

Annals of Computer Science and Information Systems
Volume 45

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

September 14–17, 2025. Kraków, Poland



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



Annals of Computer Science and Information Systems, Volume 45

Series editors:

Maria Ganzha (Editor-in-Chief),

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

Leszek A. Maciaszek,

Wrocław University of Economics, Poland and Macquarie University, Sydney, Australia

Marcin Paprzycki,

Systems Research Institute, Polish Academy of Sciences, Warsaw and Management Academy, Warsaw, Poland

Dominik Ślęzak,

University of Warsaw, Poland

Marek Bolanowski,

Rzeszów University of Technology, Rzeszów, Poland

Senior Editorial Board:

Wil van der Aalst,

RWTH Aachen University, Netherlands

Enrique Alba,

University of Málaga, Spain

Marco Aiello,

University of Stuttgart, Germany

Mohammed Atiquzzaman,

University of Oklahoma, USA

Christian Blum,

Artificial Intelligence Research Institute (IIIA-CSIC), Spain

Jan Bosch,

Chalmers University of Technology, Sweden

George Boustras,

European University Cyprus, Cyprus

Barrett Bryant,

University of North Texas, USA

Rajkumar Buyya,

University of Melbourne, Australia

Chris Cornelis,

Ghent University, Belgium

Robertas Damaševičius,

Kaunas University of Technology / Vytautas Magnus University, Lithuania

Hristo Djidjev,

Los Alamos National Laboratory, Los Alamos, NM, USA and Institute of Information and Communication Technologies, Sofia, Bulgaria

Włodzisław Duch,

Nicolaus Copernicus University, Toruń, Poland

Schahram Dustdar,

Research Division of Distributed Systems at the TU Wien, Austria and part-time ICREA research professor at UPF, Spain

Hans-George Fill,

University of Fribourg, Switzerland

Ulrich Frank,

Universität Duisburg-Essen, Germany

Ana Fred,
*Department of Electrical and Computer Engineering, Instituto Superior Técnico
(IST—Technical University of Lisbon), Lisbon, Portugal*

Giancarlo Guizzardi,
University of Twente, Netherlands

Francisco Herrera,
University of Granada, Spain

Mike Hinchey,
University of Limerick, Ireland

Janusz Kacprzyk,
Systems Research Institute, Polish Academy of Sciences, Poland

Irwin King,
The Chinese University of Hong Kong, Hong Kong

Michael Luck,
King's College London, United Kingdom

Ivan Luković,
University of Belgrade, Serbia

Marjan Mernik,
University of Maribor, Slovenia

Michael Segal,
Ben-Gurion University of the Negev, Israel

Andrzej Skowron,
University of Warsaw, Poland

John F. Sowa,
VivoMind Research, LLC, USA

George Spanoudakis,
University of London, United Kingdom

Editorial Associates:

Katarzyna Wasielewska,
Systems Research Institute Polish Academy of Sciences, Poland

Paweł Sitek,
Kielce University of Technology, Poland

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

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

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

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

Annals of Computer Science and Information Systems, Volume 45

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

ISBN 978-83-973291-9-5

ISSN 2300-5963

DOI 10.15439/978-83-973291-9-5

© 2025, Polskie Towarzystwo Informatyczne

Ul. Solec 38/103

00-394 Warsaw

Poland

Contact: sekretariat@fedcsis.org

<http://annals-csis.org/>

Cover photo:

Łukasz Kotyński,

Elbląg, Poland

Also in this series:

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

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

Volume 42: Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering **ISBN 978-83-973291-5-7**

Volume 41: Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-973291-0-2, ISBN USB: 978-83-973291-1-9**

Volume 40: Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-9-1, ISBN USB: 978-83-969601-0-8**

Volume 39: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-6-0, ISBN USB: 978-83-969601-7-7,**

ISBN ART 978-83-969601-8-4

Volume 38: Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering, **ISBN WEB: 978-83-969601-5-3**

Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**

Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2**

Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**

Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**

Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7,**

ISBN USB: 978-83-965897-7-4

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

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

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

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

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

During FedCSIS 2025 four keynote speakers delivered lectures providing a broader context for the conference participants. These presentations were:

- Damaševičius, Robertas, Kaunas University of Technology, Lithuania

Keynote title: AI-Driven Innovations in Brain Cancer Research

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

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

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

At the time, when you are reading this text, videos of the keynote presentations and of invited contributions, delivered during the FedCSIS 2025 conference, are already available on the official conference website (www.fedcsis.org). We warmly encourage you to visit the website and watch these recordings to gain additional insights and perspectives shared by distinguished speakers.

Finally, as a part of official Conference Opening, a special presentation, entitled: *Paths to Zero Emission Computing – Reducing Energy Consumption, and carbon emissions in HPC and AI environments*, was delivered by Tikiri Wanduragala, Technology Leader Lenovo Infrastructure Solutions Group (ISG), Lenovo UK and Ireland. An extended abstract, outlining main points of this presentation can be found in this volume.

FedCSIS 2025 consisted of Main Track, with five Topical Areas, and 12 Thematic Sessions. Some of Thematic Sessions have been associated with the FedCSIS conference series for many years, while some of them were 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.

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

fantastic group of student helpers. We are very grateful for your efforts!

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

Co-Chairs of the FedCSIS Conference Series

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

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

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

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

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

Preprints of Communication Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 14–17, 2025. Kraków, Poland

TABLE OF CONTENTS

COMMUNICATION PAPERS

Tree Segmentation from Low-Resolution Digital Orthophotos using a Hybrid Deep Learning Model	1
<i>Irfan Abbas, Robertas Damaševičius, Rytis Maskeliūnas, Muhammad Abdullah Sarwar</i>	
Flexible and Scalable Results Collecting in Distributed Spatial Simulations	9
<i>Piotr Aksamit, Mateusz Najdek, Wojciech Turek</i>	
Enhancing Research Data Integrity Through Blockchain: Design and Implementation of a Web Based Management System	19
<i>Sikandar Ali, Roberto Ciccocioppo, Massimo Ubaldi, Andrea Morichetta, Matteo Piersantelli</i>	
A Constraint Programming Approach for Urban Drone Trajectory Optimization	25
<i>Zahraa Asfour, Sonia Cafieri, Andrija Vidosavljevic</i>	
Effects and challenges of Governance, Lean Healthcare and Digital Transformation in Outpatient Management: An integrative review	33
<i>Lucas Beckman, Cristiana De Muylder, Olaf Reinhold</i>	
Domain-as-Particle with PSO Methods for Neural-Network Feature Weighting	41
<i>Fabio Berberi, Paolo Mercorelli</i>	
Tools for Implementing Social Innovation in the Circular Economy: Learnings from the CSS Boost project	49
<i>Cristina Alejandra Barahona Cabrera, Olaf Reinhold</i>	
Modeling and optimizing flow networks with several constraints using sequential dynamical systems	55
<i>Jens Dörpinghaus, Michael Tiemann, Robert Helmrich</i>	
Ontological support for integration computer tools in digital humanities research	63
<i>Iwona Grabska-Gradzińska, Grażyna Ślusarczyk, Barbara Strug</i>	
3D Brain Extraction from Magnetic Resonance Imaging Using Knowledge Distillation	71
<i>Kali Gürharaman, Ahmet Firat Yelkuvan, Rukiye Karakis</i>	
Enhanced GI Tract Cancer Diagnosis Using CNNs and Machine Learning Models	77
<i>Abdul Haseeb, Faheem Shehzad, Sidra Naseem</i>	
Towards OntoUML for Software Engineering: Transformation of Constraints into Various Relational Databases	85
<i>Jakub Jabůrek, Zdeněk Rybala, Petr Kroha</i>	

A Framework for Machine Unlearning Using Selective Knowledge Distillation into Soft Decision Tree	95
<i>Sangmin Kim, Byeongcheon Lee, Sungwoo Park, Miyoung Lee, Seungmin Rho</i>	
Smart Routes: Hybrid Metaheuristics for Efficient Vehicle Routing Problem	103
<i>Yehor Kovalenko, Andrei Pivavara, Joanna Ochelska-Mierzejewska</i>	
Opportunities and Challenges of LLMs as Post-OCR Correctors	111
<i>Radoslav Koyanov, Triet Ho Anh Doan</i>	
Perception and Emotional Response to AI-Generated Audiovisual Media: The Influence of Content and Context	119
<i>Nina Krzemińska, Mirosława Długosz</i>	
Static components dependency graph detection with evaluation metrics in React.js projects	127
<i>Lukasz Kurant</i>	
Exploring the entire medicinal chemistry space on the hybrid computational platform with quantum annealer and gate-based quantum circuits	135
<i>Jung-Hsin Lin</i>	
Effectiveness of metaheuristics applied to Human Resource Allocation Problem in Short-Term Employment Sector – a case study	137
<i>Paweł Myszkowski, Michał Przewozniczek, Lukasz Kopocinski</i>	
Towards a German VET Archive and its Integration into a Data Warehouse	145
<i>Thomas Reiser, Petra Steiner, Kristine Hein</i>	
A Confidence-Interval Circular Intuitionistic Fuzzy Zero Point Model for Optimizing Spare Parts Transfer in Smart Manufacturing Environments	153
<i>Velichka Traneva, Stoyan Tranev, Mihai Petrov, Venelin Todorov</i>	
Claim Frequency Estimation in Motor Third-Party Liability (MTPL): Classical Statistical Models versus Machine Learning Methods	161
<i>Ondřej Vít, Lubomír Seif, Lubomír Štěpánek</i>	
AI in theatre. Witkacy case study	167
<i>Marek Średniawa</i>	
Author Index	175

Tree Segmentation from Low-Resolution Digital Orthophotos using a Hybrid Deep Learning Model

Irfan Abbas, Robertas Damaševičius, Rytis Maskeliūnas, Muhammad Abdullah Sarwar
Centre of Real Time Computer Systems, Kaunas University of Technology
Kaunas, Lithuania

irfan.abbas@ktu.edu, robertas.damasevicius@ktu.lt, rytis.maskeliunas@ktu.lt, m.sarwar@ktu.edu

Abstract—This study presents a cost-effective tree crown segmentation framework using a hybrid deep learning model that combines a ResNet-34 encoder with a U-Net decoder. Our approach operates on low-resolution RGB Digital Orthophotos (DOPs) collected from urban and peri-urban areas in Bochum, Germany, simulating real-world data constraints. We processed 450 orthophoto-mask pairs through a comprehensive preprocessing pipeline including resizing (from 20000×20000 to 256×256), augmentation, and noise simulation. The model was trained using 10-fold cross-validation, achieving a Dice coefficient of 0.8678, Intersection over Union (IoU) of 0.7754, precision of 0.8410, and recall of 0.9103. These results demonstrate that even with downsampled imagery, reliable segmentation of tree crowns is feasible, making our approach suitable for low-cost forest inventory and precision agroforestry applications. Unlike previous studies relying on high-resolution LiDAR, this work is among the first to show robust tree crown segmentation using low-resolution orthophotos, making it accessible for widespread use in resource-constrained settings.

Index Terms—Tree segmentation, Digital Orthophotos, Remote Sensing, Forest Monitoring, Forest 4.0, Deep Learning.

I. INTRODUCTION

FOREST monitoring is essential for ensuring the sustainable management, conservation, and restoration of forest ecosystems, which are critical to biodiversity, climate regulation, and human well-being [1]. By continuously tracking changes in forest cover, composition, and health, monitoring efforts support early detection of deforestation, forest degradation, pest outbreaks, and the impacts of climate change [2], [3]. With the increasing complexity of environmental challenges and the growing demand for data-driven decision-making, traditional forest monitoring methods are evolving toward more integrated, automated, and scalable solutions. This transformation is embodied in the concept of Digital Forestry, which leverages advanced technologies such as remote sensing, artificial intelligence, Internet of Things (IoT), and geospatial analytics to enhance forest observation and analysis. Within this context, the emergence of Forest 4.0 that integrates cyber-physical systems, real-time data processing, and predictive analytics to enable proactive decision-making, optimize resource use, and ensure ecological resilience [4].

This research paper has received funding from Horizon Europe Framework Programme (HORIZON), call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) - Creation of the centre of excellence in smart forestry “Forest 4.0” No. 101059985. This research has been co-funded by the European Union under the project “FOREST 4.0 - Ekscelencijos centras tvariai miško bioekonomikai vystyti” (Nr. 10-042-P-0002).

Recent advances in satellite imagery have created new opportunities for forest monitoring, including tree segmentation on a large scale [5], [6]. Laser imaging, detection, and ranging (LiDAR)-based high-resolution [7] satellite data can provide detailed information on forest structure, including tree height, tree trunk, and tree area; however, such data are not always available and can be costly [6], [8]. Low-resolution Red-Green-Blue (RGB) satellite imagery cannot provide detailed forest information, such as forest structure and the area covered by the trees [9]. So, it is very challenging to get details of tree structure from the low-resolution LiDAR-based RGB satellite data due to mixed pixels, blurry images, shadows, overlapping, and lack of spectral information [10]. Dataset pre-processing is required to get the forest and tree structure details from the RGB images [11]. These images estimate tree cover, species, biomass, changes in forest characteristics, tree shapes, wood calculations, trunk detections, and many others [12].

We propose a Deep Learning (DL)-based model for using low-resolution RGB satellite imagery from the German forest to segment the tree areas, which comprises of ResNet-34 and U-Net models. The contributions of this study are as follows:

- We introduce a custom hybrid model that integrates a ResNet34 encoder with a U-Net decoder, combining strong feature extraction capabilities with spatial reconstruction suited for semantic segmentation tasks on low-resolution imagery.
- The model is trained and validated on DOP and nDOM datasets from the German federal state of North Rhine-Westphalia, demonstrating the feasibility of using 256×256-pixel images for high-quality segmentation.
- A comprehensive data preprocessing framework is developed, including image resizing, augmentation, compression, and noise simulation, to enhance model robustness and generalization across heterogeneous landscapes.

II. RELATED WORK

Multiple models have been developed to segment individual trees from forests and urban streets using high-resolution and low-resolution LiDAR point clouds and RGB images, and results have shown significant differences [12]. Few of these developed methods use to detect and segment the tree tops because these are the highest points in the RGB imagery and LiDAR point cloud data. By identifying and detecting the tree tops, the features of the trees are extracted and segmented. The

[13] develops a YOLO-based CCD-YOLO model to segment the individual tree using the LiDAR datasets collected from Beijing and Henan Polytechnic University, China.

Study [14] proposed a Fuzzy Center Segmentation (FCS) for ITS and used the ALS and TLS LiDAR-based datasets collected from the Jasper National Park, Montane Cordillera Ecozone, Canada, but did not segment the trees very well.

Study [15] used a watershed transformer based on U-net to perform ITS on high-resolution data sets collected from Bengaluru, India, and Gartow, Germany. The accuracy achieved by the author is 46.3% and the IoU is 71.2%. For the other dataset, the accuracy was 52% and the IoU was 72.6%, which is better than on the other data set. The datasets used in this research are also high-resolution-based.

In [16], the author developed a nonparametric approach for ITS from LiDAR data, detecting dominant and co-dominant trees for 94% and 62% for intermediate dead and overtopped trees. The overall accuracy achieved by the author is 77% which is not good for the high-resolution dataset. The author proposed the 5-step approach to segment the trees, which is time-consuming and did not provide good results. The performance of this method was also affected by the complexity of the forest terrain and the conditions of the vegetation.

In [17], the author uses a method to automate tree segmentation for individual trees from complicated urban forests. The developed method did not provide good accuracy, and the accuracy matrices are not provided, but it recognizes the difficulties of the urban areas. The suggested accuracy is lower than the natural forest areas.

In [19], a custom Individual Tree Matching (ITM) algorithm was used to compare LiDAR-detected trees with 284 field-measured reference trees. For local maxima methods, a fixed $3 \times 3\text{m}$ window applied to a non-smoothed canopy height model (CHM) achieved an F_1 -score of 0.65 with 86% of trees are detected, while methods based on [21] achieved excellent crown segmentation with mean crown *radius* $< 0.5\text{m}$ of field-measured crown radius. For non-local maxima methods, the adaptive mean shift algorithm (AMS3D) performed well with F_1 score of 0.67 and a mean crown *radius* $< 0.1\text{m}$.

In [18], the author proposed an approach to extract, detect, and segment individual crowns using multispectral airborne LiDAR data. Trees crowns are initially segmented in the spatial domain using the mean shift algorithm, under-segmented crowns are identified using a Support Vector Machine (SVM) classifier and geometric features, and the crowns identified from classification are refined using mean shift in a joint feature space with spatial and multispectral data. The experiments on a total of ten forest plots in Ontario, Canada, quantify the differences in SVM's multispectral space data, and improve the detection rate of dominant trees from 82% to 88% while having better accuracy for detection in dense and clumped forests.

Table I highlights that while high-resolution LiDAR and UAV-based imagery dominate the field, their effectiveness is highly context-dependent. Methods such as the SVM combined with Mean Shift and LiDAR demonstrated strong per-

formance in complex forest environments, particularly for clumped tree detection, achieving up to 84% segmentation accuracy. Techniques like Fuzzy Center Segmentation and the Watershed Transformer (U-Net) delivered inconsistent or suboptimal results despite using high-resolution data, pointing to limitations in handling diverse terrain and canopy structures. The Watershed Algorithm (WA), when applied to UAV LiDAR data in Eucalyptus plantations, achieved the highest F1-score (0.761), excelling particularly in low-density plots. This underscores the suitability of classical segmentation approaches in structured forest settings. Traditional methods such as the nonparametric 5-step approach, though reasonably accurate (77%), were found to be time-consuming and sensitive to terrain variations, limiting scalability. Urban ITS models consistently underperformed, highlighting persistent challenges in segmenting trees in built environments due to occlusions and background complexity. Deep learning-based methods like YOLO and U-Net offer promising segmentation capabilities but show mixed outcomes depending on dataset quality, annotation accuracy, and forest heterogeneity. The analysis reveals that no single approach outperforms others; instead, method effectiveness hinges on factors such as data resolution, forest density, canopy complexity, and the integration of geometric and machine learning strategies. These insights justify the development of hybrid models, such as the proposed U-Net with ResNet34 encoder, which aim to balance performance and generalizability, particularly when working with low-resolution RGB imagery in urban-natural mixed forest landscapes.

III. METHODOLOGY

We propose a hybrid deep learning model that combines the U-Net architecture with a ResNet34 encoder for tree segmentation from the German urban area. The model uses a U-Net architecture with the Resnet34 backbone and extracts features from the RGB satellite images. The proposed model use to train the dataset and applies the k-fold cross validation for 10 folds with the learning rate of $1e-2$ and patience of 10. Each fold runs for 100 epochs, early stopping technique is used to stop the training on the best Dice score. The proposed model consists of the U-Net framework, Resnet34 encoder, Decoder with skip connections, and output layer.

A. U-Net Framework

U-Net is a fully convolutional neural network developed for pixel-wise segmentation tasks. A U-Net is structured in a symmetric "U" shape with a contracting path (an encoder) and an expanding or decoding path with skip connections between layers with the same index in the contraction and expansion paths. The contracting path extracts contextual information when downsampling the image and increasing the depth of features, whereas the expanding or decoding path extracts spatial information when upsampling the image.

