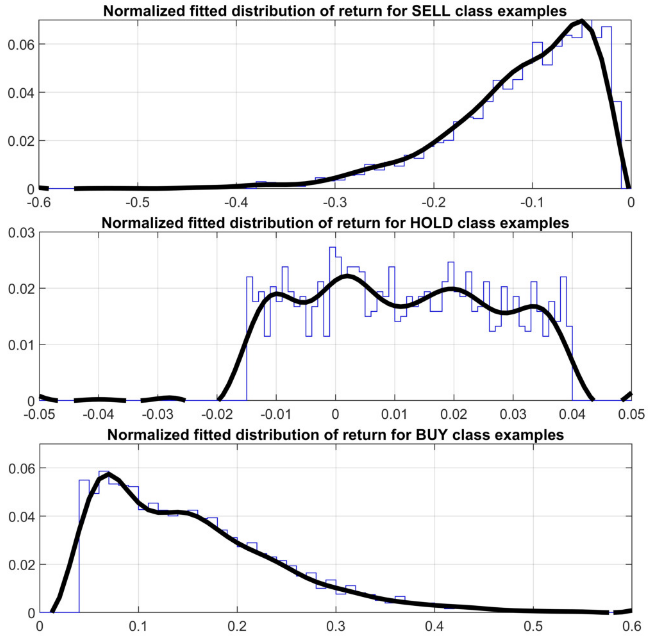


Figure 2. Gaussian process fitted normalized distribution of return within corresponding classes

Here is a strategy we undertook to address the these challenges:

1) We apply hold-class up-sampling and buy/sell-classes down sampling to roughly equalize distribution among all three classes.
2) When training the model we supply prior class distribution to be used for internal weighting designed to compensate for class imbalance.
3) We do nothing during the training process. However, when we generate predictions, we use probability prediction instead of class prediction and distribute range of values for each class in the prediction range of values according to the distribution observed in the training data.
4) Instead of using 3 classes, we split the samples into more classes: 4 classes for buy and 3 classes for sell while keeping the same single class for hold. This way, we obtain the similar number of samples for each class and the model will be balanced hence trained without numerical issues. During prediction, any prediction that falls into the 4 classes of buy, receives the buy label, any prediction falling into one of the 3 sell classes receives a sell label while no change will be observed for predictions of hold class.

We ran a number of experiments to compare the results of these four options and found that the last option (option 4) yields the best performance. As a result, in our final model, we chose to split data into 4 classes of buy, 3 classes of sell and the same class of hold.

B. *A regression model*

Even though the Competition task is nominally a classification challenge, the classes can be ordinal labelled from sell (value -1) through hold (value 0) to buy (value 1) and thereby gain monotonic relationship with the the return, which in turn allows to represent the task and model it as a regression problem either against continuous return or against only three possible target values of {-1,0,1}. As a result, our second approach is a regression model. Specifically, when running experiments for our first classification model, we found that the model often misclassified cases along the inter-class boundary, e.g., stocks with strong features of both sell and hold classes or both hold and buy classes. This effect can be easily explained when investigating continuous return distributions within each class as depicted in Figure 2.

As evident from Figure 2 there are risky cut-offs of prevalent regions around both borders of the hold class with sell and buy classes. For many examples falling into this border regions the return remains virtually the same yet some may fall to different classes. Interestingly the largest chunk of sell class density occupies small negative return right next to the hold class, while buy class tails off more slowly along the growing positive returns. Clearly the regression would be able to better represent small return differences along the classification borders as opposed to step-size changes in the classification and potentially more effectively encourage separation among two adjacent classes: sell and hold as well as hold and buy. The process of fine-tuning threshold values for the class separation is carried out as follows:

- We use Stratified k-Fold to split samples for training and validation to maintain the same distribution of classes.
- We train a regression model using the training data set and generate predictions for the validation data set.
- We compute the distribution of the training data set and use this distribution to find thresholds in the validation data set. Specifically, based on the predictions generated from the model for the validation data set, we set the thresholds so that the distribution of predictions across the three classes is the same as the distribution that we get from the training data (as we use Stratified k-Fold for the splitting earlier).

C. A combination of two binary classification models

While the second model can help to make better separation for stocks in two adjacent classes: sell and hold as well as hold and buy, it cannot address the issue that a stock in sell class is falsely classified into buy class and vice versa. Our third model is designed to address this issue. This model is a combination of the two binary classification models:

- A buy model to determine whether a stock should be bought or not (buy or not-buy). To train this model, we combine samples from sell and hold classes into a single not-buy class while keeping the samples in the buy class unchanged.
- A sell model to determine whether a stock should be sold or not (sell or not-sell). Similarly to the case of the buy model, we combine samples from buy and hold classes into a single not-sell class while keeping samples in the sell class unchanged.

These two models are trained and fine-tuned together so that when combining them for a prediction the following logic is applied:

- A buy prediction of the first model and a not-sell prediction of the second model lead to a buy result.
- A not-buy prediction of the first model and a sell prediction of the second model lead to a sell result.
- A buy prediction of the first model and a sell prediction of the second model lead to a hold result.
- A not-buy prediction of the first model and a not-sell prediction of the second model lead to a hold result.

D. A model considering stock performance

As stock performance has not been considered in any of the above models, this model is designed for this purpose. Here is how the incorporation of the continuous return (stock performance) is separately proposed for classification and regression models:

- For the classification model: as discussed in IV-A, we use 4 classes of buy, 3 classes of sell and a single class of hold. For the buy classes, we put the top 25% of the stocks according to the performance metric into the first class, the next 25% of the stocks going to the second class, and so on. Similarly, we put the top 33% of the stocks into the first sell class, the next 33% of the stocks going to the second sell class, and so on. For the hold class, there is nothing changed as it is a single class.
- For the regression model: instead of using the original values -1, 0, 1 as the target values to train a model, we use the stock performance metric or return to update the target values and stretch it within constrained (-1:1) interval such that only the best performance stock in the buy class is given value 1.0 while other stock values are updated proportionally to their performance in a range from 0.1 to 1.0. Similarly, only the stock having the worst performance in the sell class gains value -1.0 and other values are re-normalized within -1.0 to -0.1 range. Consequently, the hold class examples are also adjusted within the range of values from -0.1 to 0.1.