B. ResNet34 Encoder

ResNet34 is a residual network with 34 layers that uses residual blocks with skip connections that help to smooth the

TABLE I
COMPARISON OF INDIVIDUAL TREE SEGMENTATION (ITS) METHODS

Ref.	Method/Model	Dataset Location/Type	Resolution	Accuracy/Performance	Notes
2009, [12]	Tree top detection with RGB/LiDAR	Forest & urban areas	High/Low	Varies significantly with resolution	Tree tops used for segmentation
2025, [13]	CCD-YOLO (YOLO-based)	Beijing & Henan Univ., China	High	Accuracy not specified	YOLO-based ITS with high-res LiDAR
x2024, [14]	Fuzzy Center Segmentation (FCS)	Jasper Nat. Park, Canada (ALS/TLS)	High	Poor segmentation	Ineffective for high-res data
2024, [15]	Watershed Transformer (U-net)	Bengaluru, India & Gartow, Germany	High	India: Acc. 46.3%, IoU 71.2%; Germany: Acc. 52%, IoU 72.6%	Better in Germany
2016, [16]	Nonparametric 5-step	Forest terrain	High	Overall 77%; Dom. 94%; Co-dom. 62%	Time-consuming; terrain sensitive
2016, [17]	Automated urban ITS	Urban forests	High	Low accuracy; metrics not provided	Urban segmentation issues noted
2025, [18]	SVM + Mean Shift + LiDAR	Ontario, Canada (10 plots)	High	Det. improved 82% $\rightarrow$ 88%, Segm. 84%	Effective for clumped trees; limited spectral use
2024, [19]	ITM, AMS3D, Local Maxima	Subtropical forests	High	Local Maxima: F1=0.65, Det.=86%; AMS3D: F1=0.67, radius error < 0.1m	Limited by density (3 pts/m ²), GPS errors
2024, [20]	WA, LMA, EDCA, LSA (UAV LiDAR)	Eucalyptus plantations	High	WA: F1=0.761; CHM-based best	Sensitive to canopy/density; WA excels in low-density

gradient flow during training, as well as achieve a smooth gradient descent in smaller, deeper networks while avoiding the vanishing gradient dilemma in depth. The architecture allows for the ability to learn more complex features and tends to work better when lower-resolution imagery lacks fine detail. The model processes the input RGB images in different steps. In the first step, the model uses the 7x7 CNN layer, followed by batch normalization and the ReLU activation method. After that, a max pooling layer is used to minimize the spatial dimensions for all images. In the last four stages, residual blocks are applied to extract the high-level features. Every stage consists of multiple 3x3 convolutions to identify the shortcut, extract the best features, and repeat the process for the next stage.

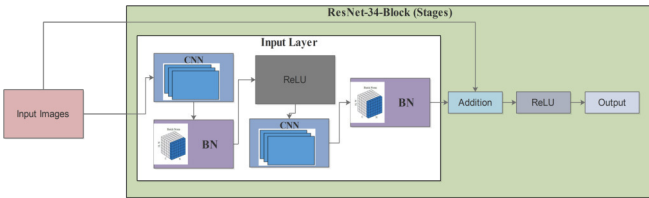


Fig. 1. The Input layer and stages of ResNet34 for Conv 7x7 layers, BatchNorm and ReLU

Figure 1 illustrates the architecture of a single ResNet-34 block, which serves as the encoder backbone within the hybrid deep learning model. The block operates on input RGB images and is composed of a series of convolutional, normalization, and activation layers structured around the concept of residual learning—a technique that enables the network to train effectively even as the number of layers increases, by preserving gradient flow and mitigating vanishing gradients. The process begins with the Input Layer, where the incoming image is passed through an initial Convolutional Neural Network (CNN) layer. This layer extracts low-level spatial features such as edges, color gradients, and textures. The resulting feature

maps are passed through a Batch Normalization (BN) layer, which normalizes the outputs across the batch dimension. This step stabilizes and accelerates the training process by reducing internal covariate shift. Next, the normalized feature maps are passed through a ReLU (Rectified Linear Unit) activation function, which introduces non-linearity and allows the model to learn complex representations. The output from ReLU then flows into another CNN layer, which further refines the extracted features. Again, a batch normalization step follows to maintain consistent learning dynamics. ResNet has the shortcut connection (or identity mapping), which allows the input to bypass one or more convolutional layers and be directly added to the output of those layers. The original input image is routed directly to the Addition block, where it is combined with the output from the second batch normalization layer. This residual connection helps preserve the identity function and ensures that the model can learn effectively even when deeper layers are less informative. After addition, another ReLU activation is applied to the combined feature map, and the result is passed to the output of the ResNet block. This output can then be forwarded to the next block in the encoder or to downstream modules, depending on the architecture. Stacked ResNet blocks across multiple stages enable the encoder to extract increasingly abstract and hierarchical features, making it effective for complex tasks like tree crown segmentation from low-resolution satellite imagery. This structure allows the model to efficiently learn both fine details and broader spatial context, which is critical for accurately delineating individual trees in heterogeneous landscapes.

C. Decoder with Skip Connections

The decoder output of the U-Net reconstructs the spatial dimensions of the image progressively with the help of deconvolutions or upsampling layers. At each deconvolution step, the encoder feature maps are connected to the corresponding decoder layer with skip connections. These skip connections

help to safeguard fine-scale spatial information that may have been lost when downsampling, and they help the model demarcate individual tree tops, particularly when they are closely spaced together in dense forest areas. The decoder consists of different blocks, and each block consists of 2x2 transposed convolution, concatenation with the corresponding encoder, and two 3x3 convolutional layers followed by batch normalization and ReLU activation.

D. Output Layer

The last layer of the model is a 1x1 convolution to map from feature maps to a single-channel binary mask representing the probability that a pixel belongs to an individual tree crown. The output is produced using a sigmoid activation function, providing output values within the range [0,1]. During post-processing, the values can be threshold to get binary segmentation maps.

E. Full Architecture

Figure 2 illustrates the complete architecture of the proposed hybrid deep learning model for individual tree segmentation, which combines a ResNet-34-based encoder with a U-Net-style decoder. This encoder-decoder framework is specifically designed for semantic segmentation tasks using low-resolution RGB aerial images, and it extracts both spatial and contextual features to generate pixel-level binary masks of tree crowns.

The architecture begins with the Input Layer, which receives preprocessed aerial images of size 256x256 pixels. These images are passed into Encoder Path, where they are processed through four sequential ResNet34-Blocks (Stage-1 to Stage-4). Each block has multiple convolutional layers, batch normalization, ReLU activations, and residual connections, enabling the network to learn hierarchical features while maintaining gradient stability. As the data flows deeper through the encoder stages, the spatial dimensions are progressively reduced while feature depth increases, capturing increasingly abstract and high-level semantic information. Skip connections are established from each ResNet stage to its corresponding decoder block, preserving fine-grained spatial features that might otherwise be lost during downsampling.

Following the encoder, the data is passed into the Decoder Path, which consists of four Decoder Blocks arranged in reverse order to the encoder stages. Each decoder block performs a combination of upsampling (via transposed convolutions or bilinear interpolation), concatenation with the corresponding encoder features (via skip connections), and convolutional layers to refine the upsampled feature maps. This path helps reconstruct the original image resolution while selectively enhancing regions corresponding to individual tree crowns. The decoder progressively restores the spatial structure of the image, integrating detailed edge information with high-level semantic understanding.

At the final stage of the architecture, the processed feature map is passed through a Final Convolutional Layer followed by a Sigmoid Activation Function, which transforms the output into a binary probability map. Each pixel in this output

represents the probability of belonging to a tree crown. A threshold (typically 0.5) is applied during post-processing to generate the final binary segmentation mask. These masks, as shown on the right side of the figure, effectively highlight tree structures within the urban or semi-natural landscape, with white or red representing detected tree areas and black for non-tree background.

This encoder-decoder model employs the powerful feature extraction capabilities of ResNet-34 alongside the spatial reconstruction strengths of U-Net, making it highly suitable for complex segmentation tasks on low-resolution imagery. The integration of skip connections is particularly important for maintaining localization accuracy in dense or heterogeneous forest and urban regions, leading to high-performance results across various segmentation metrics such as Dice coefficient, IoU, and precision-recall.

F. Performance Matrices

The following evaluation matrices are used to evaluate the performance of the proposed model:

The **Intersection-over-Union-(IoU)** is used for measuring the overlapping between mask images and the predicted mask:

$$\text{IoU} = \frac{|\text{Predicted Mask} \cap \text{Ground Truth (Mask)}|}{|\text{Predicted Mask} \cup \text{Ground Truth (Mask)}|} \quad (1)$$

where $\cap$ represents the intersection and $\cup$ represents the union between predicted segmentation and ground truth (Mask).

The **Dice Coefficient** is used for measuring the similarities between both predicted masks and ground truth:

$$\text{Dice} = \frac{2 \times |\text{Predicted Mask} \cap \text{Ground Truth}|}{|\text{Predicted Mask}| + |\text{Ground Truth}|} \quad (2)$$

where $\cap$ represents the intersection between the predicted segmentation and the ground truth.

Precision is used to compare the actual tree area and predicted tree area:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

where TP (True Positives) represents the number of correctly predicted positive samples, and FP (False Positives) represents the number of incorrectly predicted positive samples.

Recall is used to evaluate the actual tree area that is correctly detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

IV. DATASET AND SETUP

A. Data site

The data set is collected from the eastern area of Bochum, a city located in the federal state of North Rhine-Westphalia, Germany. Bochum is located in the Ruhr metropolitan area and has a strong industrial history and a diverse urban landscape. The eastern Bochum region, where the data is collected, consists of a wide range of residential areas, commercial areas, and greenery, providing a wonderful and representative sample of the region. The Ruhr area is highly urbanized but

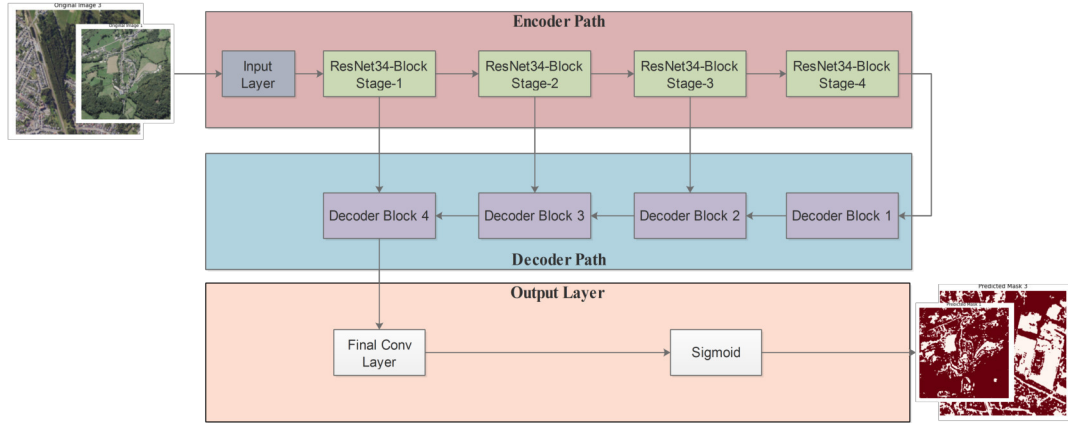


Fig. 2. A proposed hybrid model combines U-Net and Resnet34

quite dense. It consists of buildings, agricultural land, forests, and water areas. In recent decades, the region has changed significantly structurally. Heavy industries, such as coal and steel, have been replaced by other industries, causing serious changes in the landscape. This area has been affected by the closure of the Opel factory, which was driven by economic globalization in this city.

A study site is shown in Figure 3, which provides a multi-scale geographic overview of the study area, illustrating the location of Bochum city within the federal state of North Rhine-Westphalia, Germany. The rightmost panel (A) highlights North Rhine-Westphalia on the map of Germany, marking Bochum in red. Panel (B) shows a broader satellite view of the state with Bochum outlined, while panel (C) zooms into a high-resolution satellite image of Bochum, delineating its administrative boundaries in red. The study site is situated within the urban-natural landscape where the tree segmentation analysis was conducted.

B. Data Pre-processing

Data pre-processing is performed on 450 Digitale Orthophotos (DOP) images and 450 corresponding normiertes Digitales Orthophoto Modell (nDOM) images, collected from the GeoData Portal for the year 2023. Digitale Orthophotos is a German term meaning Digital Orthophotos, which are georeferenced aerial or satellite images that can be read and displayed using Geographic Information Systems (GIS) software. Both 450 DOP and 450 nDOM images are originally sized 20000x20000 pixels and converted into 256x256 pixels, with some noise added to create a blurring effect. An algorithm is proposed to convert all DOP and nDOM images into RGB format using Glymur and the Python Imaging Library (PIL/Pillow). The algorithm developed takes each image individually, processes it by converting it into Joint Photographic Experts Group (JPEG) format at 256x256 pixels, and stores it in a separate folder for further model implementation. After that, all DOPs and nDOMs are manually compared, and each DOP is renamed to match its corresponding nDOM for further tree area segmentation.

Figure 4 illustrates the data preprocessing workflow applied to DOPs and nDOMs, which are the primary input sources for the tree segmentation model. The preprocessing pipeline begins with the ingestion of raw high-resolution input images—each originally sized at $20,000 \times 20,000$ pixels. These inputs undergo several transformation steps to ensure compatibility with the deep learning model, improve computational efficiency, and simulate real-world data imperfections. The first stage, labeled "Get the Input," retrieves and parses the raw DOP and nDOM files, which are typically stored in JPEG 2000 (JP2) format. These files are then passed through a series of preprocessing operations. The raw images are resized from their original resolution to 256×256 pixels, significantly reducing memory load and speeding up training while maintaining sufficient detail for segmentation tasks. Additionally, image rotations are applied to introduce data augmentation, which enhances model generalization by exposing it to various spatial orientations of tree structures. Compression is another vital transformation step used to mimic the quality degradation commonly encountered in operational remote sensing data. This is followed by a format conversion process, where the images are transformed from JP2 to standard JPEG using the Python Imaging Library (PIL). PIL serves as the central processing engine in this pipeline, enabling all format handling, resizing, and augmentation operations. Once the core transformations are completed, the images are further processed by artificially introducing noise to simulate blurriness. This step is critical in emulating real-world conditions such as motion blur, atmospheric disturbances, or sensor imperfections, making the model more robust against such variations during inference. The final output of the pipeline consists of preprocessed JPEG and TIFF images that are uniformly scaled, augmented, and formatted. These prepared datasets serve as inputs to the hybrid deep learning segmentation model, enabling efficient training and evaluation under controlled yet realistic scenarios. The structured pipeline ensures data quality, uniformity, and robustness, all of which are crucial for achieving reliable performance in tree crown segmentation tasks using low-

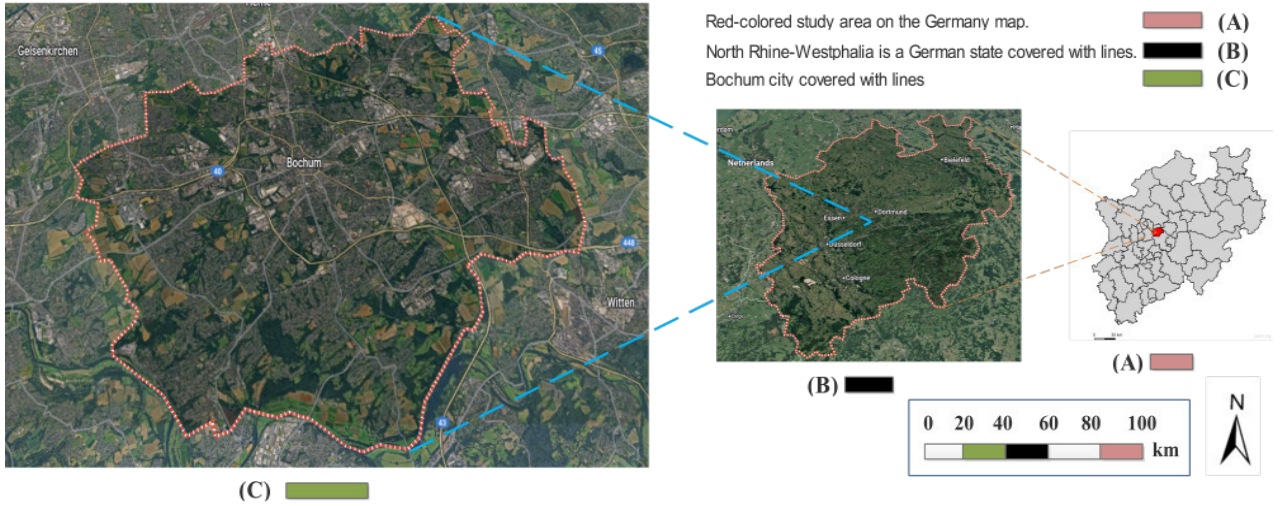


Fig. 3. (A): Indicate the study area in red color from the German Map, (B): Indicate the study area from the North Rhine-Westphalia state (C): The whole study area is located in the eastern part of Bochum

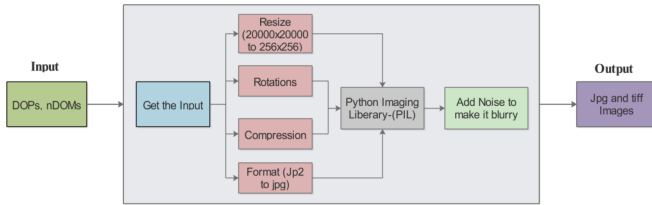


Fig. 4. Data pre-processing workflow using the PIL

TABLE III
THE SPECIFICATIONS OF EXPERIMENTAL SETUP

System Configuration	Hardware Specification
Programming Language	Python
Development Environment	Visual Studio
GPU	NVIDIA GeForce RTX 3080
CUDA Version	11.8
RAM	10 GB
Operating System	CentOS

resolution remote sensing imagery.

C. Experimental Setup

This experiment uses a DOPs- and nDOMs-based dataset for tree crown segmentation in a natural environment, aiming to evaluate the validity and effectiveness of the model's generalizability. The model was trained using hyperparameters given in Table II. The hardware specification for the experimental setup is shown in Table III

TABLE II
TRAINING HYPERPARAMETERS FOR MODEL DEVELOPMENT

Hyperparameter	Value
Learning Rate	1e-2 (with ReduceLROnPlateau)
Batch Size	16
Optimizer	Adam
Epochs	Up to 100 (early stopping on Dice score)
Validation Strategy	10-fold cross-validation

V. RESULTS AND DISCUSSIONS

This section presents and discusses the experimental results in detail. The data set is trained using the proposed model with 10-fold cross-validation. Each fold runs for up to 100 epochs, with early stopping applied to retain the best-performing model. The performance of the model is evaluated using

key metrics: Dice coefficient, Intersection over Union (IoU), Accuracy, Precision, and Recall. The model achieves its best results with a Dice score of 0.8678, an IoU of 0.7754, an accuracy of 0.8180, a precision of 0.8410, and a recall of 0.9103. With comparison of baseline values, the proposed model consistently improves the results in training and in the validation process. The results show that the proposed model performs very well. Table IV summarizes the performance of cross-validation using the same matrices as the developed model. Using a 10-fold cross-validation approach, the model's ability to generalize across different subsets of the data set is evaluated. During the training, the dice-coefficient and IoU are the most critical metrics in this situation because they quantify spatial overlap between predicted tree regions and hand-annotated tree crowns. A high Dice value indicates that the model has accurately segmented the tree areas with little over- and/or under-segmentation. The best Dice Coefficient (0.8585) and IoU (0.7598) scored in Fold 10 suggests that this instance the model is able to produce such accurate tree segmentation despite the difficulties associated with low-resolution input data.