E. The combination model

After receiving results from different four models, the final step of our multi-dimensional approach is to combine them together. To do this, we tried the two state-of-the-art methods: ensemble and stacking. Note that while these four individual models presented above share a large portion of common features, they do have separate independent features, which are only used exclusively within one model, but not in others. This way, these models are injected with the diversity that helps to combine their results better, i.e. synthetically elevate combined performance above any individual. Here is how the two combination models have been constructed:

- In the ensemble method, we first tried to use the average result from the predictions as the final result. This method suffered from a big issue of class imbalance as it always favors the hold class due to the fact that it falls in the popular middle between the buy and sell classes. To avoid this issue, instead of using the average aggregation, we chose the majority vote method, and received a significant performance gain as a result.
- In the stacking method, we trained a general model that combines the four individual sub-models together with few features. Note that in addition to the results obtained from the four sub-models, we also included features that are not the common features used by the four models when training the stacking model.

Between these two methods, while the stacking method tends to produce a better performance compared to the ensemble method, it is sensitive to changes from individual sub-models and absorbs significantly more time to complete the training process. In the end, we opted for a more stable ensemble method to produce the final predictions of our multi-dimensional model.

V. PARAMETER SELECTION

In addition to the multi-dimensional approach, we believe that selecting the right training parameters to avoid overfitting is also a key point that helped us to achieve a good score in the competition. Even though we first chose Grid Search cross-validation to search for the set of parameters maximizing the evaluation score on the validation sets, we stopped using it as soon as we realized the following two points.

- There are only 10% of the test data used for the public score evaluation, and hence the change of few results may already have a big impact on the public score.
- There is always a big gap between local training score and local validation score as well as between the local validation score and the public score evaluation.

According to our experience accumulated throughout many competitions organized on the KnowledgePit platform, in-line with these two points big re-shuffles in the leader board ranking are possible and in fact expected for such a complex and volatility-prone challenge as is the stock trends prediction. Our suspicion was later proven to be correct for this challenge as in the final evaluation on the full testing set our model surpassed all but one competitive models that turned out to be massively overfitted on the preliminary set. Given this justified expectation, we chose the parameters that may not have the best score, but produce stable results to avoid overfitting. Specifically, we chose to train our model with a narrow tree ($tree\_depth = 5$), a small ratio for feature and sample splitting during the training process ($bagging\_fraction = 0.5$, $feature\_fraction = 0.6$ and a high L1 regularization ($lambda\_l1 = 1.0$). This strategy helped us to achieve more stable result and jump from the rank beyond the top-10 in the preliminary set leader board to the 2nd in the final evaluation. Actually, we are one of the only few teams that managed to achieve a big jump in the ranking from the initial public leader board to the final private leader board. This substantial leap demonstrates the effectiveness of our strategy and adjustments made throughout the competition, reflecting our deep understanding of the challenge and the ability to optimize our solution under different evaluation conditions.

VI. Conclusion

In this paper, we have presented in details how our multi-dimensional approach was designed, implemented, and fine-tuned to achieve a very good result in predicting optimal stock trade actions. While our feature engineering could be similar or even on the modest side compared to the competitive teams', given our inexperience in stock trading, we chose to focus on the aspects of financial predictions that are critical and often overlooked: producing a range of alternative very diverse models utilizing different ML paradigms and representations to produce stable, robust yet diverse predictors of the same target function. With such approach, further boosted with conservative cross-validation and hyper-parameter fine-tuning we managed to elevate the performance further using synthetic ensemble combination scheme rather than stacking offering larger though unstable gains, while invariably guarding every design decision with the careful and conservative validation set evaluation. Our model score the 2nd place in the FedCSIS'2024 Competition and offers encouraging and optimistic outlook on the inherently difficult challenge of profitable stock market prediction.

References

[1] Aleksandar M. Rakicevi, Pavle D. Milosevic, Ivana T. Dragovic, Ana M. Poledica, Milica M. Zukanovic, Andrzej Janusz, Dominik Slezak: Predicting Stock Trends Using Common Financial Indicators: FedCSIS 2024 Data Science Challenge on KnowledgePit.ai Platform, *Proceedings of the 19th Conference on Computer Science and Intelligent Systems (FedCSIS)*, 2024.

[2] Ariyo, Adebiyi A. and Adewumi, Adewumi O. and Ayo, Charles K., "Stock Price Prediction Using the ARIMA Model," *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014, doi=: 10.1109/UKSim.2014.67.

[3] Devi, K.Nirmala and Bhaskaran, V.Murali and Kumar, G.Prem, "Cuckoo optimized SVM for stock market prediction," *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015, doi: 10.1109/ICIIECS.2015.7192906.

[4] Wijaya, Arya Yudhi and Fatichah, Chastine and Saikhu, Ahmad, "Prediction of Stock Trend Using Random Forest Optimization", *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, 2023, doi: 10.1109/ICAMIMIA60881.2023.10427958.

[5] Fauzan, Achmad and SusanAnggreainy, Maria and Nathaniel, Nicholas and Kurniawan, Afdhal, "Predicting Stock Market Movements Using Long Short-Term Memory (LSTM)", *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 2023, doi: 10.1109/AiDAS60501.2023.10284713.

[6] Nikhil, Shreyas and Sah, Rahul Kumar and Kumar Parki, Santosh and Tamang, Til Bikram and D, Somashekhara Reddy. and T R, Mahesh, "Stock Market Prediction Using Genetic Algorithm Assisted LSTM-CNN Hybrid Model", *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, doi: 10.1109/ICCCNT56998.2023.10306948.

[7] Reddy Beeram, Satyanarayana and Kuchibhotla, Swarna, "A Survey on state-of-the-art Financial Time Series Prediction Models", *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, doi: 10.1109/ICCMC51019.2021.9418313.

[8] M. Liu, L. Cen and D. Ruta, "Gradient Boosting Models for Cybersecurity Threat Detection with Aggregated Time Series Features," *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Warsaw, Poland, 2023, pp. 1311-1315, doi: 10.15439/2023F4457.

[9] D. Ruta, M. Liu and L. Cen, "Beating Gradient Boosting: Target-Guided Binning for Massively Scalable Classification in Real-Time," *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, Warsaw, Poland, 2023, pp. 1301-1306, doi: 10.15439/2023F7166.

[10] D. Ruta, M. Liu, L. Cen. FEATURE ENGINEERING FOR PREDICTING FRAGS IN TACTICAL GAMES. *Proc. Int. Conf. 2023 IEEE International Conference on Multimedia and Expo*, 2023.
FEATURE ENGINEERING FOR PREDICTING FRAGS IN TACTICAL GAMES

[11] D. Ruta, M. Liu, L. Cen and Q. Hieu Vu. Diversified gradient boosting ensembles for prediction of the cost of forwarding contracts. *Proc. Int. Conf. 2022 17th Conference on Computer Science and Intelligence Systems*, 2022.

[12] Q. Hieu Vu, L. Cen, D. Ruta and M. Liu. Key Factors to Consider when Predicting the Costs of Forwarding Contracts. *Proc. Int. Conf. 2022 17th Conf. on Computer Science and Intelligence Systems*, 2022.

[13] D. Ruta, L. Cen, M. Liu and Q. Hieu Vu. Automated feature engineering for prediction of victories in online computer games. *Proc. Int. Conf on Big Data*, 2021.

[14] Q. Hieu Vu, D. Ruta, L. Cen and M. Liu. A combination of general and specific models to predict victories in video games. *Proc. Int. Conf. on Big Data*, 2021.

[15] D. Ruta, L. Cen and Q. Hieu Vu. Deep Bi-Directional LSTM Networks for Device Workload Forecasting. *Proc. 15th Int. Conf. Comp. Science and Inf. Sys.*, 2020.

[16] L. Cen, D. Ruta and Q. Hieu Vu. Efficient Support Vector Regression with Reduced Training Data. *Proc. Fed. Conf. on Comp. Science and Inf. Sys.*, 2019.

[17] D. Ruta, L. Cen and Q. Hieu Vu. Greedy Incremental Support Vector Regression. *Proc. Fed. Conf. on Computer Science and Inf. Sys.*, 2019.

[18] Q. Hieu Vu, D. Ruta and L. Cen. Gradient boosting decision trees for cyber security threats detection based on network events logs. *Proc. IEEE Int. Conf. Big Data*, 2019.

[19] L. Cen, A. Ruta, D. Ruta and Q. Hieu Vu. Regression networks for robust win-rates predictions of AI gaming bots. *Int. Symp. Advances in AI and Apps (AAIA)*, 2018.

[20] Q. Hieu Vu, D. Ruta, A. Ruta and L. Cen. Predicting Win-rates of Hearthstone Decks: Models and Features that Won AAIA'2018 Data Mining Challenge. *Int. Symp. Advances in Artificial Intelligence and Apps (AAIA)*, 2018.

[21] L. Cen, D. Ruta and A. Ruta. Using Recommendations for Trade Returns Prediction with Machine Learning. *Int. Symp. on Methodologies for Intelligent Sys. (ISMIS)*, 2017.

[22] A. Ruta, D. Ruta and L. Cen. Algorithmic Daily Trading Based on Experts' Recommendations. *Int. Symp. on Methodologies for Intelligent Systems (ISMIS)*, 2017.

[23] Q. Hieu Vu, D. Ruta and L. Cen. An ensemble model with hierarchical decomposition and aggregation for highly scalable and robust classification. *12th Int. Symposium Advances in AI and Applications (AAIA)*, 2017.

[24] L. Cen and D. Ruta. A Map based Gender Prediction Model for Big E-Commerce Data. *The 3rd IEEE Int. Conf. on Smart Data*, 2017.

[25] D. Ruta and L. Cen. Self-Organized Predictor of Methane Concentration Warnings in Coal Mines. *Proc. Int. Joint Conf. Rough Sets, LNCS*, Springer, 2015.

Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 191–197 DOI: 10.15439/2024F7639 ISSN 2300-5963 ACSIS, Vol. 41

Unused information in token probability distribution of generative LLM: improving LLM reading comprehension through calculation of expected values.

Krystian Zawistowski
0009-0002-9589-9030
Samsung Research Poland
Warsaw
Email: k.zawistowsk@samsung.com, krystian.zawistowski@zoho.com

***Abstract*—LLM text decoding is key component for perceived LLMs quality. We demonstrate two experiments showing that decoding methods could be improved by manipulation of token probabilities. First, we test few LLM on SummEval summary scoring dataset, to measure reading comprehension. We compare scores from greedy decoding to expected values over the next token distribution[1]. We scale logits by large temperature to increase the entropy of scores. This allows strong improvement of performance on SummEval (in terms of correlations to human judgement). We see improvement from 6-8% to 13-28% for 7B Mistral and from 20%-46% to 37%-56% for Mixtral, beating GPT 4 0314 result on two metrics. Part of the gain seems related to positional bias. Secondly, we use probability-based tree sampling algorithm, to examine all most probable generations for given prompt.**