Accuracy measures the total number of correctly classified pixels (trees and background) as a proportion. Although accuracy is not important, it can be informative and less sensitive when using skewed data. The Fold 10 again has the highest

TABLE IV
THE TRAINING AND VALIDATION RESULTS FOR EACH FOLD

Folds	Dice Co-efficient	IoU	Accuracy	Precision	Recall
Fold 1	0.7937	0.6613	0.7473	0.6928	0.9438
Fold 2	0.8133	0.7094	0.7577	0.7938	0.8949
Fold 3	0.8445	0.7322	0.7820	0.7847	0.9231
Fold 4	0.8448	0.7383	0.7729	0.7597	0.9667
Fold 5	0.7895	0.6799	0.7575	0.7716	0.8523
Fold 6	0.7895	0.6573	0.7260	0.6770	0.9547
Fold 7	0.8294	0.7102	0.7658	0.7560	0.9255
Fold 8	0.8363	0.7243	0.7834	0.7831	0.9118
Fold 9	0.7823	0.6459	0.7415	0.6736	0.9417
Fold 10	0.8585	0.7598	0.8091	0.8150	0.9169

accuracy (0.8091) supporting the model performance. The precision value measures the numbers of the predicted pixels for trees that were correct, so this is key in minimizing false positives, for example, with roads or shadows on trees. The model also achieved its highest precision (0.8150) in Fold 10, demonstrating that it accurately differentiates trees from other land features: houses, land area, roads, and other objects.

In recall, the false negative refers to the correctly detected number of tree pixels, as this is relevant in ecological or forestry applications when missing tree areas could affect the estimation of canopy coverage. Fold 4 had the best recall (0.9667), which means that the model detected nearly every actual tree pixel in Fold 4, although it was possible to add false positives. Although Fold 10 consistently produced the highest in all metrics used in this study, Fold 3 and Fold 4 consistently perform well after Fold 10 and gain the second highest results. Fold 9 had the lowest results for Dice (0.7823) and IoU (0.6459); it is probable that Fold 9 had more challenging image conditions, such as shadowing, overlapping trees, and/or poorer annotation quality. In the same vein, precision was at its lowest in Fold 9 (0.6736), which may indicate more confusion with non-tree elements.

The proposed hybrid model demonstrates solid and steady performance through every fold, with some minor variations that may be explained by the natural variance of low-resolution orthophoto imagery. The consistently high Dice and IoU scores indicate that the hybrid deep learning model is well-tailored for tree crown segmentation tasks with the inevitable constraints of resolution and noise in the input data. These results underscore the viability of deep learning models, particularly ResNet34 with a U-Net decoder, to use low-resolution, remote sensing data to extract important spatial features for forestry and ecological monitoring.

We tested our part of the dataset on the trained model and visualized the results including the original image, mask image, predicted mask, and overlapped image. The results for each sample are also measured with the in performance matrices. The results for the 15 samples are given in Table V. The model performs very well on sample 8 and achieved the highest Dice coefficient of 0.8995, IoU of 0.8174, accuracy of 0.8558, and precision of 0.8655, but sample 5 achieves the highest recall. During testing, the second sample produces the

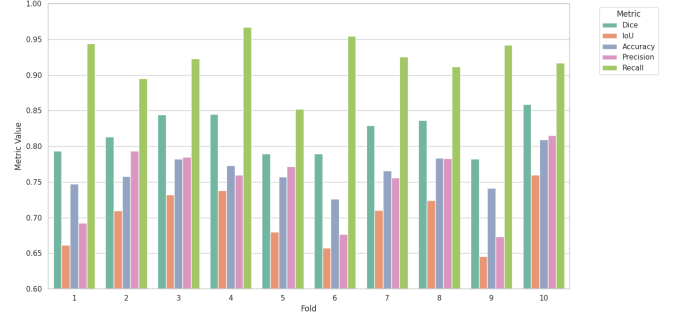


Fig. 5. Borplot visualization of model training and validation segmentation performance Metrics (Dice Coefficient, IoU, Accuracy, Precision, and Recall) across 10 folds

second largest results with a Dice coefficient of 0.8812, and an IoU of 0.7877.

TABLE V
THE TESTING RESULTS FOR 15 DIFFERENT SAMPLES

Samples	Dice Co-efficient	IoU	Accuracy	Precision	Recall
Sample 1	0.7220	0.5650	0.6410	0.5681	0.9903
Sample 2	0.6393	0.4693	0.5780	0.4732	0.9850
Sample 3	0.7328	0.5783	0.5942	0.5860	0.9777
Sample 4	0.7268	0.5709	0.6609	0.5798	0.9854
Sample 5	0.8812	0.7877	0.8009	0.7903	0.9958
Sample 6	0.7162	0.5579	0.5734	0.5594	0.9952
Sample 7	0.8328	0.7135	0.7645	0.7638	0.9155
Sample 8	0.8995	0.8174	0.8558	0.8655	0.9362
Sample 9	0.8600	0.7544	0.8464	0.8052	0.9228
Sample 10	0.7705	0.6266	0.7006	0.6332	0.9837
Sample 11	0.7901	0.6531	0.6609	0.6571	0.9906
Sample 12	0.7707	0.6269	0.6332	0.6325	0.9862
Sample 13	0.8588	0.7525	0.7890	0.7600	0.9871
Sample 14	0.8047	0.6733	0.7294	0.6836	0.9780
Sample 15	0.7380	0.5833	0.6977	0.5867	0.9900

The results are summarized visually in Figure 6. The model achieves high recall, meaning it is very sensitive to identifying tree regions, though at the cost of reduced precision. The Dice Coefficient and Accuracy metrics confirm its overall reliability in segmentation tasks, while the IoU values reflect room for improvement in terms of spatial precision, especially under complex visual conditions. These results validate the effectiveness of the hybrid ResNet34–U-Net architecture in tree segmentation from low-resolution aerial images, while also suggesting directions for refinement—particularly in improving boundary sharpness and reducing false positives.

VI. CONCLUSIONS

We proposed a hybrid deep learning model that combines the ResNet34 encoder with a U-Net decoder architecture to segment individual trees from low-resolution RGB orthophoto images (DOPs) over the German urban and semi-urban landscape. Despite the limitations of low-resolution imagery, the model demonstrated high segmentation performance, achieving a Dice coefficient of 0.8678, IoU of 0.7754, accuracy of 0.8180, precision of 0.8410, and an exceptional recall of

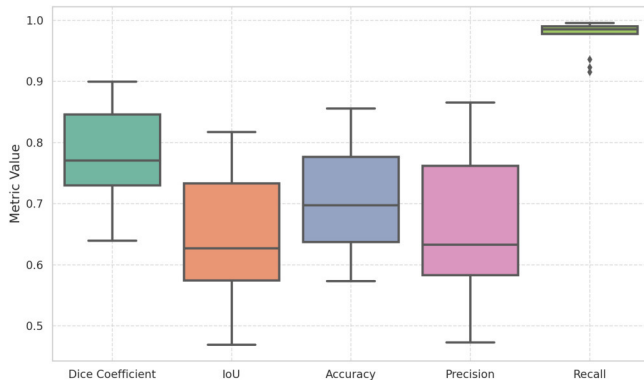


Fig. 6. Boxplot visualization of segmentation performance metrics (Dice Coefficient, IoU, Accuracy, Precision, and Recall)

0.9103. These results confirm the model's strong ability to detect and delineate tree crowns across diverse environments, including dense forest patches and complex urban settings. The preprocessing pipeline—consisting of format conversion, resizing, augmentation, and noise injection—played a critical role in preparing the dataset and enhancing model generalizability. Visualizations of predicted masks and overlay comparisons further validated the model's effectiveness, showing a high degree of alignment with ground truth annotations, particularly in structured and less cluttered regions.

Comparative analysis with existing state-of-the-art individual tree segmentation (ITS) methods revealed that our approach is competitive even against models relying on high-resolution LiDAR or UAV data, making it a cost-effective alternative for large-scale forest monitoring in data-limited regions. The model performed well overall, but challenges remain in improving segmentation precision and handling complex urban-object boundaries, suggesting opportunities for future work on post-processing refinement and attention-based enhancements.

DATA AVAILABILITY

The data set is collected from the GeoData Portal of the Federal State of North Rhine-Westphalia under the data license "Deutschland - Zero - Version 2.0" for the year 2023.

REFERENCES

- [1] D. Rajasugunasekar, A. K. Patel, K. B. Devi, A. Singh, P. Selvam, and A. Chandra, "An integrative review for the role of forests in combating climate change and promoting sustainable development," *International Journal of Environment and Climate Change*, 2023.
- [2] R. Damaševičius and R. Maskeliūnas, "Modeling forest regeneration dynamics: Estimating regeneration, growth, and mortality rates in lithuanian forests," *Forests*, vol. 16, no. 2, 2025.
- [3] —, "Adaptive sensor clustering for environmental monitoring in dynamic forest ecosystems," *Peer-to-Peer Networking and Applications*, vol. 18, no. 3, 2025.
- [4] R. Damaševičius, G. Mozgeris, A. Kurti, and R. Maskeliūnas, "Digital transformation of the future of forestry: an exploration of key concepts in the principles behind forest 4.0," *Frontiers in Forests and Global Change*, vol. 7, 2024.
- [5] T. Mijit, E. Firkat, X. Yuan, Y. Liang, J. Zhu, and A. Hamdulla, "Lr-seg: A ground segmentation method for low-resolution lidar point clouds," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 347–356, 2024.
- [6] D. Joshi and C. Witharana, "Vision transformer based unhealthy tree crown detection and evaluation of annotation uncertainty," 2025.
- [7] L. Wallace, A. Lucieer, and C. S. Watson, "Evaluating tree detection and segmentation routines on very high resolution uav lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7619–7628, 2014.
- [8] M. D. Hossain and D. Chen, "Remote sensing image segmentation: Methods, approaches, and advances," *Remote Sensing Handbook, Volume II*, pp. 117–144, 2025.
- [9] H. Chen, W. Li, J. Gu, J. Ren, H. Sun, X. Zou, Z. Zhang, Y. Yan, and L. Zhu, "Low-res leads the way: Improving generalization for super-resolution by self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 857–25 867.
- [10] B. Koonce, "Resnet 34," in *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*. Springer, 2021, pp. 51–61.
- [11] J. Shen, Q. Xu, M. Gao, J. Ning, X. Jiang, and M. Gao, "Aerial image segmentation of nematode-affected pine trees with u-net convolutional neural network," *Applied Sciences*, vol. 14, no. 12, p. 5087, 2024.
- [12] J. Reitberger, C. Schnörr, P. Krzystek, and U. Stilla, "3d segmentation of single trees exploiting full waveform lidar data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 6, pp. 561–574, 2009.
- [13] Y. Liu, A. Zhang, and P. Gao, "From crown detection to boundary segmentation: Advancing forest analytics with enhanced yolo model and airborne lidar point clouds," *Forests*, vol. 16, no. 2, p. 248, Jan. 2025.
- [14] Z. Xi, C. Hopkinson, and L. Chasmer, "Supervised terrestrial to airborne laser scanner model calibration for 3d individual-tree attribute mapping using deep neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 209, pp. 324–343, 2024.
- [15] M. Freudenberg, P. Magdon, and N. Nölke, "Individual tree crown delineation in high-resolution remote sensing images based on u-net," *Neural Computing and Applications*, vol. 34, no. 24, p. 22197–22207, Aug. 2022.
- [16] H. Hamraz, M. A. Contreras, and J. Zhang, "A robust approach for tree segmentation in deciduous forests using small-footprint airborne lidar data," *International journal of applied earth observation and geoinformation*, vol. 52, pp. 532–541, 2016.
- [17] C. Zhang, Y. Zhou, and F. Qiu, "Individual tree segmentation from lidar point clouds for urban forest inventory," *Remote Sensing*, vol. 7, no. 6, pp. 7892–7913, 2015.
- [18] S. Li, S. Zhao, Z. Tian, H. Tang, and Z. Su, "Individual tree segmentation based on region-growing and density-guided canopy 3d morphology detection using uav lidar data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [19] T. Saeed, E. Hussain, S. Ullah, J. Iqbal, S. Atif, and M. Yousaf, "Performance evaluation of individual tree detection and segmentation algorithms using als data in chir pine (*pinus roxburghii*) forest," *Remote Sensing Applications: Society and Environment*, vol. 34, p. 101178, 2024.
- [20] Y. Yan, J. Lei, J. Jin, S. Shi, and Y. Huang, "Unmanned aerial vehicle–light detection and ranging-based individual tree segmentation in eucalyptus spp. forests: Performance and sensitivity," *Forests*, vol. 15, no. 1, p. 209, 2024.
- [21] M. Dalponte and D. A. Coomes, "Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data," *Methods in ecology and evolution*, vol. 7, no. 10, pp. 1236–1245, 2016.

Flexible and Scalable Results Collecting in Distributed Spatial Simulations

Piotr Aksamit
AGH University of Krakow
paksamit@student.agh.edu.pl

Mateusz Najdek
AGH University of Krakow
najdek@agh.edu.pl

Wojciech Turek
AGH University of Krakow
wojciech.turek@agh.edu.pl

Abstract—Among many distributed spatial simulation systems each has its own approach to the problem of results collecting and analysis. The volume of results can be huge, while not all results are finally needed. The presented solution is to provide a unified form of defining the range of data to be collected and the methods for efficiently collecting them during the simulation runtime. Simulation results can be represented as a stream of records, where every record has the same structure. This observation means, that simulation can specify one or more data schema, equivalent of the *CREATE TABLE* command in an SQL database. Then data selection and analysis comes down to writing proper *SELECT* statements. The paper describes three main parts of the proposed, SQL-inspired results collecting method: parsing and query analysis, distributed computing and integrating all parts together. The method has been integrated with the *HiPUTS*, a distributed urban traffic simulator.

I. INTRODUCTION

COMPUTER simulation is a powerful method for conducting research on spatial-temporal systems. Spatial simulation can be successfully used for understanding behaviors of individual animals forming flocks [1], for predicting crowded locations during buildings evacuations [2], for understanding geological structures created ages ago by microscopic sea shellfish [3], for optimizing urban traffic [4] or public transport [5]. In general, these methods are based on mathematical models of a physical environment and physical entities existing in the environment. The simulation algorithm updates the state of the entities and the environment as the simulated time passes.

All spatial simulations generate large volumes of data during the simulation process, as the state of all the entities changes over the simulated time. Storing the complete record of these changes can significantly influence the performance of the computation. Considering an exemplary simulation of 1 million people in a city, each characterized with e.g. 30 bytes of variable state, we can expect over 1 terabyte of results after simulating only 1 hour of their life at 10 frames per second. Such volumes cannot be stored in memory and require time-consuming disk operations.

Typically, in order to draw the necessary conclusions from the simulation, not all the results are needed. The researcher may need only the state changes record of a selected subset of entities, which is an option available in popular simulators [6]. More often the desired result of the simulation computation is an aggregated value of selected state changes. This includes

simple statistics and distributions of observed states over time and space, which are typically the basis for drawing valuable conclusions. Such aggregations can be computed after the simulation process finishes, however, this approach again requires all the results to be collected. It would be far more convenient to be able to compute the needed aggregates during the simulation. This approach is relatively hard to generalize, therefore it typically requires implementing specific extensions to the simulator's source code.

The problem becomes even more complex when the distribution of the simulation computation is considered. Simulating complex models of large-scale scenarios requires using many computing nodes simultaneously in order to receive the results in a reasonable time and to fit the model in the available memory. Many contemporary spatial simulation tools support distributed computing [7], [8], [9], [10]. Although distributed collection of results can be an efficient and scalable solution, calculations of aggregated values by separate computing nodes, which are responsible for different fragments of the simulated environment, is a significant challenge.

In the presented work we address the problem of simulation results collecting and aggregating during distributed computing of spatial simulations. We propose a general approach, based on the Structured Query Language, SQL, for defining the range of required data. The proposed approach does not require the user to extend the simulator's source code. In fact it does not require knowing any programming language. It is based on a standard and relatively simple query syntax, used for defining a data model and querying the model afterwards. The query syntax allows for defining simple and complex queries, including spatial and temporal limits, selecting entities of specific features or states, identifying relations between entities and computing many types of aggregates.

We present an abstract, Java-based implementation of this approach, which can be integrated with many different simulation tools. For evaluation purposes it has been integrated with an urban traffic simulator. In this important field, the simulation is a basis for the majority of research [4].

The proposed solution is also capable of working in distributed environments. While preserving the simplicity of queries, all the features have been successfully implemented for working with distributed processing of simulation models.

In the following section we present the existing approaches to the problem of results collecting in distributed spatial sim-

ulations. The next two sections describe the proposed solution from the user perspective and its internal implementation details.

II. EXISTING APPROACHES

The problem of results collecting, aggregating and processing is inextricably connected with all large-scale simulations, therefore it is not novel. An in-deep analysis of its elements and possible solutions can be found in [11], where authors identify four crucial challenges: selection, collection, storage and retrieval. The two considered strategies for data selection: all vs partial, have their advantages and drawbacks. Collecting all data can be justified only when rolling-back the simulation is required or detailed after-action review is needed. In other cases selection should be performed during the simulation runtime. In the context of data collection the problem of scalability has been pointed out with distributed collection as the possible solution.

An agent-based architecture for collecting simulation data has been presented in [12]. It provides services that facilitate data collection and analysis within a distributed simulation. The discrete event simulation models are considered in this case. The proposed approach is compared to the "baseline" methodology, where sub-models report the data to a central database for output analysis. Not surprisingly, a significant improvement over the centralized approach has been reported.

A few years later a similar problem was addressed using the Web Services, which apparently gained popularity in this period of time. The authors point out that "The data collection system should place minimal stress on the simulation infrastructure from both a computational load and communications overhead perspective".

A comparison between centralized, partially and fully distributed data collection methods has been presented in [13]. The evaluation is limited to the recommended, fully decentralized method, which achieves $O(N)$ complexity, guaranteeing good scalability.

The issue of aggregating simulation results has been identified in [14], where a geo-distributed simulation has been considered. The proposed solution involved using Hadoop, a map-reduce computational framework for processing the data. The same platform has been used in [15], where data processing was executed in distributed manner without the need for centralization. Obviously not all types of aggregations can be computed this way.

Unfortunately, the conclusions, design concepts and particular solutions described above, have not been incorporated into the existing spatial simulation tools. The problem of data collecting and processing seems to receive less attention than the modeling, simulation and computations distribution. In the Flame simulator [8], a user can configure two types of outputs: "snapshots" of the complete simulation state or "agents", which results in saving selected entities only. In addition, the number of iterations between saved states can be specified. No aggregations are available. In the D-MASON framework [7], each computing node (each model partition) writes its

results to its own file. One can use a statistics output streams to save self-implemented aggregates. The Pandora simulation system [10] saves all the data from of the environment and agents collected during each time step in two files. The authors provide a separate program (implemented in Python) to further analyze the data.

In the REPAST HPC [9] the most advanced data collection functionalities can be found. A user can define so called "Aggregate data collection", which use several reduction functions (sum, min, max, etc) to compute basic aggregates in the fly. The aggregates can compute the values from remote computing nodes using MPI and store them in a single file. Unfortunately, the configuration of data collection has to be defined in the source code, using the provided API.