I. INTRODUCTION

GENERATIVE LLMS are trained on large text corpora as estimators of next token probability conditional on prior text. Then sampling from such probability distribution is performed, or token with largest probability is chosen (greedy decoding). Typically, one introduces parameter T – called temperature. Let l_i be the logit for i-th token. Then token probability is as follows:

$$p_i = \frac{e^{\frac{l_i}{T}}}{\sum_j e^{\frac{l_j}{T}}} \tag{1}$$

Question arises: what the T should be? Typically $T \in (0, 1]$, with greedy decoding as limit in 0 and larger values corresponding to greater diversity (but also greater randomness). Research [1] shows that human generated text often does not correspond to modelled highest probability. Human choice of words is not guided by greatest probability, as "humans optimize against stating the obvious". The author of this observation, H. P. Grice in [2] gave following example: suppose that I meet a man with a gas tank asking me to sell some gas and I answer "There's gas station over the corner". I said only a bit of information that doesn't asnwer directly, while there's lots of implicit information not being said (gas station is open, it seels gasoline and has it available and you can go there buy some). Such concepts might be pretty foreign to LLMs: when we asked Mixtral Instruct "How to get gasoline in Fresno", it gave us long instruction on finding gas station on a map, choosing best gas station, operating the pump safely, paying and so on, despite the fact that most of it is irrelevant to the problem of getting gasoline in Fresno specifically.

Thus, a tradeoff arises. Probability maximization with small temperature doesn't give us natural, relevant responses. Large T is not ideal either, introducing more randomness, as low probability token might be either very informative or very wrong. Is however a single fixed value of T sufficient, even for specific use case?

We conjecture that decoding should be more dynamically controlled to more fully utilize the information in the distribution.

II. SUMMARY EVALUATION WITH EXPECTED VALUE DECODING.

A. Expected value decoding.

Currently, greedy decoding is often used for label based QA[2]. We want to find whether relative probabilities of few most probable tokens are informative. We test whether greedy decoding can be outperformed by calculation of expected value. We evaluate our approach on SummEval [4] dataset. It contains 1600 article summaries with human annotations for relevance, fluency, coherence and consistency of a summary. We compare our result against known LLM-based evaluations ([5], [6]). LLM is asked to evaluate relevance (or other feature) on Likert scale (from 1 to 5). We use MCQ prompts from [6], where LLM answer is A, B, C, D or E (A is 1 – worst, E is 5

---

[1] Source code released here: https://github.com/kzawisto/unused_information_llm

[2] For example in LM Evaluation Harness, standard set of LLM benchmarks [3]

©2024, PTI

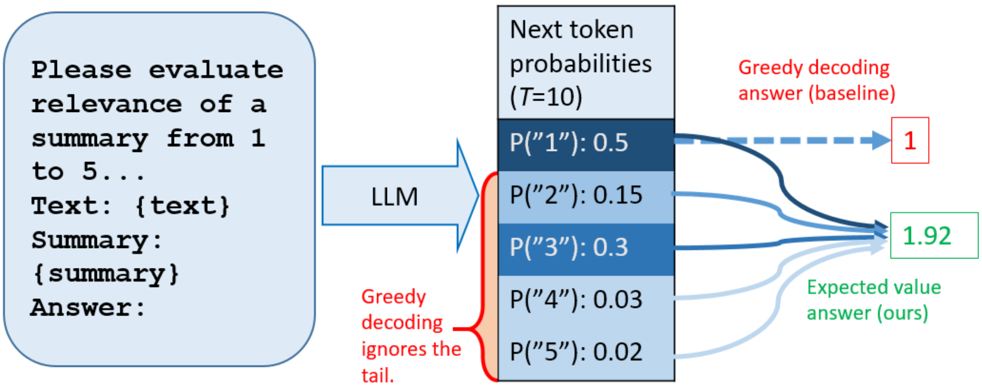

Please evaluate relevance of a summary from 1 to 5...
Text: {text}
Summary: {summary}
Answer:
LLM
Next token probabilities (T=10)
P("1"): 0.5
P("2"): 0.15
P("3"): 0.3
P("4"): 0.03
P("5"): 0.02
Greedy decoding answer (baseline)
1
1.92
Expected value answer (ours)
Greedy decoding ignores the tail.

Fig. 1. Conceptual diagram of presented approach: instead of answering with most probable token, we calculate expected value for temperature $T = 10$ to utilize residual information in next-token distribution.

– best). Let $p(A), p(B)$... be probability of token "A", "B"... Expected value score (that we use) is calculated as follows:

$$E(s) = p(A) + 2p(B) + 3p(C) + 4p(D) + 5p(E), \qquad (2)$$

while greedy score (a standard baseline) is:

$$s_{max} = F\left(\text{arg max}_{t\in\{A,B,C,D,E\}} p(t)\right), \qquad (3)$$

where F is mapping $\{A \to 1, B \to 2...\}$. $E(s)$ depends on T, as (1) shows, while s_{max} does not. However, our initial experiments show that for $T \in (0, 1]$ these values are very close to each other, $E(s) \approx s_{max}$. LLM is spuriously certain about its answer and assigns near 100% probability to selected answer. To fix that we increased entropy of score distribution by setting very high $T = 10$ (i.e. we want for these scores to have smooth continuous distribution over the [1, 5] interval). Conceptual diagram is presented on Figure 1.

B. Results

We evaluated Pearson correlation of scores to human judgements, the scores being calculate either with greedy method (3) or expected value method 2 with $T = 10$. We saw strong improvement in metric correlation to human judgement. In addition, $E(s)$ scores from Mixtral 8x7B Instruct [7] beat GPT3.5 and nearly match GPT4 results from [6], see Table I.

Furthermore, strong improvements were produced for small and quantized LLM too. We evaluated 3 LLMs, from 7B to 47B parameters and all metrics are consistently improved. We compare scores for quantized and float16 Mixtral Instruct in Table II. Surprisingly, quantized Mixtral performance is only slightly worse than float16 version.

Up to 4.4 times improvement is achieved for Mistral v0.2 Instruct 7B from [8] (from 6.4% to 28.4% on relevance). We show these results in Table III. In Table IV we show result for SOLAR 10.7B Instruct [9]. While authors of the model reported it to outperform much larger Mistral 7x8B on some benchmarks, we see nothing similar for SummEval. Consistently with Mistral results, for SOLAR the largest gain was observed for relevance evaluation (from 19% to 43%). For every model we show results for float16 inference and also for model checkpoints quantized with use of GPTQ [10].

We see that summarization metrics, being relevant automated metrics for reading comprehension, strongly improve with the number of parameters. Also, quantized 4-bit LLMs are very strong performers proportionally to their size and outperform similar size float16 models (for instance, quantized Mixtral, having 24GB in parameters and Mixture-of-Experts architecture strongly outperforms float16 SOLAR with 21 GB of parameters). For this reason quantized LLMs might be viable, cost-effective option for RAG and other similar use cases. This phenomenon is similar to emergent abilities of LLMs [11] where larger sizes lead to qualitative improvement in LLM performance. In this case too, quantized LLMs [12] retain large portion of their emergent capabilities.

Gains are particularly strong for relevance and consistency: this is important for systems that rely on reading comprehension, like RAG expert systems. We used vLLM [13] and Transformers [14] for implementation.

C. Positional bias.

Previously it was reported that LLM preference for candidate responses might be altered [15] by simply reordering the responses in the prompt. This effect is called positional bias. Our experimental setup might be affected by it, as we use multi choice question answering prompts from [6].

We modified our approach as follows: we evaluate our score for two nearly identical MCQ prompts that differ by the order of answer candidates. One prompt has answer candidates in A, B, C... order, the other in E, D, C... order. Having done that, we average the scores for two prompts, doing that for every example we evaluate.

We performed this experiment for Mistral 7B for relevance evaluation and results can be found in Table V. Quite interestingly averaging out positional bias produces strong improvement for greedy decoding, while there's no big difference for $E(s)$ decoding. Furthermore stronger improvement is found

TABLE I
PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL: MIXTRAL-INSTRUCT AND PRIOR RESULTS FOR OPENAI MODELS.

Metric	GPT3.5 0301 [6]	GPT4 0314 [6]	ChatGPT [5]	Mixtral E(s), T=10	Mixtral (greedy)
Fluency	0.431	**0.600**	0.384	0.392	0.222
Relev.	0.395	0.461	0.459	**0.555**	0.457
Consist.	0.484	**0.618**	0.516	0.506	0.397
Coher.	0.416	0.428	0.438	**0.485**	0.427

TABLE II
PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL FOR MIXTRAL 8X7B V 0.1 INSTRUCT.

Metric	Fp16, greedy	Fp16, E(s)	Int4, greedy	Int4, E(S)
Fluency	0.222	0.392	0.235	**0.405**
Relev.	0.457	0.555	0.464	**0.564**
Consist.	0.379	**0.506**	0.293	0.470
Coher.	0.428	**0.485**	0.388	0.438

TABLE III
PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL FOR MISTRAL 7B V 0.2 INSTRUCT.

Metric	Fp16, greedy	Fp16, E(s)	Int4, greedy	Int4, E(S)
Fluency	0.060	**0.134**	-0.045	0.061
Relev.	0.064	**0.284**	0.074	0.264
Consist.	0.061	**0.252**	0.076	0.249
Coher.	0.084	**0.199**	0.042	0.176

TABLE IV
PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL FOR SOLAR 10.7B INSTRUCT.

Metric	Fp16, greedy	Fp16, E(s)	Int4, greedy	Int4, E(S)
Fluency	0.187	0.240	0.187	**0.251**
Relev.	0.192	**0.427**	0.165	0.364
Consist.	0.298	0.331	0.156	0,194
Coher.	0.305	**0.362**	0.200	0.267

for Fp16 model, than for Int4 model. Table VI shows few more experiments for Nous Hermes DPO Mistral 7B model, where either prompt candidates put in ascending (A, B, C... order), reversed (E, D, C... order) or random order. NaN correlations indicate that the model had predicted identical result for all test examples. One new result here is that random order is remarkably bad, with many correlations dropping by 30% or more. From human point of view prompt says exact same thing, but this is not the same for LLMs, which cannot generalize when the structure is altered. Also, NousHermes Mistral, undergoing more extensive finetuning and alignment, outperforms Mistral Instruct on metrics related to logical reasoning, but underperforms on fluency and coherence.

This suggests that gains from $E(s)$ method might be related to positional bias, but details of it are not clear without further research.

This looks related to spurious certainity of LLM we already mentioned, our conjecture of temperature misconfiguration and improved results for very high $T = 10$. LLM, when having no good candidate hypothesis, seems to overreact to weak signals - instance of this problem is positional bias. While averaging out provides specific solution to positional bias, setting large temperature provides general solution: as LLM might take into account more candidate hypotheses, which presently are dominated by overreaction to spurious signal.

These problems could be related to the use of softmax function in attention heads. Neural net limitations with respect to softmax and the rank of matrix under it were brought to attention by [16] (which proposes high rank RNN). Similar problems might reemerge in case of transformers and attention, which use relatively small matrices for attention heads. For softmax it does not matter whether signal is weak or strong, only whether it is the strongest among provided candidates. It is also true however, that weak signal supression can be learned by the attention head in the pretraining process, especially when bias matrices are added to Q and K – so it is impossible to tell more without further study.

D. Statistical analysis.

We evaluate our results on 1600 samples from SummEval dataset, calculating Pearson correlation to human judgement evaluations on four metrics: fluency, relevance, consistency, coherence. We recalculate correlations for 1600 ChatGPT-evaluated samples provided by [5] and 1200 samples evaluated by GPT3.5 0301 and GPT4 0314, provided by [6].