In the field of urban traffic simulation, which is an exemplary use case considered in this work, the problem of data collection and analysis also receives relatively little attention. One of the most popular simulators, SUMO [16], offers a wide variety of so called *outputs*, which cover selected types of entities, e.g. all cars states or all lane-change events. It also offers a fixed list of aggregates, like average trip speed or route length. The user cannot select specific fragments of space or specify more complex operations on the simulation results.

Another popular tool for urban traffic simulation is MAT-Sim [17], which is constructed over a simpler, queue-based traffic model. It allows storing the simulation results in several csv files, containing information about all simulated trips. Further analysis is to be performed outside the simulation.

SMARTS [18] is an urban traffic simulator designed to perform distributed computations. It implements several advanced features related to model partitioning and local communication between workers. It demonstrates good scalability of simulation distribution. However, the results collecting mechanisms assumes sending cars' trajectories to a centralized server, which stores these in files. Such approach turns out to be a scalability bottleneck when hundreds of cores are used simultaneously.

The problem of results collecting and aggregating in distributed spatial simulations is clearly visible in the domain. Unfortunately, it does not receive enough attention, leaving a user with a complex problem to solve.

III. PROPOSED RESULTS COLLECTING SUBSYSTEM

Every simulation generates different types of data. This means that the simulation defines its own data schema, composed of the state description of all simulated entities, which changes over time. This schema can be queried using SQL-like *SELECT* statements.

The proposed solution is abstract and can be applied in various simulations. In order to adopt it to a particular case, we need to define multiple "dictionary" tables, where data do not change during simulation, and one "events" table that keeps both references to "dictionary" tables and changeable variables. The concept will be presented using a specific case of a urban traffic simulation.

We have combined our idea with the *HiPUTS* Simulator [19], which is a distributed urban traffic simulator¹. It allows us to divide the environment model of simulation (road map) into separate fragments and assign the responsibility of computation over those fragments to separate processes, which can use many nodes of a computing cluster. The simulation process is divided into small time quanta called steps. In each step, the number of cars can be different in separate workers, as they can move freely from one to the other.

The environment model in the simulation is composed of two key components: lanes and junctions. It is depicted as a directed graph, where the edges represent the lanes and the nodes correspond to the junctions. Every lane starts and ends with junction. Each junction has a fixed, known position, while each lane is a straight segment linking junctions. As a result, every lane has a defined length, and junctions are aware of the circular arrangement of incoming and outgoing lanes. From the vehicle's perspective, it tracks details such as its current lane and position, speed, acceleration, and other characteristics like its length, maximum speed, and designated route, enabling it to navigate accordingly. The state of vehicle changes during the simulation due to natural movement, so parameters such as e.g. speed, acceleration and position on lane are modified and represent singular state of car in a specific time unit of the simulation. This specific car is simulated by the corresponding worker, depending on its current location.

Here we have three "dictionary" tables - *Cars*, *Junctions* and *Lanes* and one "events" table - *Events* that tracks all changes. The dictionary tables contains static data that are constant during the simulation and they are not modified. The event table, in turn, contains dynamic data, such as information about changes of individual vehicles in simulation. The simulation model definition of SQL "dictionary" and "events" tables is presented in Listing 1.

To unambiguously define car in time and place we need to use triple:

(*WorkerId*, *StepNumber*, *CardId*)

```

1 CREATE TABLE Cars
2 (
3     CarId bigint PRIMARY KEY,
4     MaxSpeed numeric NOT NULL,
5     Length numeric NOT NULL
6 );
7
8 CREATE TABLE Junctions
9 (
10    JunctionId bigint PRIMARY KEY,
11    Longitude numeric NOT NULL,
12    Latitude numeric NOT NULL
13 );
14
15 CREATE TABLE Lanes
16 (
17    LaneId bigint PRIMARY KEY,
18    IncomingJunctionId bigint NOT NULL
19    references Junctions(JunctionId),
20    OutgoingJunctionId bigint NOT NULL

```

```

21    references Junctions(JunctionId),
22    Length numeric NOT NULL
23 );
24
25 CREATE TABLE Events
26 (
27    WorkerId text NOT NULL,
28    StepNumber bigint NOT NULL,
29    CarId bigint NOT NULL
30    references Cars(CarId),
31    LaneId bigint NOT NULL
32    references Lanes(LaneId),
33    PositionOnLane numeric NOT NULL,
34    Speed numeric NOT NULL,
35    Acceleration numeric NOT NULL,
36    PRIMARY KEY(WorkerId, StepNumber, CarId)
37 );

```

Listing 1. Equivalent simulation model definition in SQL tables.

For simplicity let's define the following view (Listing 2), that combines all listed table above 1. Data analysis is ultimately reduced to executing *SELECT* queries on the *SimulationDatas* view. As previously mentioned simulations can be run on multiple computing instances, referred to as *Workers*, with each instance tasked with processing a specific portion of the map. The configuration file allows to define the format in which the analysis results should be exported. Currently the module supports exporting data in CSV [20] and Parquet [21], [22] formats, which has been found to be well-suited for data visualization thanks to their structure and the ability to facilitate detailed analysis across different contexts.

```

1 CREATE VIEW SimulationDatas AS
2
3 SELECT
4     e.*, c.MaxSpeed, c.Length,
5     l.Length AS LaneLength,
6     ji.JunctionId AS IncomingJunctionId,
7     ji.Longitude AS IncomingJunctionLongitude,
8     ji.Latitude AS IncomingJunctionLatitude,
9     jo.JunctionId AS OutgoingJunctionId,
10    jo.Longitude AS OutgoingJunctionLongitude,
11    jo.Latitude AS OutgoingJunctionLatitude
12 FROM Events e
13 JOIN Lanes l ON l.LaneId = e.LaneId
14 JOIN Junctions ji
15     ON ji.JunctionId = l.IncomingJunctionId
16 JOIN Junctions jo
17     ON jo.JunctionId = l.OutgoingJunctionId
18 JOIN Cars c ON c.CarId = e.CarId

```

Listing 2. Definition of SQL view for data in simulation.

Example. A trivial simulation has 2 steps. In first step there are 2 cars (*C1*, *C2*), in second step 3 cars (*C2*, *C3*, *C4*). Cars *C1*, *C2* are always on worker *W1*, rest on worker *W2*. Then *Events* table has 5 rows, two rows from first step and three rows from second step. Then *Events* table looks as presented in Table 1 (for simplicity we present only columns: *WorkerId*, *StepNumber*, *CardId*):

Available syntax:

```

1 SELECT [expression [ [ AS ] column_name ] ] [, ...]
2 [FROM SimulationDatas]
3 [WHERE condition]
4 [GROUP BY grouping_element [, ...] ]
5 [HAVING condition]

```

¹<https://github.com/hiputs/HiPUTS>

TABLE I
EXEMPLARY OVERVIEW OF FRAGMENT OF SOME COLLECTED EVENTS
DATA

WorkerId	StepNumber	CarId	...
W1	1	C1	...
W1	1	C2	...
W1	2	C2	...
W2	2	C3	...
W2	2	C4	...

```

6 [ORDER BY expression [ ASC | DESC ] [ NULLS { FIRST
  | LAST } ] [, ...] ]
7 [LIMIT count]
8 [OFFSET start]
9
10 where grouping_element is one of:
11   (
12     expression,
13     ( expression, [, ...] )
14   )
15 where expression is one of the following:
16 - number, text, NULL,
17 - scalar function e.g. COALESCE, CONCAT, ABS
18 - aggregates e.g. MIN, MAX, SUM, AVG, COUNT
19 - arithmetic operators: +, -, *, /, %
20 - text operators: ||
21 - conditional operators: OR, AND
22 - relational operators: =, <, >, <=, >=, <>,
23   LIKE, REGEX, BETWEEN, IN, IS [NOT] NULL,
24   IS [NOT] DISTINCT
25 - conditional statement:
26   CASE
27     WHEN expression_1 THEN result_1
28     [WHEN ...]
29     [ELSE else_result]
30   END
31 - column name: (defined by simulation)
32 - casting: expression::type or CAST(expression,
33   type)
34 - parentheses expression: (expression)
35 - signed expression: +, -, ~

```

Listing 3. Available syntax for collecting and aggregating data using this method.

IV. SYNTAX DESCRIPTION

The syntax for our system is inspired by PostgreSQL² and follows a familiar structure for querying and manipulating data. In the following, we describe each clause supported by our solution and provide explanations for the available expressions and operators.

A. SELECT Clause

The SELECT clause specifies the data to retrieve from the database. Each expression can optionally be renamed using the AS keyword:

- `expression [ [ AS ] column_name ]`: Defines what data to retrieve (e.g., column names, computed values).
- `AS column_name`: Renames the result of the expression.

²<https://www.postgresql.org/docs/current/sql-select.html>

B. FROM Clause

The FROM clause specifies the source of the data. In our case, `SimulationDatas` defines the combined "dictionary" tables and "events" table.

C. WHERE Clause

The WHERE clause filters the rows returned by the SELECT clause according to a specified condition:

- `WHERE condition`: Applies filters using relational (=, <, >, etc.) and logical operators (AND, OR, etc.).

D. GROUP BY Clause

The GROUP BY clause aggregates rows with the same values in specified columns:

- `GROUP BY grouping_element`: Groups the rows by one or more columns (or expressions), often used with aggregate functions.

E. HAVING Clause

The HAVING clause filters the aggregated results produced by GROUP BY:

- `HAVING condition`: Similar to WHERE, but operates on aggregated data.

F. ORDER BY Clause

The ORDER BY clause specifies the sorting order of the results:

- `ORDER BY expression [ ASC | DESC ] [ NULLS { FIRST | LAST } ], [, ...]`: Orders the results by an expression or multiple expressions in ascending (ASC - default) or descending (DESC) order, and determines the placement of NULL values.

G. LIMIT and OFFSET Clauses

- `LIMIT count`: Limits the number of rows returned by the query.
- `OFFSET start`: Skips the first start rows before returning the rest.

H. Expressions and Operators

Various expressions and operators can be used within the above clauses to perform more complex calculations and conditions:

1) *Scalar Functions*: Scalar functions perform operations on individual values. Currently implemented functions:

- `ABS(expression)`: Returns the absolute value of a numeric expression. e.g. `ABS(-5)` returns 5.
- `CEIL(expression)`: Rounds a numeric expression up to the nearest integer. e.g. `CEIL(4.3)` returns 5.
- `COALESCE(expression1, expression2, ...)`: Returns the first non-NULL expression from the list. If all expressions are NULL, it returns NULL. e.g. `COALESCE(NULL, 'default', 'value')` returns 'default'.
- `CONCAT(expression1, expression2, ...)`: Concatenates multiple expressions into a single string.

e.g. `CONCAT('Hello', ' ', 'World')` returns 'Hello World'.

- `CONCAT_WS(separator, expression1, expression2, ...)`: Concatenates multiple expressions into a single string skipping NULL values, with a specified separator between them. e.g. `CONCAT_WS('-', '2024', '10', '23')` returns '2024-10-23'.
- `FLOOR(expression)`: Rounds a numeric expression down to the nearest integer. e.g. `FLOOR(4.7)` returns 4.
- `LENGTH(expression)`: Returns the number of characters in a string. e.g. `LENGTH('OpenAI')` returns 6.
- `LOWER(expression)`: Converts all characters in a string expression to lowercase. e.g. `LOWER('Hello World')` returns 'hello world'.
- `ROUND(expression, decimal_places)`: Rounds a numeric expression to the specified number of decimal places. e.g. `ROUND(3.14159, 2)` returns 3.14.
- `SQRT(expression)`: Returns the square root of a numeric expression. e.g. `SQRT(16)` returns 4.
- `SUBSTR(expression, start_position, length)`: Extracts a substring from a string expression starting at `start_position` and continuing for `length` characters. e.g. `SUBSTR('HiPUTS', 2, 3)` returns 'iPU'.
- `UPPER(expression)`: Converts all characters in a string expression to uppercase. e.g. `UPPER('hello world')` returns 'HELLO WORLD'.

2) *Aggregate Functions*: Aggregate functions operate on sets of rows and return a single result:

- `MIN()`, `MAX()`: Return the minimum or maximum value.
- `SUM()`: Adds numeric values.
- `AVG()`: Computes the average.
- `COUNT()`: Counts rows.

3) *Operators*: **Arithmetic Operators**:

- `+`, `-`, `*`, `/`, `%`: Perform mathematical operations.

Text Operators:

- `||`: Concatenates strings.

Conditional Operators:

- `AND`, `OR`: Combine conditions.

Relational Operators:

- `=`, `<`, `>`, `<=`, `>=`, `<>`: Compare values.

Specialized Operators:

- `LIKE`, `REGEX`: Perform pattern matching.
- `BETWEEN`, `IN`, `IS [NOT] NULL`: Check ranges, membership, or null values.
- `IS [NOT] DISTINCT`: Check for distinction between two values.

I. Conditional Statements

The CASE expression allows conditional logic:

- `CASE WHEN expression THEN result [WHEN ...] [ELSE else_result] END`: Evaluates conditions and returns the corresponding result.

J. Casting and Parentheses

Casting:

- `expression::type`: Casts an expression to a specific type.
- `CAST(expression AS type)`: Another way to perform the cast.

Parentheses and Signed Expressions:

- Parentheses control the order of operations within expressions.
- Signed expressions: `+`, `-`, `~`: Apply positive, negative, or bitwise complement to an expression.

Below are presented some example of usage of this solution:

1) Problem 1: What was speed of car with id = 3.

```
1 SELECT Speed, WorkerId, StepNumber
2 FROM SimulationDatas
3 WHERE CarId = 3
```

Listing 4. Query for problem 1.

2) Problem 2: Collect velocity of car with id = 3 when it was on lanes L1, L2 and L3.

```
1 SELECT LaneId, Speed
2 FROM SimulationDatas
3 WHERE LaneId IN ('L1', 'L2', 'L3')
```

Listing 5. Query for problem 2.

3) Problem 3: For each car what was its average speed in every 10 steps.

```
1 SELECT
2     CarId,
3     StepNumber / 10 AS Start,
4     StepNumber / 10 + 9 AS End,
5     AVG(Speed)
6 FROM SimulationDatas
7 GROUP BY CarId, StepNumber / 10
```

Listing 6. Query for problem 3.

4) Problem 4: Calculate average/minimum/maximum speed separately for very slow - up to 5m/s, slow - to 10m/s, medium - to 15m/s, fast - to 20m/s, very fast - above 20m/s

```
1 SELECT
2     CASE
3         WHEN Speed < 5 THEN 'Very slow'
4         WHEN Speed < 10 THEN 'Slow'
5         WHEN Speed < 15 THEN 'Medium'
6         WHEN Speed < 20 THEN 'Fast'
7         ELSE 'Very fast'
8     END,
9     MIN(Speed) AS min,
10    AVG(Speed) AS avg,
11    MAX(Speed) AS max
12 FROM SimulationDatas c
13 GROUP BY
```

```

14 CASE
15     WHEN Speed < 5 THEN 'Very slow'
16     WHEN Speed < 10 THEN 'Slow'
17     WHEN Speed < 15 THEN 'Medium'
18     WHEN Speed < 20 THEN 'Fast'
19     ELSE 'Very fast'
20 END

```

Listing 7. Query for problem 4.

5) Problem 5: Which cars had average speed greater than 100 in 10 subsequent steps

```

1 SELECT
2     CarId, AVG(Speed) AS speed,
3     StepNumber / 10 AS stepnumber_from,
4     StepNumber / 10 + 9 AS step_number_to
5 FROM SimulationDatas c
6 GROUP BY CarId, StepNumber / 10
7 HAVING AVG(Speed) > 100

```

Listing 8. Query for problem 5.

K. Optimization

1) *Skipping simulation step*: Consider the following example.

```

1 SELECT StepNumber, AVG(Speed)
2 FROM SimulationDatas
3 WHERE StepNumber % 100 = 0
4 GROUP BY StepNumber

```

WHERE clause in this example does not depend on vehicle, lane or junction parameters. When for the given simulation step *WHERE* clause is always false, we can skip whole step without checking condition for each vehicle in this step. Before every simulation step we can execute the following procedure.

Algorithm 1 Procedure checking whether *WHERE* clause is always false

```

1: condition ← WHERE clause
2: if condition is empty then
3:     return false
4: end if
5:
6: for each column ∈ SimulationDatas do
7:     change all occurrences of column in condition with
       "NULL"
8: end loop
9:
10: condition ← evaluated condition
11: if condition is empty then
12:     return false
13: end if
14:
15: return condition

```

By conducting a preliminary check before starting the calculations for each simulation step, it can greatly enhance the

program's performance. This approach allows the aggregator to determine whether collection of results for a particular step are necessary. If they are not essential—for example, when it is only interested in data calculated every 100 simulation steps—the aggregator can skip unnecessary calculations during the intervening steps. This optimization conserves computational resources, reduces execution time, and improves overall efficiency by focusing only on the simulation steps that yield relevant data.

2) *Short-circuit evaluation*: Our module stops calculating boolean expression combined by *AND*, *OR* operators, when value can be predicted by already computed sub-expressions for given patterns:

- **OR** - when expression or part of expression consists of multiple sub-expressions joined by *OR* we can predict result whenever any or sub-expression is *true*
- **AND** - when expression or part of expression consists of multiple sub-expressions joined by *AND* we can predict result whenever any or sub-expression is *false*

In most programming languages, this optimization is a common practice to improve computational performance when evaluating logical expressions with multiple conditions connected by logical operators like **AND** or **OR**. By implementing short-circuit evaluation, the program determines the result of the entire expression based on the initial conditions. If the outcome is already clear after evaluating the first few conditions, it skips the unnecessary computation of the remaining ones. This not only saves processing time but also enhances the overall efficiency of the program. For instance, in an **AND** operation, if one condition evaluates to false, the entire expression is false, and there is no need to check the subsequent conditions. This technique is especially beneficial in complex logical statements where some conditions might be resource-intensive to evaluate.

V. IMPLEMENTATION

Query computation consists of several steps:

- 1) parsing - check, whether query has valid syntax
- 2) analyze - check query type, it can be *InlineResult* - which means that result of data does not depends from simulation e.g. *SELECT 1 + 2*. Otherwise we have to split the query into two parts. The first one, that will be calculated during simulation, and the second one that will combine results together.

Consider the following query:

```

1 SELECT CarId / 2, 3 * AVG(Speed + 3)
2 FROM Simulation
3 WHERE CarId % 10 = 0
4 GROUP BY CarId / 2
5 HAVING MIN(Speed) > 10

```

Here we can specify the following parts:

- 1) *AVG(Speed + 3)* - in order to calculate average, we will calculate separately sum and count of expression *Speed + 3*

- 2) *GROUPBYCarId/2* - we have to calculate each aggregate independently by expression *CarId/2*.
- 3) $3 * AVG(Speed + 3)$,
 $MIN(CASEWHEN CarId > 0 THEN Speed$
 $ELSE 2 * Speed END)$ - in this query we have two aggregate functions, during simulation we will not know value of this expressions until simulation finishes and we gather data together. After compacting data we can then execute remaining scalar expressions. For expression $3 * AVG(Speed + 3)$ we will calculate $SUM(Speed + 3)$, $COUNT(Speed) + 3$ during simulation. After simulation finishes we will determine value of $AVG(Speed + 3)$ and after combining data from every computational node we can calculate $3 * AVG(Speed + 3)$
- 4) *WHERE CarId % 10 = 0* - where clause can be calculated during simulation, because it contains only scalar expressions.