We evaluate statistical significance with use of bootstrap method. We randomly shuffle series of human evaluation metric x_i and we do it 10000 times. For every random shuffle $\hat{x}_i$ and Pearson correlation coefficient r we calculate $\bar{x}_i = \frac{r}{\sqrt{1-r^2}}\hat{x}_i + x_i$. Clearly for large sample size Pearson correlation $corr(x_i, \bar{x}_i) \to r$. We examine the empirical cumulative distribution $P(\rho|r, x_i)$ of $corr(x_i, \bar{x}_i)$.

TABLE V

PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL FOR MISTRAL 7B V 0.2 INSTRUCT - POSITIONAL BIAS ANALYSIS.

Metric	Fp16, greedy	Fp16, E(s)	Int4, greedy	Int4, E(S)
Relevance - standard.	0.064	0.284	0,074	0.264
Relevance - average.	0.245	0.295	0.161	0.307

TABLE VI

PEARSON CORRELATIONS TO HUMAN JUDGEMENT ON SUMMEVAL FOR MISTRAL 7B NOUS HERMES DPO FP16 - POSITIONAL BIAS ANALYSIS.

Metric	Greedy, ascending	Greedy, reverse	Greedy, random	E(S), ascending	E(S), reversed	E(S), random
Relevance	0.32	0.36	0,092	0.44	0.42	0.2
Consistency	0.38	0.33	0.17	0.5	0.42	0.3
Fluency	0.097	NaN	0	0.12	0.0001	0.015
Coherence	0.16	NaN	0.04	0.24	0.06	0.061

With this we seek to evaluate, whether the difference of two sample correlations r_1 for sample 1 and r_2 for sample 2 is statistically significant. Significant difference of r_1 and r_2 would correspond to r_2 being unlikely result if real correlation for sample 2 was r_1:

$$P(\rho < r_2 | r_1, x_i) < 5\%$$

and

$$1 - P(\rho > r_2 | r_1, x_i) < 5\%,$$

according to p-value testing methods.

Our estimates suggests that our results in Table I for consistency evaluation and relevance evaluation with Mixtral are significantly better that GPT model results, as far as statistical significance is concerned.

In addition, almost all improvements of $E(s)$ method over greedy method provide significant difference in correlation. Only exception is consistency evaluation for SOLAR for Fp16 model in Table IV, which is not statistically significant (the difference is 3.3% while significance threshold corresponds to 3.4%).

III. TREE-BASED SAMPLING

To further develop our hypothesis we propose an LLM inference analysis method that, for a given prompt, seeks to find all probable completions that could be generated by nucleus sampling - to give complete, nearly deterministic picture, what LLM outputs could be for given prompt. As a foundation we use tree-search based sampling algorithm. We use priority queue mechanism, where most probable completions are evaluated first (like in Dijkstra algorithm). Tree sampling (a.k.a beam search) is broadly implemented approach[3] in generative language models. Recently we saw very similar algorithm to ours [17] applied to compiler optimization (highest probability output produces superior compiler parametrization). Other controlled beam search techniques used for improved natural laguage generation can be found in [18], [19].

We utilize priority based tree sampling to find all possible or most probable completions for nucleus sampling. Algorithm 1 shows this procedure in pseudocode. This algorithm has exponential asymptotic complexity: every iteration produces N new sequences without fixed lower bound for N (LLM tokenizers have tens of thousands of tokens), leading to exponential divergence K^N for K new tokens. One could decrease N by adjusting T and $\hat{p}$. We notice that for some prompts N is small number and large values of N indicate a qualitative change in the text generation (such as going from direct answer to user query, to additional not needed remarks). As an example of this we evaluate following prompt for Mixtral instruct[4]:

```
<s> [INST]Please provide one original,
creative paraphrase for sentence
"My name is John Kennedy"
and write new line after it[/INST]
Answer:\n\n"
```

Outputs with their evaluated probabilities can be found in Table VII. We used nucleus sampling threshold $\hat{p} = 0.9$ and temperature $T = 2$ and we show outputs with $p > 0.1\%$. Temperature is large, and reason is that for smaller temperatures only first, most probable output would be generated, while now it is generated with 73% probability. Output distribution is uneven, with top output 41 times more probable than second most probable output and very fat tailed with with about 23% of probability mass distributed among very unlikely outputs ($p < 0.1\%$).

We did not get any diversity of the paraphrase, despite asking for it explicitly, we got only two options: “I go by the name of Kennedy John” and “I go by the name of JFK”. At the same time LLM becomes unpredictable in the tail of the distribution as various additional comments follow after requested text. Output is thus not diverse and diversity we get provides little benefit: it might be a problem, when long unpredictable output follows the answer, evading usual stopping mechanisms of the inference (here we stop inference on two new lines in a row.)

This type of behavior can be easily explained with reference to the content of training corpora for LLM. Specific tasks like “paraphrase this sentence” or specific sentences like “My name is John Kennedy” are likely rare in the large internet crawl corpora. At the same time, casual conversation is more frequent, so LLM can generate diverse full sentence answers (but this is not very useful for instruction-following tasks).

---

[3] Available in popular library Transformers [14].

[4] The prompt contains no new lines, but line wrapping was added for clarity.

Algorithm 1 Tree-crawling topP algorithm

```
Require:
  $t_1...t_n$ {Prompt input sequence.}
  $\phi(t_1..t_n) \to l_i$ {LLM that maps token sequence to next-token log-probabilities}
  $\hat{p} \in (0, 1)$ {TopP probability threshold}
  $\alpha$ {Minimum loglikelihood of completion.}
  $StopTokens$ {Tokens that terminate inference, such as newline or end-of-sentence.}
  $MaxSteps$ {Max number of LLM evaluations}
Ensure:
  $Queue \leftarrow [(t_1..t_n, 0)]$ {Priority queue ordered by second argument}
  $Complete \leftarrow []$ {Generated sequences terminated on stop tokens.}
  $Incomplete \leftarrow []$ {Generated sequences terminated on minimum logprobability $\alpha$}
  $i \leftarrow 0$
  while $Queue$ not empty and $i < MaxSteps$ do
    $seq_i, lproba \leftarrow Queue.pop()$
    $l_i \leftarrow \phi(seq)$
    for all $token, l \in TopPCandidates(l_i, \hat{p})$ do
      $el \leftarrow (concat(seq, token), lproba + l)$
      if $token \in StopTokens$ then
        $Complete.append(el)$
      else if $l + x \geq \alpha$ then
        $Queue.append(el)$
      else
        $Incomplete.append(el)$
      end if
    end for
    $i \leftarrow i + 1$
  end while

  return $Complete, Incomplete$
```

TABLE VII
TREE SAMPLING ALGORITHM OUTPUT FOR MIXTRAL INSTRUCT, COMPLETED SENTECES.

Probability	LLM output
0.73	I go by the name of Kennedy John."
0.018	I go by the name of JFK"
0.005	I go by the name of JFK" is a creative paraphrase for "My name is John Kennedy."
0.005	I go by the name of JFK" is a creative paraphrase for the sentence "My name is John Kennedy."
0.003	I go by the name of JFK" is a creative paraphrase of "My name is John Kennedy."
0.002	I go by the name of JFK" is a possible creative paraphrase for "My name is John Kennedy."
0.001	I go by the name of JFK" is a possible paraphrase for "My name is John Kennedy."
0.001	I go by the name of JFK" followed by, "What an honor to make your acquaintance!"

Results point to causes of few problems of LLMs that we believe to be fairly widespread in applications based on prompting LLM and parsing their output.

1) Repetitive output – LLM output might be often identical, as there is single completion with very large probability. LLM might sometimes provide little advantage compared to retrieval based or rule based system.
2) Stopping instability – when LLM completes desired output, many different unwanted follow-up comments might be produced, disturbing rule based inference stopping mechanism.
3) Uncontrollability – when LLM is asked to do something, it can ignore instruction.
4) Hallucination - false or otherwise unwanted outputs might be produced by LLM in rare cases, while being undetected in tests using standard decoding.

These issues are rarely detected by commonly used accuracy-based benchmarks. There are some generic automated metrics such as MAUVE for diversity [20], but goal of our method is to analyze these issues in specific use cases and provide guidelines on configuration and further analysis. Our algorithm allows to analyze influence of modified prompts and system configuration on output probabilities, allowing greater degree of reliability and objectivity in the development, as different prompts, LLMs or sampling algorithm can be compared. On more general terms it seems that decoding heuristics like

TopP seem to fail our expectations, where entropy is extremely small or extremely large. Instead other approaches might be investigated: a) Scaling T for entropy extremes of next-token probability distribution. b) Taboo sampling – tree sampling with penalty for token and substring repetition. c) Stopping generation for large entropy spikes (as that would likely result in unwanted output). This might allow new improvements in few important KPI for LLM, such as output diversity, controllability and safety.

Among related work we may mention Mirostat [21], a sampling algorithm similar to TopP. Fixed perplexity objective allows to avoid both incoherence for large p and repetition for small p, while similar approach of [22] seeks to generate text with locally constant information content. Also vLLM [13] implements repetition penalty to stabilize low p inference. Another tradeoff however remains, as high Shannon information makes no difference between highly informative word in human terms and unwanted token overrepresented in the training dataset (as Shannon information is defined as the inverse of probability). In addition, prioritizing largest possible probability is very useful for some use cases, such as multi-choice QA [3] or LLM for source code generation (see Fig. 7 in [23]) or compiler configuration [17]. Balancing this tradeoff for humal language is a problem with no general solution yet known to us. We thus seek to provide a tool for analyzing specific problems, such as prompting, interaction between LLM and rule based scripts or stopping LLM inference reliably.

IV. CONCLUSIONS

We show LLM decoding method that improves performance for answers given on quantitative scale: such as "evaluate relevance of summary on a scale from 1 to 5". On SummEval summary evaluation dataset the method produces strong improvements, with open source LLM nearly matching much larger GPT3.5 and GPT4, with GPT4 0314 outperformed on relevance and coherence. Such improvement supports our hypothesis that the temperature might be not optimally configured in standard LLM decoding, as token probabilities do not reflect real world probabilities and small and large temperatures serve different purposes.

We demonstrate new LLM analysis method using priority based tree sampling algorithm, useful for study of some niche problems in LLM, such as the diversity and controllability of the output.

We show reading comprehension metrics for few different LLM with sizes 7B, 10.7B and 47B with float16 (half precision) inference and 4 bit GPTQ quantization. Summarization metrics strongly improve with the number of parameters, and quantized 4-bit LLM are effective in proportion to their size (which is of interest for RAG on low-powered systems).

REFERENCES

[1] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. doi: 10.48550/arXiv.1904.09751. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH

[2] H. P. Grice, "Logic and conversation," in *Syntax and Semantics: Vol. 3: Speech Acts*, P. Cole and J. L. Morgan, Eds. New York: Academic Press, 1975. doi: 10.1163/9789004368811_003 pp. 41–58.