In general computational nodes will keep the following tuples:

(*GroupingKey*, *Aggregate*₁, *Aggregate*₂, ..., *Aggregate*_N)

(*CarId/2*, *AVG(Speed + 3)*, *MIN(Speed)*)

After simulation finishes we have to take all tuples and compact them together. Our aggregator uses following aggregate functions: *SUM*, *MIN*, *MAX*, *COUNT*, *AVG*. Compacting *SUM*, *MIN*, *MAX*, *COUNT* is pretty straightforward, only for *AVG* we need to keep separately *COUNT* and *SUM*.

A. Memory control

We cannot predict amount of data that will be gathered, so we have to exchange data between RAM and hard drive. Json/XML serializers wouldn't be efficient, that's why we decided to define custom binary serialization and deserialization.

Simple types like *integer*, *text*, *bigint*, *boolean* have implemented serialization/deserialization methods in every language, so let's focus on more complicated data structures.

Consider the following query, where *CarId* is *bigint*, *Speed* is *double*.

```
1 SELECT CarId, AVG(Speed)
2 FROM SimulationDatas
3 GROUP BY CarId
```

During simulation computational nodes have to keep *CarId* and *AVG(Speed)*. For *CarId* = 5, current value of *AVG(Speed)* = 2.5, where *SUM(Speed)* = 10, *COUNT(Speed)* = 4, tuple

(*CarId*, *AVG(Speed)*)

will be represented by the following sequence of bits.

Bits	Meaning
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000101	5 - bigint
01000000 00100100 00000000 00000000 00000000 00000000 00000000 00000000	10 - double
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000100	4 - bigint

B. Configuration

The analyzer can be launched and configured using a dedicated configuration file that contains its parameters definition. Example definition is presented in Listing 9.

```
1 analyzerConfiguration:
2   bufferSize: 1000
3   storageType: in-memory
4   numberOfSegments: 8
5   nodeSizeOfIndexTree: 16
6   levelsInIndexTree: 4
7   export:
8     enabled: true
9     format: parquet
10    path: /path/to/export/directory
11    configExportEnabled: false
12  queries:
13    - SELECT MIN(Speed) + MAX(Speed)
14      FROM SimulationDatas
15      WHERE CarId::bigint IN (2, 34)
16      GROUP BY CarId
17      HAVING COUNT(1) > 5
18    - SELECT AVG(Speed) FROM SimulationDatas
```

Listing 9. Example configuration file for analyzer.

The configuration file is a crucial component that allows users to customize and control the behavior of the application. It consists of different parameters that define how the module operates, enabling users to tailor the performance and functionality to meet specific needs. These parameters cover various aspects and are defined as:

- 1) **bufferSize** - The size of the buffer when writing data from the Worker to a temporary file and reading data from it by the PostMaster.
- 2) **storageType** - An enumerated value indicating whether the data processed by the Workers is stored on the hard drive (drive) or in RAM (in-memory).
- 3) **numberOfSegments** - A numerical value specifying the number of independent segments into which HashMap structures are divided, ensuring concurrent operation of the program. Each segment has separate locks for reading and writing. Too few segments will cause bottlenecks during parallel usage, while too many will increase memory overhead.
- 4) **nodeSizeOfIndexTree** - The HashMap structure uses an IndexTree instead of an array to store hash keys. IndexTree is similar to a SparseArray. This means unused elements do not occupy memory space. The structure does not perform rehashing operations, so it cannot be larger than the initial value.
- 5) **levelsInIndexTree** - The number of levels in the IndexTree structure.

- 6) **allocateStartSize** - The initial size of the `HashMap` structure in bytes.
- 7) **allocateIncrement** - The number of bytes by which the size of the `HashMap` structure is increased.
- 8) **queries** - An array of SQL queries.
- 9) **export**:
 - **enabled** - A boolean value indicating whether the results should be exported to a file.
 - **format** - An enumerated value specifying the data export format (`parquet` or `csv`).
 - **path** - The relative path to the folder where the results will be exported. In the specified folder, a new folder with the export date is created, into which a file named `result<query_number>` with the extension specified in the configuration file is generated.
 - **configExportEnabled** - A boolean value indicating whether the configuration file should be exported along with the data.

When choosing parameters, it is important to assess that the maximum number of elements in the `HashMap` structure will be equivalent to:

$$\text{numberOfSegments} \times \text{nodeSizeOfIndexTree}^{\text{levelsInIndexTree}}$$

VI. THIRD PARTY LIBRARIES

Our implementation uses two mainly used libraries *JSqlParser* and *MapDb*

A. JSqlParser

For parsing we used publicly available library called *JSqlParser*³. It is an open-source library written in Java that enables analysis, manipulation, and processing SQL queries in both text and object formats.

Some of the key features of this library include:

- SQL Query Parsing - the ability to convert SQL queries from their textual form into an object-oriented data structure that is easy to work with.
- Query Manipulation - allows to modify, delete, or add parts of SQL queries using an object interface.
- SQL Query Generation - enables to create SQL queries using objects and convert them back into text form.

With *JSqlParser*, developers can effortlessly manipulate and analyze SQL queries within their applications.

B. MapDb

For memory control We use library *MapDb*⁴ that provides Maps, Sets and other collections backed by off-heap or on-disk storage. It is a hybrid between java collection framework and embedded database engine. Library is concurrent-safe and provides support for *ACID* transactions. Due to the fact that it can use on-disk storage, library requires to provide serialization & deserialization implementation of used items.

Storage type and other parameters that *MapDb* enables can be configured through config file.

VII. CONCLUSIONS AND FURTHER WORK

In this paper, we presented a SQL-inspired method for flexible and scalable result collection in distributed spatial simulations. By representing simulation data as streams of records conforming to defined schemas, we enable users to specify the range of data to be collected and the computations to be performed using standard SQL query syntax. This approach abstracts the complexities of data collection and aggregation in distributed environments, allowing users to focus on analysis without modifying the simulator's source code or writing additional programs.

We implemented this method within HiPUTS, a distributed urban traffic simulator, demonstrating its practicality and efficiency. Our evaluation showed that the approach introduces minimal time and memory overhead while providing significant flexibility in data collection and aggregation. The optimization techniques, such as skipping simulation steps, short-circuit evaluation and custom binary serialization & deserialization, further enhance performance by reducing unnecessary computations.

The proposed method addresses a significant gap in existing spatial simulation tools, which often lack advanced mechanisms for data selection and aggregation, especially in distributed settings. By enabling SQL-like queries, our solution simplifies the process of obtaining meaningful insights from large-scale simulations, which is crucial for researchers and practitioners working with complex models and massive datasets.

For future work, we plan to explore several directions.

A significant area we intend to explore is the utilization of simulation results during runtime by different entities within the simulation, such as vehicles in a traffic model, or by the load balancer itself. By enabling entities to access aggregated or filtered simulation data in real-time, we can enhance the fidelity and adaptability of the simulation. For example, vehicles could adjust their behaviors based on current traffic conditions derived from aggregated data, leading to more realistic modeling of traffic flow and congestion patterns. This dynamic interaction would allow for the simulation of advanced scenarios, such as adaptive cruise control or real-time route optimization.

Similarly, the load balancer could use real-time simulation metrics to dynamically adjust the distribution of computational load across workers. By monitoring the simulation results, the load balancer can identify hotspots or regions with increased computational demands and reallocate resources accordingly. This approach can improve the efficiency and scalability of the simulation by ensuring balanced workloads and minimizing processing delays.

Machine Learning Integration, where exploring the incorporation of machine learning techniques for predictive analytics within the simulation framework could enable advanced func-

³<https://github.com/JSQLParser/JSqlParser>

⁴<https://mapdb.org/>

tionalities like anomaly detection and trend prediction directly during simulation runtime.

In conclusion, the proposed method addresses a critical challenge in distributed spatial simulations by offering a user-friendly, efficient, and scalable solution for data collection and aggregation. By empowering users to specify precisely what data they need and how it should be processed, we facilitate more effective and focused analysis. We believe that further development along the outlined directions will enhance the system's capabilities and broaden its adoption, ultimately contributing to more advanced and insightful simulation studies across various fields.

ACKNOWLEDGMENTS

The research presented in this paper was funded by the National Science Centre, Poland, under the grant no. 2019/35/O/ST6/01806. We gratefully acknowledge Poland's high-performance Infrastructure PLGrid ACK Cyfronet AGH for providing computer facilities and support.

REFERENCES

- [1] R. De Nicola, L. Di Stefano, O. Inverso, and S. Valiani, "Modelling flocks of birds from the bottom up," in *Leveraging Applications of Formal Methods, Verification and Validation. Adaptation and Learning*, T. Margaria and B. Steffen, Eds. Cham: Springer Nature Switzerland, 2022. ISBN 978-3-031-19759-8 pp. 82–96.
- [2] M. De Iuliis, E. Battagazzorre, M. Domaneschi, G. P. Cimellaro, and A. G. Bottino, "Large scale simulation of pedestrian seismic evacuation including panic behavior," *Sustainable Cities and Society*, vol. 94, p. 104527, 2023. doi: <https://doi.org/10.1016/j.scs.2023.104527>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210670723001385>
- [3] P. Topa, Łukasz Faber, J. Tyszcza, and M. Komosinski, "Modelling ecology and evolution of foraminifera in the agent-oriented distributed platform," *Journal of Computational Science*, vol. 18, pp. 69–84, 2017. doi: <https://doi.org/10.1016/j.jocs.2016.07.009>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877750316301168>
- [4] S. S. S. M. Qadri, M. A. Gökçe, and E. Öner, "State-of-art review of traffic signal control methods: challenges and opportunities," *European transport research review*, vol. 12, pp. 1–23, 2020.
- [5] M. B. K. Kubiak and R. Długosz, "Solutions for planning smart hybrid public transportation system—pożnan agglomeration as a case study of satellite towns' connections," in *Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems*, 2019, p. 67.
- [6] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.
- [7] G. Cordasco, V. Scarano, and C. Spagnuolo, "Distributed mason: A scalable distributed multi-agent simulation environment," *Simulation Modelling Practice and Theory*, vol. 89, pp. 15–34, 2018. doi: <https://doi.org/10.1016/j.simpat.2018.09.002>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1569190X18301230>
- [8] M. Holcombe, S. Coakley, and R. Smallwood, "A general framework for agent-based modelling of complex systems," in *Proceedings of the 2006 European conference on complex systems*, vol. 1. European Complex Systems Society Paris, France, 2006.
- [9] N. Collier, J. Ozik, and C. M. Macal, "Large-scale agent-based modeling with repast hpc: A case study in parallelizing an agent-based model," in *Euro-Par 2015: Parallel Processing Workshops: Euro-Par 2015 International Workshops, Vienna, Austria, August 24–25, 2015, Revised Selected Papers 21*. Springer, 2015, pp. 454–465.
- [10] X. Rubio-Campillo, "Pandora: a versatile agent-based modelling platform for social simulation," *Proceedings of SIMUL 2014, The Sixth International Conference on Advances in System Simulation*, pp. 29–34, 2014.
- [11] P. A. Wilcox, A. G. Burger, and P. Hoare, "Advanced distributed simulation: a review of developments and their implication for data collection and analysis," *Simulation Practice and Theory*, vol. 8, no. 3, pp. 201–231, 2000. doi: [https://doi.org/10.1016/S0928-4869\(00\)00023-9](https://doi.org/10.1016/S0928-4869(00)00023-9). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0928486900000239>
- [12] S. S. Y. Xu and F. Ciarallo, "An agent-based data collection architecture for distributed simulations," *International Journal of Modelling and Simulation*, vol. 24, no. 2, pp. 55–64, 2004.
- [13] E. Kaya and F. E. Sevilgen, "A fully distributed data collection method for hla based distributed simulations," in *Proceedings of the 2009 Summer Computer Simulation Conference*, 2009, pp. 337–347.
- [14] K.-T. Yao, R. F. Lucas, C. E. Ward, G. Wagenbreth, and T. D. Gottschalk, "Data analysis for massively distributed simulations," in *Interservice/Industry Training, Simulation, and Education Conference (IITSEC)*, 2009, pp. 2–32.
- [15] Y. Wu and G. Gong, "A fully distributed collection technology for mass simulation data," 06 2013. doi: 10.1109/ICCIS.2013.438 pp. 1679–1683.
- [16] H. Chen, K. Yang, S. G. Rizzo, G. Vantini, P. Taylor, X. Ma, and S. Chawla, "Qarsumo: a parallel, congestion-optimized traffic simulator," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 578–588.
- [17] A. Horni, K. Nagel, and K. W. Axhausen, "Introducing matsim," in *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016, pp. 3–7.
- [18] K. Ramamohanarao, H. Xie, L. Kulik, S. Karunasekera, E. Tanin, R. Zhang, and E. B. Khunayn, "SMARTS: Scalable microscopic adaptive road traffic simulator," *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, pp. 1–22, 2016.
- [19] W. T. Mateusz Najdek, Natalia Brzozowska, "Hiputs: Super-scalable simulation of microscopic continuous urban traffic model," in *Proceedings of the 39th ECMS International Conference on Modelling and Simulation*, 2025, pp. 476–485.
- [20] Y. Shafranovich, "Common Format and MIME Type for Comma-Separated Values (CSV) Files," RFC 4180, Oct. 2005. [Online]. Available: <https://www.rfc-editor.org/info/rfc4180>
- [21] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, "Dremel: interactive analysis of web-scale datasets," *Proceedings of the VLDB Endowment*, vol. 3, no. 1–2, pp. 330–339, 2010.
- [22] D. Vohra and D. Vohra, "Apache parquet," *Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools*, pp. 325–335, 2016.

Enhancing Research Data Integrity Through Blockchain: Design and Implementation of a Web Based Management System

Sikandar Ali*

*School of Advanced Studies, Center of Neuroscience
University of Camerino
Email:sikandar.ali@unicam.it

Roberto Ciccocioppo[†], Massimo Ubaldi[†],
Andrea Morichetta[‡], Matteo Piersantelli[§]

[†] School of Pharmacy, Center for Neuroscience, Pharmacology Unit

[‡] School of Science and Technology, Computer Science Division

^{†‡} University of Camerino, [§] AM Microsystems Srl.

Email: {roberto.ciccocioppo@unicam.it, massimo.ubaldi@unicam.it, andrea.morichetta@unicam.it, info@am-microsystems.com}

Abstract—In this digital age, data driven research has become indispensable. Ensuring secure and transparent experimental data management remains a key challenge. Blockchain technology with its decentralized architecture, immutability and resistance to tampering, presents a viable solution to these problems. This paper presents a blockchain based web management system designed to streamline the handling and integrity of experimental research data. Leveraging the capabilities of blockchain technology enables the system to ensure data integrity. A critical aspect of research is data, by securely storing file hashes on decentralized network. This blockchain based web-system allows users to login with its credentials and upload experimental data in form a csv file and associate it with respective universities and supervisors of the user. Later the system allows the supervisor to login with its credentials to access the file for further analysis. Upon file upload, the system will calculate the hash of the file and stored on blockchain network. This approach guarantees the immutability and authenticity of the uploaded research data. The system prevents tampering and ensures transparency of the experimental data, which is paramount in academic research. The proposed system addresses the growing need for secure and efficient data management in research, providing a reliable solution for maintaining data integrity throughout the research cycle.

I. INTRODUCTION

BLOCKCHAIN is a digital distributed ledger. The functionality of blockchain is continually increasing because of decentralized networks, lack of reliance on trust, unchangeable storage and ability to share information anonymously. In experimental research, particularly in behavioral neuroscience, maintaining the integrity and authenticity of data is very important. This paper presents a web-based system designed to store and manage experimental data in csv file, which can range in GBs. These files capture critical data which are translated into results, such as the drinking pattern of rats, observed through a micro controller-based system. This experimental data includes parameters like device ID, timestamp and liquid consumption, which are crucial as every recorded value impacts the final analysis. These datasets, often stored as

CSV file, must remain unaltered to ensure unbiased, reliable results. After all, this data forms the basis for published finding, peer review, future studies. Even minor alteration could mislead entire research communities wasting time and resources. To address these challenges, we have developed a blockchain based web management system that not only sure data integrity but also make sure the traceability of data. Blockchain locks records in place, making unauthorized changes impossible. Every adjustment if allowed gets permanently logged, so anyone can trace the data's history. This transparency builds trust, journals and collaborators can verify that the numbers/data haven't been tweaked post experiment. More importantly, it enforces reproducibility, a cornerstone of good science. If other labs can't replicate results because the original data was altered, the entire study lost credibility. With research ethics under increasing scrutiny, blockchain offers a simple fix, which is immutable proof that data stays fair from collection to publication. The data stored in blockchain is immutable and transparent for the whole network [1]. Blockchain technology is changing and enables alternative approaches like tamper proof credentialing and decentralized learning system [2]. Currently many institutions still rely on collecting data either on paper or centralized data management systems. However, this form of medical system is highly prone to privacy breach [3]. Therefore, the transformation of centralized data management system to decentralized data management system is an irresistible trend [4]. However, most of the system stores and maintain their experimental data either on their server or on papers it also wastes a lot of resources [5]. Therefore, in this paper a blockchain based system has been proposed, which maintains experimental data integrity and the key feature of the system is blockchain technology for data authentication. When file uploaded on server it will calculate the hash of the file and stored on blockchain network. The data (hash of the file) will be stored via smart contracts. The smart contract is a digital set of rules that are automatically executed when predetermined terms and conditions are met. The smart contract was written in solidity language and deployed on

This work was supported by AM Microsystems srl

network, so web system can communicate with blockchain through it. The system calculates the hash of experimental data file and stored on server as well as on blockchain networks and then for authentication, it compares both hashes of a file which is stored on server as well as on blockchain network. If the hash matches that means data is authenticated successfully and file is unaltered. Otherwise, the system will flag the file as altered while showing error message on screen. This ensures that any unauthorized modification to data is immediately detected. This system provides a robust mechanism of data integration. This system combines server-based technology for storage and access control whereas blockchain technology is used for authentication.

II. LITERATURE REVIEW

Continuous and rapid development of blockchain technologies, Different features have been proposed for web-based system for maintaining the integrity[6]. A blockchain based web 3.0 system has been developed to manage the educational documents and certificates, particularly the issuance of student degree. The certificates and degrees have been issued to students without the use of centralized system. The verification process has been conducted by using Ethereum smart contracts and degree has been issued to students and as it's a decentralized system, it maintains data integrity [7]. Similarly, In another blockchain-based health contract management system has been proposed for dealing the health-related certificates by interacting with smart contracts. In this study, a blockchain based system has been proposed by the author to negotiate contracts between health insurance companies and end user. Once the contract has been finalized then it will be stored on blockchain network, because of its immutable feature [8]. In another related paper, Author has developed a time release encryption system, in which information has been stored using asymmetric key encryption without needing help from external agents. However, the study has some limitations such as changing the difficulty based on prime numbers . Additionally, the researchers discovered some new methods such as proof of semantics to develop web [9]. A decentralized blockchain based medical record system has been proposed to handle EHR. Med-Rec architecture has a modular design in which authorization, administration privileges and data sharing are among the participants [10]. Med-block was a block chain-based hybrid architecture to protect electronic health records. The architectural node of the system is divided into storing nodes, submission nodes and endorsement nodes [11]. A generic blockchain based architecture has been proposed for storing patient electronic health records [12]. EHR blockchain architecture secures the environment and stops, tempering of electronic health record by tracking all the events in blockchain networks [13]. In their study, researcher has proposed Blockchain based system called men shared. The system was able to minimize the risk of data privacy and can be used to solve the problem of data sharing among data custodians in an untrusted environment[14]. An alternative method was used by researchers, in which a private blockchain

based data sharing scheme has been proposed. This system has used a consortium blockchain to save the index of security [15]. A data management and sharing platform has been proposed by combining artificial intelligence and blockchain based technologies. This platform has used the transparency of the zone chain for data tracking, and it also ensures the data remain unaltered [16]. This paper has proposed hybrid approach, ensures compatibility with existing management systems, maintains operational efficiency and provides tamper proof verification of data existence at a specific time. The blockchain timestamps server as definitive, publicly auditable evidence of data priority, effectively deterring misappropriation and protecting intellectual property claims prior to the dissemination of results.

III. METHODOLOGY

The blockchain based web management system has been developed for collecting experimental data. The front end of the web system has been developed by using html5 and css with Bootstrap. The back end up of the system has been developed in PHP core. The smart contract has been written in solidity language and deployed on Ethereum blockchain via sepolia testnet. Smart contracts will act as a bridge between blockchain network and web-based system. The system is mainly divided into two parts, one experimental data management system and the other part is validating the experimental data through blockchain technology to maintain its integrity.

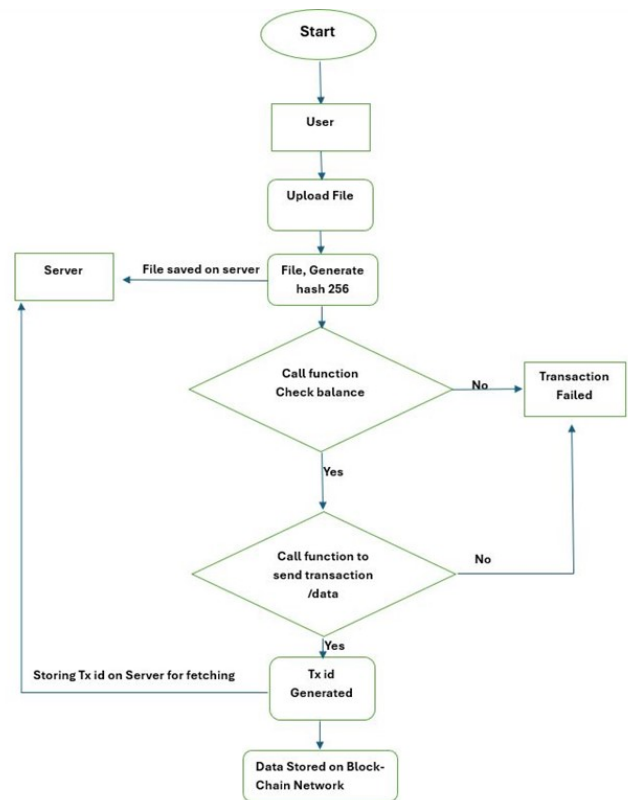


Fig. 1. Shows Flow chart for Storing Hash on Blockchain Network.

Figure 1 shows the flow chart in which the authorized user (student) can login with their credentials and upload their experimental data file on server with other details related to experiment. The file will be selected, and other experimental data will be written and uploaded. The system will calculate the hash of the file before uploading the file to the server. The file is then uploaded on the server and the database has been updated. Calculated hash will then be sent on blockchain network by calling a function. This function will communicate with smart contract, which was written in solidity language and deployed on ethereum testnet sepolia. The transaction (TX) has different parameters like nonce, gas price, gas limit, to (smart contract), value, data and chain id. The check balance function will be called which make sure the gas price would be enough for the transaction. If the gas price is not enough then it will generate errors of not enough gas price. If the price is enough then the system will generate a new transaction and signed with private key and sent onto the network. If the transaction is not successful then it will generate errors, otherwise the system will generate TX (Transaction) id through consensus mechanism, which will be stored in database.

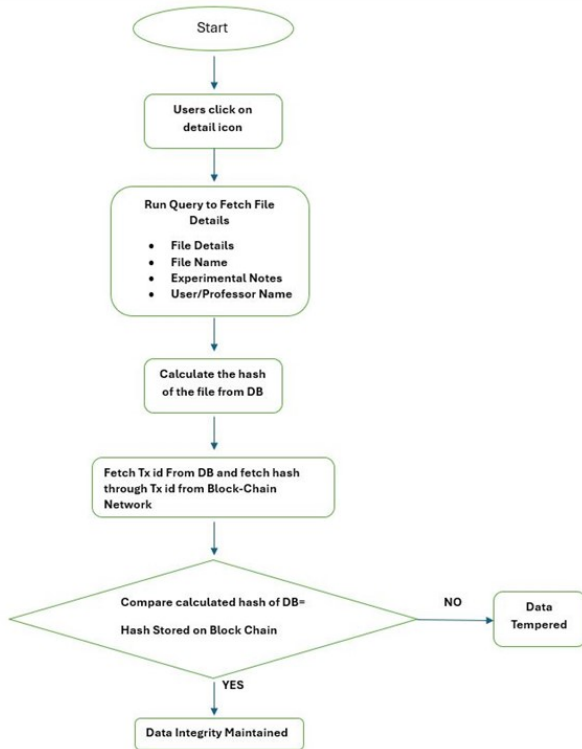


Fig. 2. Shows Flow chart of Data Integrity Process.

Figure 2 shows the flow chart for web management system for data authentication. The user (Professor->hierarchy level two) will click on details. It will run the SQL query to fetch the details related to that file such as file name, experimental notes and supervisor name. The system calculates the hash of the fetched file from server/database and also fetch its transaction id. The system will use web3 API to retrieve the

hash stored on blockchain network by using tx id. The system will compare the calculated hash of the file which is retrieved from the server with the hash stored on Ethereum blockchain network. If they matches, the system will display message of data integrity, which means data is intact, otherwise it will display an error message saying data is altered. The system has been divided into three level hierarchies, which were administration (admin), student and Professor. These three different hierarchies have different permission levels.

IV. IMPLEMENTATION

This section explains the implementation of the tool as well as explains the different hierarchical level and how GUI base authentication process works for authenticating the experimental data.

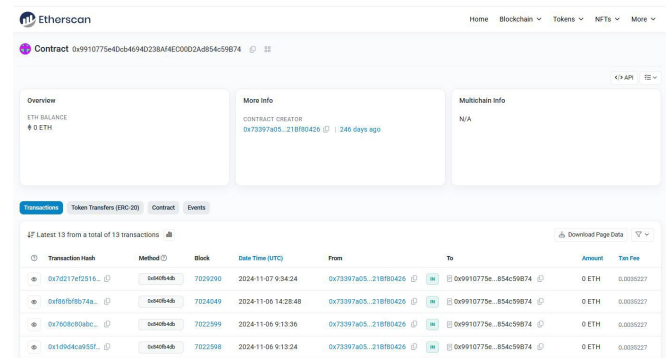


Fig. 3. Shows the successful deployment of smart contracts on blockchain networks.

Figures 3 shows the successful deployment of smart contracts in blockchain networks. To ensure smooth communication between the web-based management system and the blockchain network, the smart contract was written. The smart contract was developed to facilitate the hash storage and retrieval of hash from the blockchain network.

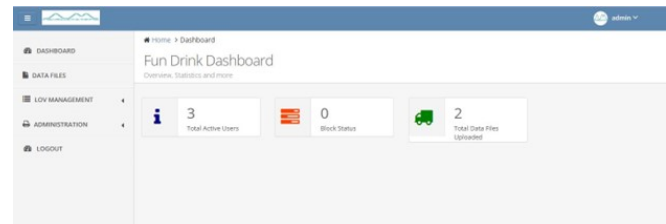


Fig. 4. shows the GUI of admin.

Figure 4 shows the admin graphical user interface. The administration has the ultimate power in this system. The admin can see all the data uploaded in the system and can see the detailed information such as total no of active users, total number of files uploaded, and has authority to block users. He can update the institution lists as well as create new users and assign them one of the hierarchies according to their position.

Fig. 5. Shows the GUI of student's experimental data entry Form.

Figure 5 shows the graphical user interface of the student view. The authorized student can log into their account and upload the conducted experimental file by selecting his supervisor. The user can also add more details like experiment name, total number of hours, experiment start and end dates, and can add notes in experimental detail section. Once the file has been uploaded then user will no longer has access or authority to delete file from system. Only user has access to read the files he has uploaded. He can only view them.

Fig. 6. Shows the GUI of Professor View Portal.

Figure 6 shows the graphical user interface of the professor's view, in which he can look into details of uploaded file with blockchain authentication. This detailed view shows the details of the experiments like the name of the student who has conducted this test, his email, his institution, uploaded date and the professor's name with additional experimental notes and it shows the statement that this file has been authenticated through a blockchain network, which shows the data integrity. The professor can analysis the file after downloading it or on the web by clicking analysis web icons for short files up to 10,000 data points.

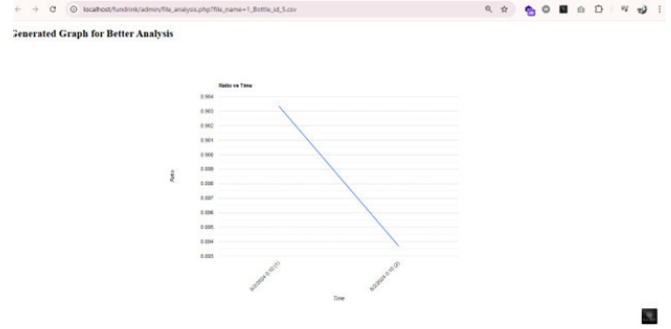


Fig. 7. Shows the data analysis of file on website.

Figure 7 shows the GUI of web-based tool analysis. This web-based system can be used to analysis the experimental data. The system can plot the graphs up to ten thousand points but after that it will run out of cache memory, therefore it halts. Detailed analysis of the file could be done through MATLAB or by using python-based script for analysis by using Matplotlib libraries.

V. VALIDATION OF TOOL

This section shows the validation of the implemented web management system that shows the error free implementation of blockchain technology, which ensures the integrity of experimental data stored by the user. This section presents the different key stages of processes like generation of Transaction id (TX ID), hash comparison and data integrity verification process.

A. Transaction ID Generation

The system will calculate the hash upon uploading the experimental data and store the file hash on blockchain network. A transaction id TX will be generated as proof that data (hash) has been recorded on blockchain network. Figure 8 shows the generated transaction id on blockchain network, and this transaction id has been stored in database for further reference.

Fig. 8. Shows transaction records on block chain network.

B. Verification on the Block Chain Network

To verify the integrity of the stored experimental data. The system will retrieve the hash stored on the blockchain networks

and compare it with the hash of the file stored on the server. Figure 9 shows a message stating that the hash has been successfully matched, indicating that the experimental data have not been altered.

Block Chain File Status:

Data has been Authenticated from Block Chain Network

Fig. 9. Shows the message conforming hash match and experimental data integrity.

If the hash doesn't match in that case the system will detect the potential tampering in experimental data and alert the user by message on screen. Figure 10 shows message, which states that the blockchain hash doesn't match with hash of the file stored on server.

Block Chain File Status:

Data has been plagiarized and not been Authenticated from Block Chain Network

Fig. 10. Shows the message indicating data tempering.

C. System Performance

The system has successfully maintained the data integrity of experimental data by using blockchain based web system. The use of transaction id for fetching the data and hash comparison method ensures that any unauthorized modification to the experimental data can be detectable. The results confirm the system has achieved its primary goal of data integrity.

VI. CONCLUSION

The development of blockchain based web management system has been successfully implemented for storing and managing experimental data. This system addresses the critical challenges of maintaining data integrity by using blockchain technology. The use of cryptographic hashing provides the robust mechanism for detecting unauthorized modification in data. The system ensures the integrity of data by verifying it. The validity of the system has been conducted through series of tests like calculation of hash while uploading file secondly generation of transaction id then storing the cryptographic hash on blockchain network and verification of data integrity through hash comparison. The validation process shows that the system successfully identifies the tampered data, while conforming to the integrity of experimental data files, which were unaltered. This system has a strong need in academic and research fields where the reliability of experimental data is paramount. In future work. This system should focus on the scalability of the system so that the system can handle large datasets. In addition, exploring the implementation of other decentralized technologies that could improve the security and efficiency of the system. Overall, this project shows the

potential of blockchain technology in ensuring data security / integrity and providing a particular solution for managing experimental data in research environment.

VII. ACKNOWLEDGMENT

We thank AM Microsystems for the collaboration to this study. The author Sikandar Ali was supported by fellowship from the Eureka program of the Region Marche, Italy. The work was supported by PRIN-PNRR2022 -P202274WPN (to RC), PRIN-2022 20227HRFPJ (to RC), MNESYS (PE0000006)- Project AMSUD 2024, PRIN 2022X9X5MS (to MU).

REFERENCES

- [1] Z. Liu, Y. Xiang, J. Shi, P. Gao, H. Wang, X. Xiao, B. Wen, Q. Li, and Y.-C. Hu, "Make web3.0 connected," *IEEE transactions on dependable and secure computing*, vol. 19, no. 5, pp. 2965–2981, 2021.
- [2] J. Bhattacharya, "What is web 3.0? the future of the internet," Single Grain, 2022, accessed: Dec. 2022. [Online]. Available: <https://www.singlegrain.com/Web3.0/web-3-0/>
- [3] X. Zhang and Y. Wang, "Retracted article: Research on intelligent medical big data system based on hadoop and blockchain," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 7, 2021.
- [4] H. Li, D. Han, and M. Tang, "A privacy-preserving storage scheme for logistics data with assistance of blockchain," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4704–4720, 2021.
- [5] J. Xu, K. Xue, S. Li, H. Tian, J. Hong, P. Hong, and N. Yu, "Healthchain: A blockchain-based privacy preserving scheme for large-scale health data," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8770–8781, 2019.
- [6] S. K. Shawon, H. Ahammad, S. Z. Shetu, M. Rahman, and S. A. Hossain, "Diucerts dapp: a blockchain-based solution for verification of educational certificates," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 1–10.
- [7] Z. Sun, D. Han, D. Li, X. Wang, C.-C. Chang, and Z. Wu, "A blockchain-based secure storage scheme for medical information," *EURASIP Journal on Wireless Communications and Networking*, vol. 2022, no. 1, p. 40, 2022.
- [8] E. Chondrogiannis, V. Andronikou, E. Karanastasis, A. Litke, and T. Varvarigou, "Using blockchain and semantic web technologies for the implementation of smart contracts between individuals and health insurance organizations," *Blockchain: Research and Applications*, vol. 3, no. 2, p. 100049, 2022.
- [9] F. Yang and X. Yuan, "Toward timed-release encryption in web3 an efficient dual-purpose proof-of-work consensus," *arXiv preprint arXiv:2205.09020*, 2022.
- [10] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in *2016 2nd international conference on open and big data (OBD)*. IEEE, 2016, pp. 25–30.
- [11] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "Medblock: Efficient and secure medical data sharing via blockchain," *Journal of medical systems*, vol. 42, pp. 1–11, 2018.
- [12] A. F. da Conceição, F. S. C. da Silva, V. Rocha, A. Locoro, and J. M. Barguil, "Eletronic health records using blockchain technology," *arXiv preprint arXiv:1804.10078*, 2018.
- [13] G. Yang and C. Li, "A design of blockchain-based architecture for the security of electronic health record (ehr) systems," in *2018 IEEE International conference on cloud computing technology and science (CloudCom)*. IEEE, 2018, pp. 261–265.
- [14] Q. Xia, E. B. Sifah, A. Smahi, S. Amofa, and X. Zhang, "Bbds: Blockchain-based data sharing for electronic medical records in cloud environments," *Information*, vol. 8, no. 2, p. 44, 2017.
- [15] A. Zhang and X. Lin, "Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain," *Journal of medical systems*, vol. 42, no. 8, p. 140, 2018.

- [16] S. Zhang, A. Kim, D. Liu, S. C. Nuckchady, L. Huang, A. Masurkar, J. Zhang, L. Tseng, P. Karnati, L. Martinez *et al.*, “Genie: a secure, transparent sharing and services platform for genetic and health data,” *arXiv preprint arXiv:1811.01431*, 2018.

A Constraint Programming Approach for Urban Drone Trajectory Optimization

Zahraa Asfour, Sonia Cafieri, Andrija Vidosavljevic
ENAC, Université de Toulouse

Email: {zahraa.asfour, sonia.cafieri, andrija.vidosavljevic}@enac.fr

Abstract—We address the optimization of drone trajectories in urban environments. We introduce a Constraint Programming formulation for a version of the problem where the vehicles are at the same flight level and their path to fly by is known. We show that, compared to an approach based on a Mixed-integer Linear Optimization model introduced previously, the Constraint Programming approach allows better performances to be achieved in the solution of the problem for a specific instance structure.

I. INTRODUCTION

THE growing use of drone delivery in urban environments requires safe and efficient low-altitude trajectory planning and optimization [4]. We address the problem of optimizing drone trajectories by adjusting ground delays and cruise speeds for flights operating along fixed horizontal paths at a common flight level. The main constraints come from a key operational requirement, and consists in maintaining a sufficient separation between pairs of drones at each time along their trajectories. Although the general problem of optimizing drone trajectories involves deciding horizontal paths and flight-levels too for the considered vehicles, the version addressed in this paper is worth exploring as it may be exploited in decomposition approaches for solving large-scale drone trajectory optimization problems.

In [2], a mixed-integer linear optimization model is introduced for drone trajectory optimization in urban environments, where the selection of horizontal paths, flight-levels, and ground delays are used as optimization levers. The Urban Drone Trajectory Model (UDTM) introduced in [1] extends this by also incorporating cruise speed decisions. In [8], optimization is based on adjusting cruise speeds and ground delays, with additional considerations for flight priorities. This is extended in [9], where scheduled take-off uncertainty is addressed too. In that work, the problem is solved using the adversarial Benders decomposition. The subproblem, formulated as a constraint programming model, which retains the same optimization levers as in [8], is solved by three different heuristics.

In this paper, we introduce a Constraint Programming (CP) formulation for the considered problem, as an alternative to the mixed-integer linear optimization model (MILP) from our earlier work [1]. Two optimization criteria are considered. The first minimizes the total deviation, defined as the sum of differences between actual and scheduled landing times across all flights. It reflects a system-level efficiency objective. The

second minimizes the maximum individual deviation, aiming to limit the worst-case deviation. This criterion is relevant when coordination between multiple agents is limited [3]. We show that the CP formulation performs better than the MILP formulation on a specific instance structure, particularly when minimizing the maximum deviation.

The remainder of the paper is organized as follows. Section II presents the problem, recalls the MILP model from earlier work, and introduces the CP model. Section III presents and discusses numerical results. Section IV concludes the paper and discusses possible directions for future research.

II. CP AND MILP FORMULATIONS

This section defines the problem under consideration, recalls the MILP model previously introduced in [1], and presents the CP model proposed in this study.

A. Problem statement

We consider a set of flight intentions, each defined by a departure point, an arrival point, a scheduled take-off time, a horizontal path, and an authorized cruise speed range, bounded between a minimum and a maximum value for each segment of the path. All flights operate at a single cruising flight-level and follow fixed horizontal paths in an urban environment modeled as a directed graph, as proposed in [2], where arcs represent street segments and nodes represent their intersections. Drones are assumed to be of the same type.