[3] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, "A framework for few-shot language model evaluation," 12 2023. doi: 10.5281/zenodo.10256836. [Online]. Available: https://zenodo.org/records/10256836

[4] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, 2021. doi: 10.1162/tacl_a_00373. [Online]. Available: https://aclanthology.org/2021.tacl-1.24

[5] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou, "Is ChatGPT a good NLG evaluator? a preliminary study," in *Proceedings of the 4th New Frontiers in Summarization Workshop*, Y. Dong, W. Xiao, L. Wang, F. Liu, and G. Carenini, Eds. Singapore: Association for Computational Linguistics, Dec. 2023. doi: 10.18653/v1/2023.newsum-1.1 pp. 1–11. [Online]. Available: https://aclanthology.org/2023.newsum-1.1

[6] C. Shen, L. Cheng, X.-P. Nguyen, Y. You, and L. Bing, "Large language models are not yet human-level evaluators for abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023. doi: 10.18653/v1/2023.findings-emnlp.278 pp. 4215–4233. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.278

[7] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," 2024. doi: 10.48550/arXiv.2401.04088

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. doi: 10.48550/arXiv.2310.06825

[9] D. Kim, C. Park, S. Kim, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim, C. Ahn, S. Yang, S. Lee, H. Park, G. Gim, M. Cha, H. Lee, and S. Kim, "Solar 10.7b: Scaling large language models

with simple yet effective depth up-scaling," 2024. doi: 10.48550/arXiv.2312.15166

[10] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," 2023. doi: 10.48550/arXiv.2210.17323

[11] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022. doi: 10.48550/arXiv.2206.07682

[12] P. Liu, Z. Liu, Z.-F. Gao, D. Gao, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen, "Do emergent abilities exist in quantized large language models: An empirical study," 2023. doi: 10.48550/arXiv.2307.08072

[13] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th Symposium on Operating Systems Principles*, ser. SOSP '23. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. ISBN 9798400702297 p. 611–626. [Online]. Available: https://doi.org/10.1145/3600006.3613165

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020. doi: 10.18653/v1/2020.emnlp-demos.6 pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[15] C. Shen, L. Cheng, X.-P. Nguyen, Y. You, and L. Bing, "Large language models are not yet human-level evaluators for abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023. doi: 10.18653/v1/2023.findings-emnlp.278 pp. 4215–4233. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.278

[16] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," 2018. doi: 10.48550/arXiv.1711.03953

[17] D. Grubisic, C. Cummins, V. Seeker, and H. Leather, "Priority sampling of large language models for compilers," 2024. [Online]. Available: http://arxiv.org/abs/2402.18734

[18] C. Meister, M. Forster, and R. Cotterell, "Determinantal beam search," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021. doi: 10.18653/v1/2021.acl-long.512 pp. 6551–6562. [Online]. Available: https://aclanthology.org/2021.acl-long.512

[19] L. Vilnis, Y. Zemlyanskiy, P. Murray, A. Passos, and S. Sanghai, "Arithmetic sampling: parallel diverse decoding for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023. doi: 10.5555/3618408.3619870

[20] K. Pillutla, L. Liu, J. Thickstun, S. Welleck, S. Swayamdipta, R. Zellers, S. Oh, Y. Choi, and Z. Harchaoui, "Mauve scores for generative models: Theory and practice," 2023. doi: 10.5555/3648699.3649055

[21] S. Basu, G. S. Ramachandran, N. S. Keskar, and L. R. Varshney, "{MIROSTAT}: A {neural} {text} {decoding} {algorithm} {that} {directly} {controls} {perplexity}," in *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2007.14966

[22] C. Meister, T. Pimentel, G. Wiher, and R. Cotterell, "Locally typical sampling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 102–121, 2022. doi: 10.48550/arXiv.2202.00666. [Online]. Available: https://api.semanticscholar.org/CorpusID:252918666

[23] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, "Evaluating large language models trained on code," 2021. doi: 10.48550/arXiv.2107.03374

Author Index

Abomelha, Fares 1
Addabbo, Pia 111
Ahmad, Tanwir 151
Almazni, Yahya 9
Alsulami, Rasha 19
Asselborn, Thomas 29

Baum, Myriam 45
Beloff, Natalia 9, 19

Cavalli, Ana 151
Cen, Ling 185
Čepek, Miroslav 77
Charvat, Karel 77
Coronato, Antonio 111
Curcio, Enza 169

Dimitrieski, Vladimir 177
Dörpinghaus, Jens 29
Đukić, Marija 177

Fiorino, Mario 111

Geldhauser, Carina 37
Gerhards, Christian 45
Gheorghiță, Ștefana 117
Gioia, Maricarla Di 169
Grüner, Lea 83
Gurova, Silvi-Maria 125

Hildebrand, Robert 69
Huertas, Carlos 53

Jedlička, Karel 77
Jíša, Nikolas 61

Kausar, Faraz 29
Kepka, Michal 77
King, Brannon 69
Kordić, Slavica 177
Kothari, Kush 105
Koutenský, František 77
Krüger, Kai 83
Kuzmych, Lyudmyla 95
Kyrylov, Yuriy 95

Lazarova, Meglena 125
Liu, Ming 185
Luković, Ivan 177

Mallouli, Wissam 151
Malyshev, Konstantin 37
Matson, Eric 159
Mehta, Bhavya 105
Melzer, Sylvia 29
Möller, Ralf 29
Mombelli, Aurélien 131

Naeem, Muddasar 111
Nambiar, Reshmika 105
Newbury, Paul 1
Nica, Anca-Maria 117

Ostromsky, Tzvetan 125

Passaro, Anna 169
Pergl, Robert 61
Pihrt, Jiří 77

Quilliot, Alain 131

Reiser, Thomas 139
Ruta, Dymitr 185
Rybář, Vojtěch 77

Sadovykh, Andrey 151
Schneider, Martin 151
Seceleanu, Cristina 151
Shrawne, Seema 105
Šimánek, Petr 77
Smith, Eunsun 159
Stasolla, Fabrizio 169
Steiner, Petra 139

Tomić, Miroslav 177
Truscan, Dragos 151

Voloshyn, Mykola 95
Vu, Quang Hieu 185

Wang, Yinxuan 159
White, Martin 9, 19

Zawistowski, Krystian 191
Zullo, Antonio 169