This study focuses on optimizing drone trajectories while avoiding Potential Loss of Separation (PLOS) during the cruise phase. A PLOS is defined as a situation in which the time difference between the arrivals of two drones at the same node is less than the minimum separation time. In this work, the minimum separation time is considered node-specific. It is derived from the minimum required separation distance and the minimal authorized cruise speed on the arcs leading to that node.

The goal is to determine, for each drone, a ground delay before takeoff and a cruising speed for each arc of its path, such that all separation between pairs of drones during the cruise phase are satisfied.

B. Model parameters

The main parameters used in both the MILP and CP formulations are summarized below.

Flight parameters

$\mathcal{F}$	set of all flights.
T_f	scheduled take-off time of flight $f \in \mathcal{F}$.
$\delta^{\max} \in \mathbb{R}^+$	maximum ground delay before take-off.

Horizontal path parameters

k_f	horizontal path assigned to flight $f \in \mathcal{F}$.
$K = \bigcup_{f \in \mathcal{F}} \{k_f\}$	set of all horizontal paths.
f_k	flight of path $k \in K$.
$\mathcal{N}_k$	set of nodes visited along path $k \in K$.
n_k^d	departure node of path $k \in K$.
n_k^a	arrival node of path $k \in K$.
$p_{k,n}$	predecessor of node $n \in \mathcal{N}_k \setminus \{n_k^d\}$ on path $k \in K$.
$d_{k,n}^{\min}, d_{k,n}^{\max}$	minimum and maximum travel time along arc $(p_{k,n}, n)$, based on speed bounds.
Γ_k	nominal landing time for path $k \in K$.
$N = \bigcup_{k \in K} \mathcal{N}_k$	set of all nodes visited by at least one flight.

For each pair of departure and arrival nodes associated with flight $f \in \mathcal{F}$, the horizontal path k_f is either selected from a predefined set of available paths (see Section III-B1), or computed using the A* algorithm (see Section III-B2). Note that, for each flight, the nominal landing time corresponds to the landing time obtained without ground delay and assuming the maximum allowed speed on all arcs.

Potential loss of separation parameters

$\mathcal{P}$	set of PLoS.
$k_i^1, k_i^2 \in K$	the two paths involved into the PLoS $i \in \mathcal{P}$.
$f_i^1, f_i^2 \in \mathcal{F}$	the two flights involved into the PLoS $i \in \mathcal{P}$.
$n_i \in \mathcal{N}_{f_i^1} \cap \mathcal{N}_{f_i^2}$	PLoS point where the paths k_i^1 and k_i^2 intersect within the PLoS $i \in \mathcal{P}$.
$M_i^{12}, M_i^{21} \geq 0$	big-M constants used to linearize the separation constraints for PLoS $i \in \mathcal{P}$.
$s_n \geq 0$	minimum separation time at node $n \in N$.
$\mathcal{V}_n$	set of all paths involved in a PLoS at node $n \in N$.

The set $\mathcal{P}$ includes all PLoS, identified by detecting intersecting nodes between the horizontal paths of different flights. Further details on how PLoS are identified, and on how the Big-M constants are computed, can be found in [1]. These constants are computable since all flight times are bounded by the earliest and latest possible arrival times at each node.

Note that in some cases, a pair of flight path share multiple consecutive nodes, giving rise to several detected PLoS. These PLoS are referred to as *correlated PLoS*. In such situation, the order of passage must be maintained across all involved nodes to ensure separation.

Correlated PLoS parameters

$\mathcal{O}$	set of all correlated PLoS.
$i_o^* \in \mathcal{P}$	first (reference) PLoS in correlated PLoS $o \in \mathcal{O}$.
$\mathcal{L}_o \subset \mathcal{P}$	set of remaining PLoS for $o \in \mathcal{O}$.
$k_o^1, k_o^2 \in K$	the two paths involved into the correlated PLoS $o \in \mathcal{O}$.

C. MILP model

The MILP formulation introduced in our previous work [1], corresponding to the UDTM-FPFL variant with a single cruising flight level, is recalled below. Two optimization criteria are considered: (i) minimizing the sum of deviations (referred to as the *SumDev* criterion), and (ii) minimizing the maximum individual deviation (denoted as *MaxDev*). Both formulations share variables and constraints and differ in the objective function.

The decision variable of the SumDev-MILP models are the following.

$t_{k,n}, \forall k \in K, \forall n \in \mathcal{N}_k$	continuous variable representing the arrival time of flight f_k at node n .
$u_i \in \{0, 1\}, \forall i \in \mathcal{P}$	binary variable equal to 1 if f_i^1 passes before f_i^2 at node n_i , 0 otherwise.

The SumDev-MILP formulation aims at minimizing the total deviation. For each path $k \in K$, the deviation is defined as the difference between the arrival time at the final node, t_{k,n_k^a} , and its nominal landing time Γ_k . Since Γ_k is a fixed value for each flight, this objective corresponds to minimizing the sum of individual delays. The formulation is defined as follows:

$$\begin{aligned}
\min \quad & \sum_{k \in K} (t_{k,n_k^a} - \Gamma_k) \quad [\text{SumDev-MILP}] \\
\text{s.t.} \quad & t_{k,n_k^d} \geq T_{f_k}, \quad \forall k \in K \quad (1) \\
& t_{k,n_k^d} \leq T_{f_k} + \delta^{\max}, \quad \forall k \in K \quad (2) \\
& t_{k,n} - t_{k,p_{k,n}} \geq d_{k,n}^{\min}, \quad \forall k \in K, \forall n \in \mathcal{N}_k \setminus \{n_k^d\} \quad (3) \\
& t_{k,n} - t_{k,p_{k,n}} \leq d_{k,n}^{\max}, \quad \forall k \in K, \forall n \in \mathcal{N}_k \setminus \{n_k^d\} \quad (4) \\
& t_{k_i^2, n_i} - t_{k_i^1, n_i} \geq s_{n_i} - M_i^{12}(1 - u_i), \quad \forall i \in \mathcal{P} \quad (5) \\
& t_{k_i^1, n_i} - t_{k_i^2, n_i} \geq s_{n_i} - M_i^{21}u_i, \quad \forall i \in \mathcal{P} \quad (6) \\
& u_{i_o^*} = u_i, \quad \forall o \in \mathcal{O}, \forall i \in \mathcal{L}_o \quad (7) \\
& t_{k,n} \in \mathbb{R}^+, \quad \forall k \in K, \forall n \in \mathcal{N}_k \\
& u_i \in \{0, 1\}, \quad \forall i \in \mathcal{P}.
\end{aligned}$$

Constraints (1)–(2) define the feasible time window for the start of the cruise phase, taking into account a ground delay before take-off bounded between 0 and $\delta^{\max}$. Constraints (3)–(4) make sure the travel time along each arc to remain within admissible bounds. For any node n (excluding the departure node) on a path k , the arrival time $t_{k,n}$ depends on the arrival

time at the preceding node $p_{k,n}$ and the time required to traverse the arc $(p_{k,n}, n)$. The minimum and maximum travel times, denoted respectively by $d_{k,n}^{\min}$ and $d_{k,n}^{\max}$, are derived from the maximum and minimum authorized cruise speeds on that arc and arc length. Constraints (5)-(6) ensure that a minimum time separation s_{n_i} is maintained between the two flights involved in each PLoS $i \in \mathcal{P}$. These constraints model a disjunctive condition: either flight f_i^1 passes before flight f_i^2 at node n_i , or the inverse. This disjunction is handled by introducing a binary variable u_i , which determines the order of passage, and big-M constraints. Finally, Constraint (7) ensures that for each correlated PLoS, the order of passage is same across all associated nodes.

The MaxDev-MILP formulation aims at minimizing the maximum individual deviation across all flights. To this end, an additional continuous variable $\Delta = \max_{k \in K} (t_{k,n_k^a} - \Gamma_k)$ is introduced representing the maximum deviation, and it is modeled with Constraint (8). The formulation is defined as follows:

$$\begin{aligned} \min \quad & \Delta \quad [\text{MaxDev-MILP}] \\ \text{s.t.} \quad & \Delta \geq t_{k,n_k^a} - \Gamma_k, \quad \forall k \in K \\ & \text{Constraints (1) to (7)} \\ & t_{k,n} \in \mathbb{R}^+, \quad \forall k \in K, \forall n \in \mathcal{N}_k \\ & u_i \in \{0, 1\}, \quad \forall i \in \mathcal{P} \\ & \Delta \in \mathbb{R}^+. \end{aligned} \quad (8)$$

D. CP model

An alternative formulation of the SumDev-MILP and MaxDev-MILP is proposed using CP formalism. Unlike the MILP formulation, which combines continuous and binary variables with linear constraints, the CP model is based on interval and integer variables and supports global constraints and logical operators.

a) Decision variables:

Interval variables. One fixed-duration interval variable, denoted $task_{k,n}$, defined for each path $k \in K$ and each visited node $n \in \mathcal{N}_k$, models the occupancy of node n by flight f_k for a duration s_n , corresponding to the minimum separation time required at that location. The start of this interval, $\text{StartOf}(task_{k,n})$, represents the arrival time of flight f_k at node n , and is equivalent to the continuous variable $t_{k,n}$ in the MILP formulation.

Integer variable. A single integer variable Δ is introduced to represent the maximum deviation. It is used only in the MaxDev version to express the upper bound over all deviations.

b) Constraints:

Each CP constraint corresponds to a MILP counterpart and

captures the same temporal or logical relationship.

Precedence and speed limit constraints:

$$\text{StartOf}(task_{k,n_k^d}) \geq T_{f_k}, \quad \forall k \in K \quad (9)$$

$$\text{StartOf}(task_{k,n_k^d}) \leq T_{f_k} + \delta^{\max}, \quad \forall k \in K \quad (10)$$

$$\begin{aligned} \text{EndBeforeStart}(task_{k,p_{k,n}}, task_{k,n}, d_{k,n}^{\min} - s_{p_{k,n}}), \\ \forall k \in K, n \in \mathcal{N}_k \setminus \{n_k^d\} \end{aligned} \quad (11)$$

$$\begin{aligned} \text{StartBeforeEnd}(task_{k,n}, task_{k,p_{k,n}}, s_{p_{k,n}} - d_{k,n}^{\max}), \\ \forall k \in K, n \in \mathcal{N}_k \setminus \{n_k^d\}. \end{aligned} \quad (12)$$

Constraints (9) and (10) ensure that the cruise phase starts within an admissible time window. These constraints are equivalent to the MILP Constraints (1) and (2). Constraints (11) and (12) impose time bounds on the travel between two consecutive nodes. They rely on the following global constraints, defined for two interval variables $task1$ and $task2$:

- Constraint (11) ensures that

$$\text{StartOf}(task_{k,n}) \geq \text{EndOf}(task_{k,p_{k,n}}) - s_{p_{k,n}} + d_{k,n}^{\min}$$

- Constraint (12) ensures that

$$\text{StartOf}(task_{k,n}) \leq \text{EndOf}(task_{k,p_{k,n}}) - s_{p_{k,n}} + d_{k,n}^{\max}$$

Since $\text{EndOf}(task_{k,p_{k,n}}) = \text{StartOf}(task_{k,p_{k,n}}) + s_{p_{k,n}}$, Constraints (11) and (12) together imply that the time between the start of two consecutive tasks lies within $[d_{k,n}^{\min}, d_{k,n}^{\max}]$. They are thus equivalent to constraints (3) and (4) in the MILP formulation.

Separation constraints:

$$\begin{aligned} \text{NoOverlap}(\text{SequenceVar}([task_{k,n}], \forall k \in \mathcal{V}_n)), \\ \forall n \in N. \end{aligned} \quad (13)$$

Constraint (13) uses the global constraint NoOverlap to enforce time-based separation at each PLoS point. It is applied to the sequence variable $\text{SequenceVar}([task_{k,n}]_{k \in \mathcal{V}_n})$, constructed from the set of interval variables $\{task_{k,n}\}_{k \in \mathcal{V}_n}$, where each interval models the presence of a flight at node n for a fixed duration s_n . The SequenceVar structure allows the solver to reason globally about the ordering of intervals on a shared unary resource. Combined with NoOverlap, it ensures that no two intervals in the set overlap in time, thereby preventing simultaneous occupancy of the node. This constraint is equivalent to the disjunctive separation Constraints (5)–(6) in the MILP formulation, but provides a more compact representation.

Order preservation constraint in correlated PLoS:

$$\begin{aligned} \text{SameSequence}(\text{SequenceVar}([task_{k_o^1, n_i^*}, task_{k_o^2, n_i^*}]), \\ \text{SequenceVar}([task_{k_o^1, n_i}, task_{k_o^2, n_i}])) , \\ \forall o \in \mathcal{O}, \forall i \in \mathcal{L}_o. \end{aligned} \quad (14)$$

Constraint (14) ensures that the relative order in which two flights cross a sequence of PLoS points remains consistent

when these nodes are part of the same correlated PLoS. This consistency is enforced by the global constraint SameSequence, which compares two ordered lists of interval variables, each representing the presence of the same pair of flights (paths) at different nodes. The first list defines the reference order at node n_{i_o} , while the second list corresponds to any other node $n_i, \forall i \in \mathcal{L}_o$ involved in the same correlated PLoS set. The constraint guarantees that if one flight precedes the other at the reference node, the same order is maintained at all other nodes. This constraint plays the same role as MILP Constraint (7).

Maximal deviation constraints:

$$\Delta \geq \text{StartOf}(\text{task}_{k,n_k^a}) - \Gamma_k, \forall k \in K. \quad (15)$$

Constraint (15) defines an upper bound Δ on the individual deviations from nominal values. It plays the same role as MILP constraint (8) in the *MaxDev* objective.

c) Model formulations:

The SumDev-CP formulation which minimizes the total deviation, is given by:

$$\begin{aligned} \min \quad & \sum_{k \in K} (\text{StartOf}(\text{task}_{k,n_k^a}) - \Gamma_k) \quad [\text{SumDev-CP}] \\ \text{s.t.} \quad & \text{Constraints (9) to (14).} \end{aligned}$$

The MaxDev-CP formulation which minimizes the maximum individual deviation, is given by:

$$\begin{aligned} \min \quad & \Delta \quad [\text{MaxDev-CP}] \\ \text{s.t.} \quad & \text{Constraints (9) to (15).} \end{aligned}$$

III. EXPERIMENTAL RESULTS

This section presents the numerical results obtained for the MILP and CP formulations introduced in Sections II-C and II-D.

A. Experimental setup

All experiments were conducted using the Gurobi [5] solver for the MILP model and the IBM CP Optimizer [6] (via Docplex) for the CP model. For MILP formulations, Gurobi applies a Branch-and-Bound algorithm combined with cutting planes and presolve routines to explore the search space. For the CP model, the IBM CP Optimizer employs a depth-first search strategy enhanced by constraint propagation, domain reduction, and dynamic variable ordering. Default solver parameters were used in all cases. A single thread was enforced with a time limit of one hour. All runs were executed on a high-performance computing system equipped with 80 cores running at 2.10 GHz and 1 TB of memory.

Each experiment uses a set of generated flight intentions, as detailed in Section III-B. Each flight intention includes a departure and arrival node, a fixed horizontal path, a scheduled take-off time, and a cruise speed range between 4 m/s and 10 m/s.

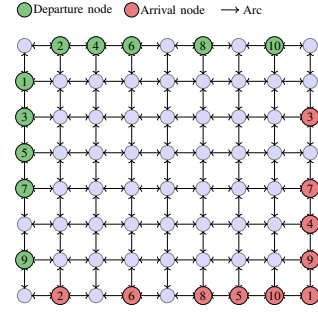


Fig. 1. Grid graph showing fixed departure (green) and arrival (red) node pairs.

B. Test instances

Two types of test instances are considered to evaluate the models. The first type is based on synthetic grids specifically designed to induce a high density of interconnected PLoS. By interconnected PLoS, we refer to situations where resolving one PLoS directly activate another. The second type consists of Vienna instances built on the urban network of the city of Vienna.

1) Instances with high density of interconnected PLoS:

These instances are built on a synthetic grid graph, designed to define the structure of flight interactions.

The graph consists of 72 nodes and 254 arcs, with a fixed arc length per instance. Five arc lengths are considered: 60 meters, 250 meters, 500 meters, 1000 meters, and 2000 meters. To control traffic density, 10 representative flights have been designed. Their departure and arrival nodes are illustrated in Fig.1, and their scheduled take-off times (in seconds) are provided in TABLE I. These 10 representative flights are replicated to generate instances with 20, 30, 40, 50, and 60 flight intentions. Each additional group of 10 flights differs by a 4-second shift in the scheduled take-off time. A maximum ground delay $\delta^{\max}$ of 60 seconds (1 minute) is allowed for all flights.

TABLE I
SCHEDULED TAKE-OFF TIMES (IN SECONDS) FOR THE FIRST 10 FLIGHTS, RELATIVE TO THAT OF THE FIRST FLIGHT.

Pair	1	2	3	4	5	6	7	8	9	10
Scheduled take-off time (s)	0	6	13	7	8	12	12	16	31	32

For each number of flights and arc length, five instances are generated by randomly selecting, for each flight, a horizontal path from a predefined set of available paths. This results in a total of 25 Grid-based instances for each number of flights. The instance generation process is computationally inexpensive, with an average time of less than one second per instance.

2) *Vienna instances:* These instances are based on the urban network of the city of Vienna. The graph contains 4,441 nodes and 7,287 arcs, and spans a metropolitan area approximately 12 kilometers in diameter (see Fig. 2). Departure and

arrival nodes are selected randomly, with the aim of producing flight paths that are spatially dispersed across the network.



Fig. 2. Vienna city graph.

We consider test sets with 50, 100, 150, 200, 250, and 300 flights. For each flight, the shortest horizontal path between a selected departure/arrival nodes is computed using the A* algorithm. Fifty representative instances are generated per number of flights. Each instance is produced in less than two seconds on average.

C. Results on instances with high density of interconnected PLoS

The results obtained on the considered Grid-based instances highlight clear differences in performance between models as the number of flights increases.

For the MaxDev objective, the CP formulation maintains consistent performance across all traffic levels and arc lengths, as shown in TABLE II, which presents the average solving time (in seconds) for each combination of arc length and number of flights. All instances are solved except for a single case with 60 flights. The only unresolved instance corresponds to the most constrained configuration, involving 60 flights and an arc length of 60 meters. In contrast, the number of instances solved to the optimality by the MaxDev-MILP decreases with increasing number of flights. As shown in TABLE III, which reports the number of optimally solved instances for each combination of number of flights and arc length, the MILP model solves only 21 out of 25 instances at 40 flights, 14 at 50 flights, and 4 at 60 flights. The most significant computational difficulties are observed for short arc lengths, particularly 60 meters. In such cases, the limited traversal time induces narrow feasible domains for the temporal variables. This considerably restricts the range of possible time shifts available to satisfy separation constraints, especially since a minimum separation distance of 32 meters—adopted from the work of [7]—is imposed between drones. In contrast, the CP model handles these tightly-constrained instances more efficiently. This behavior is clearly reflected in the solving times reported in TABLE II. For instance, at 50 flights with

60-meter arcs, MaxDev-CP solves all five instances in under 360 seconds on average, while MaxDev-MILP exceeds 3400 seconds.

TABLE II
AVERAGE SOLVING TIME (IN SECONDS) FOR CP AND MILP MODELS ON GRID-BASED INSTANCES, FOR DIFFERENT NUMBERS OF FLIGHTS AND ARC LENGTHS, WITH A ONE-HOUR TIME LIMIT.

20 flights					
Arc length (meters)	60	250	500	1000	2000
MaxDev-MILP	0.74	0.06	0.04	0.04	0.05
MaxDev-CP	2.81	0.09	0.06	0.06	0.07
SumDev-MILP	3600.0	0.34	0.35	0.35	0.38
SumDev-CP	3600.0	13.52	13.33	13.26	13.67
30 flights					
Arc length (meters)	60	250	500	1000	2000
MaxDev-MILP	203.74	0.88	0.84	0.83	0.93
MaxDev-CP	7.18	1.52	1.50	1.50	1.60
SumDev-MILP	3600.0	2966.43	3150.54	3020.26	2928.27
SumDev-CP	3600.0	3600.0	3600.0	3600.0	3600.0
40 flights					
Arc length (meters)	60	250	500	1000	2000
MaxDev-MILP	3596.55	5.03	4.94	4.95	5.14
MaxDev-CP	20.52	2.94	2.96	2.94	2.94
SumDev-MILP	3600.0	3600.0	3600.0	3600.0	3600.0
SumDev-CP	3600.0	3600.0	3600.0	3600.0	3600.0
50 flights					
Arc length (meters)	60	250	500	1000	2000
MaxDev-MILP	3439.21	1484.84	1509.12	1527.03	1533.80
MaxDev-CP	356.58	18.11	17.60	17.32	26.75
SumDev-MILP	3600.0	3600.0	3600.0	3600.0	3600.0
SumDev-CP	3600.0	3600.0	3600.0	3600.0	3600.0
60 flights					
Arc length (meters)	60	250	500	1000	2000
MaxDev-MILP	3600.0	3002.21	3055.06	3011.09	2996.38
MaxDev-CP	1031.30	212.06	236.13	287.37	299.82
SumDev-MILP	3600.0	3600.0	3600.0	3600.0	3600.0
SumDev-CP	3600.0	3600.0	3600.0	3600.0	3600.0

Fig. 3 provides a detailed comparative analysis of the objective values and lower bounds obtained by the MaxDev-MILP and MaxDev-CP across Grid-based instances. Fig. 3a reports the differences in objective values (MaxDev-MILP minus MaxDev-CP) for each arc length and number of flights, while Fig. 3b displays the corresponding differences in lower bounds. Each color encodes a specific arc length category, and for each number of flights (40, 50, and 60), five instances are considered per arc length.

As shown in Fig. 3a, for 40 and 50 flights, the objective values returned by MaxDev-MILP and MaxDev-CP are identical across all instances when the arc length is 250 meters or greater. However, at 60 flights, deviations appear even

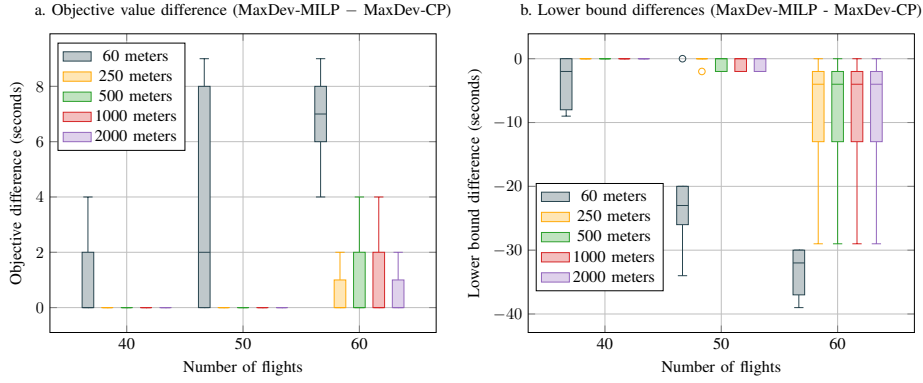


Fig. 3. Analysis of the MaxDev objective and lower bound values obtained by MaxDev-MILP and MaxDev-CP across Grid-based instances with a one-hour time limit.

TABLE III
NUMBER OF GRID-BASED INSTANCES SOLVED TO OPTIMALITY WITHIN THE ONE-HOUR TIME LIMIT, FOR VARYING NUMBERS OF FLIGHTS AND ARC LENGTHS.

Number of flights	20					30					40					50					60				
	Arc length (meters)					60	250	500	1000	2000	60	250	500	1000	2000	60	250	500	1000	2000	60	250	500	1000	2000
MaxDev-MILP	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	1/5	5/5	5/5	5/5	5/5	1/5	4/5	3/5	3/5	3/5	0/5	1/5	1/5	1/5	1/5
MaxDev-CP	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	4/5	5/5	5/5	5/5	5/5
SumDev-MILP	0/5	5/5	5/5	5/5	5/5	0/5	2/5	2/5	2/5	2/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5
SumDev-CP	0/5	5/5	5/5	5/5	5/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5

for longer arcs, with several non-zero differences up to 4-9 seconds observed across all arc categories. The largest discrepancies occur for 60-meter arcs, where all instances display objective gaps, indicating that the MILP formulation fails to match the optimal solution achieved by CP. Fig. 3b further highlights the limitations of MILP in these configurations. For 60-meter arcs under high traffic levels (50 and 60 flights), the differences between the MILP lower bounds and the lower bounds returned by CP reach up to 30 seconds.

For the SumDev objective, both SumDev-MILP and SumDev-CP reach the one-hour time limit in instances with 30 flights, and make solving time a less informative indicator for their comparison (see TABLE II). Fig. 4d shows the mean time at which the last bound update occurs, illustrating a clear distinction in the resolution strategies applied by the two approaches. In SumDev-CP, bounds are typically updated within the first few seconds and remain unchanged thereafter. At 20 flights, for instance, the last update occurs on average after 25 seconds. In contrast, MILP continues to refine bounds until the time limit.

Fig. 4a illustrates the number of instances where SumDev-CP achieves a lower SumDev objective than SumDev-MILP, across all considered number of flights. As shown in Fig. 4a, the number of instances where CP outperforms MILP varies with the number of flights. At 40 flights, CP outperforms MILP in 13 instances versus 12. At 50 flights, the advantage shifts to MILP, with 14 instances where it provides better solutions

compared to 6 for CP. However, at 60 flights, CP outperforms MILP in 19 out of 25 instances. Fig. 4b shows the distribution of objective value differences (SumDev-MILP minus SumDev-CP). For traffic levels ranging from 20 to 50 flights, MILP tends to yield better solutions than CP for arc lengths of 250 meters and above. Most of the differences are small or negative, indicating slightly better objective values obtained by MILP. In contrast, for 60-meter arcs, CP outperforms MILP. This is especially clear at 40 and 50 flights, where the differences in favor of CP reach up to +80 and +56 seconds respectively. The MILP model appears less effective in handling short arcs. At 60 flights, CP yields better results than MILP across all arc lengths in the tested instances. In particular, for 60-meter arcs, the objective difference reaches values as high as +183 seconds in favor of CP, and the median difference remains positive for nearly all arc configurations. Fig. 4c displays the difference in lower bound values obtained by SumDev-MILP and SumDev-CP across all arc lengths and numbers of flights. As observed, SumDev-MILP tends to provide tighter bounds than CP throughout all configurations. This confirms the benefit of the continuous relaxation exploited by MILP.

Overall, the CP model demonstrates a clear advantage under the MaxDev objective in high-density of interconnected PLoS settings, both in terms of solving time and number of instances solved. It systematically outperforms the MILP formulation across all traffic levels tested. For SumDev, CP is more

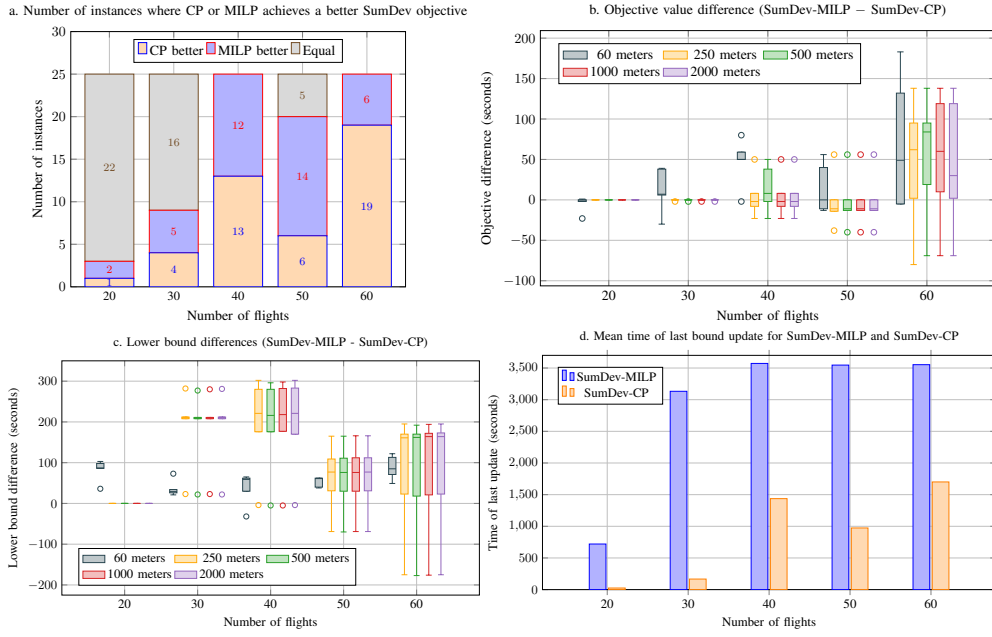


Fig. 4. Analysis of the SumDev objective and lower bound values obtained by SumDev-MILP and SumDev-CP across Grid-based instances with a one-hour time limit.

effective in generating high-quality feasible solutions early in the search, while MILP provides tighter lower bounds and proves optimality more quickly.

D. Results on Vienna instances

The experiments conducted on the Vienna instances exhibit a distinct computational behavior from the Grid-based instances described in Section III-C. These instances are characterized by a lower density of PLoS and a more dispersed network structure (see Fig. 2 and Fig. 5).

TABLE IV
AVERAGE SOLVING TIME (IN SECONDS) FOR CP AND MILP MODELS ON VIENNA INSTANCES, WITH A ONE-HOUR TIME LIMIT.

Number of flights	50	100	150	200	250	300
MaxDev-MILP	≤ 0.01	0.03	0.07	0.13	0.21	0.39
MaxDev-CP	0.02	0.07	0.17	0.34	0.62	1.01
SumDev-MILP	0.01	0.04	0.08	0.16	0.29	0.49
SumDev-CP	72.02	1080.05	2592.05	3600.00	3600.00	3600.00

TABLE IV shows the average solving times for CP and MILP models on Vienna instances, under a one-hour time limit. Under the MaxDev objective, both MaxDev-MILP and MaxDev-CP solve all instances up to 300 flights to the optimality. Solving times, however, differ. MaxDev-MILP remains consistently faster, with average times below 0.4 seconds compared to 1.01 seconds for CP at 300 flights. This behavior contrasts with the Grid-based instances, where CP was faster under the same objective. The observed shift can be attributed

to the reduced number of interconnected PLoS in the Vienna network.

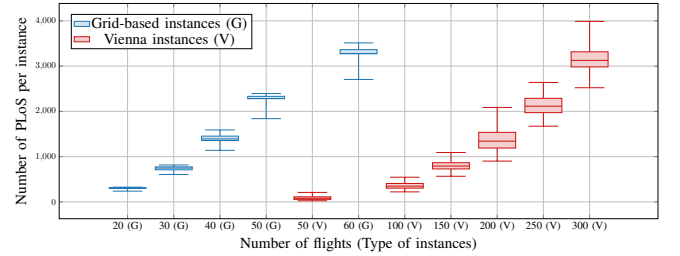


Fig. 5. Boxplot of the number of PLoS per instance as a function of the number of flights

For the SumDev objective, the MILP formulation successfully solves all instances within short computation times, independently of the traffic level. In contrast, the CP model loses its ability to prove optimality beyond 150 flights. This behavior is illustrated in Fig.6a, which displays the number of instances solved to optimality by SumDev-CP as a function of the number of flights. The corresponding differences in lower bounds between the two models are shown in Fig.6b, where the gap increases with problem size, reaching a median of 9.5 seconds at 300 flights.

Despite the inability to certify optimality with a one-hour time limit, both approaches return identical objective values across all traffic levels, as presented in Fig.6c. This observation suggests that the solutions produced by the CP model are optimal. Moreover, the CP formulation updates its lower bound very early during the search, as illustrated in Fig.6d: the average time of the last update remains below 0.02 seconds

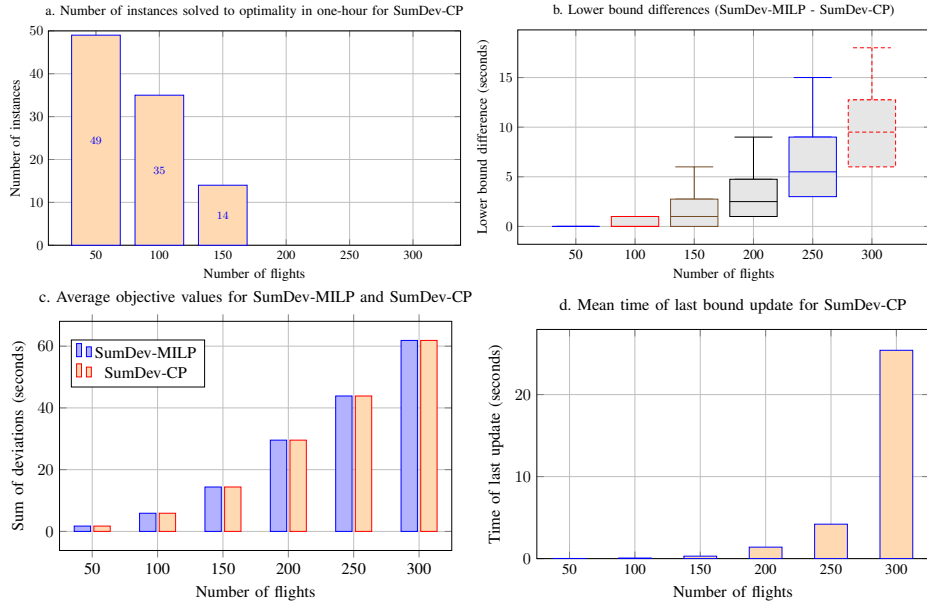


Fig. 6. Analysis of the SumDev objective and lower bound values obtained by SumDev-MILP and SumDev-CP across Vienna instances with a one-hour time limit.

for 50 flights and under 30 seconds for 300 flights.

In summary, the Vienna instances are solved efficiently across all numbers of flights by the MILP formulation. For the MaxDev objective, MILP and CP formulations exhibit comparable performance, with all instances solved to optimality and low solving times. For SumDev, MILP achieves optimality in all cases, while CP finds solutions with the same objective value but exhibits increasing optimality gaps as traffic increases. The early stagnation of CP bound updates limits its ability to certify optimality in large-scale configurations.

IV. CONCLUSION AND FUTURE RESEARCH

This paper compared MILP and CP formulations for optimization of urban drone trajectories under two deviation-based objectives. Numerical experiments on Grid-based and Vienna instances revealed that CP outperforms MILP for MaxDev in instances with high density of interconnected PLoS, while MILP provides tighter lower bounds for SumDev and scales better on Vienna instances. CP proves effective in quickly identifying good solutions, though it struggles to improve bounds. These results confirm the complementary strengths of both approaches, depending on the objective and instance structure. CP can serve as an effective tool for generating initial feasible solutions to warm-start MILP models. Future research directions include the integration of these models into decomposition approaches.

ACKNOWLEDGMENT

The authors thank Carla Juvin for insightful discussions on Constraint Programming formulations.

REFERENCES

- [1] Zahraa Asfour, Sonia Cafieri, and Andrija Vidosavljevic. Optimization of unmanned air vehicles trajectories in urban air mobility. Preprint, May 2025.
- [2] Denis Bereziat, Sonia Cafieri, and Andrija Vidosavljevic. Metropolis II: Centralised and strategical separation management of UAS in urban environment. In *Proceedings of the 12th SESAR Innovation Days, Budapest, Hungary, 5-8 December 2022*, 2022.
- [3] Geoffrey Ding and Hamsa Balakrishnan. Lexicographic min-max fairness in task assignments. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 3408–3414. IEEE, 2023. doi: 10.1109/CDC49753.2023.10384128.
- [4] Malik Doole, Joost Ellerbroek, and Jacco Hoekstra. Estimation of traffic density from drone-based delivery in very low level urban airspace. *Journal of Air Transport Management*, 88:101862, 2020. doi: 10.1016/j.jairtraman.2020.101862.
- [5] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, , 2025. <https://www.gurobi.com/> (accessed: February 20, 2025).
- [6] IBM ILOG CPLEX Optimization Studio CPLEX User's Manual, , 2025. <https://www.ibm.com/fr-fr/products/ilog-cplex-optimization-studio> (accessed: May 14, 2025).
- [7] Andres Morfin Veytia, Calin Andrei Badea, Niki Patrinopoulou, Ioannis Daramouskas, Joost Ellerbroek, Vaios Lappas, Vassilios Kostopoulos, and Jacco Hoekstra. U-Space Utilisation of Airspace under Various Layer Function Assignments and Allocations. *Drones*, 7(7):444, 2023. doi: 10.3390/drones7070444.
- [8] Mercedes Pelegrin, Claudia d'Ambrosio, Rémi Delmas, and Youssef Hamadi. Urban Air Mobility: From complex tactical conflict resolution to network design and fairness insights. *Optimization Methods and Software*, 38(6):1311–1343, 2023. doi: 10.1080/10556788.2023.2241148.
- [9] Tom Portoleau and Claudia D'Ambrosio. A robust two-stage model for the urban air mobility flight scheduling problem. In *Proceedings of the 8th International Symposium on Combinatorial Optimization, ISCO 2024, La Laguna, Tenerife, Spain, May 22–24, 2024*. Springer, 2024. doi: 10.1007/978-3-031-60924-4_26.

Governance, Lean Healthcare, and Digital Transformation in Outpatient Management: An Integrative Review for Data-Driven Health System

Lucas Frota Beckman
0000-0002-2790-0311
Ceuma University and FUMEC
University, PDMA
e-mail:
lfbeckman2013@gmail.com

Cristiana F. De Muijder
0000-0002-0813-0999
FUMEC University
PDMA,
e-mail: crismuijder@hotmail.com;
cristiana.muijder@fumec.br

Olaf Reinhold
0000-0003-1977-1641
Cooperative State University
Saxony, e-mail:
olaf.reinhold@scrc-leipzig.de

Abstract—This study presents an integrative literature review exploring the intersection between governance, Lean Healthcare, and digital transformation in the management of specialized outpatient clinics. The increasing complexity of healthcare demands strategies that improve operational efficiency, ensure regulatory compliance, and support sustainable service delivery. The review analyzes data-driven models, digital technologies, and lean methodologies applied to optimize resource allocation, reduce waiting times, and improve patient experience. Findings highlight the relevance of predictive analytics, artificial intelligence, and interoperable health information systems as enablers of smarter healthcare services. Key challenges include system fragmentation, resistance to organizational change, and technological infrastructure gaps. The study proposes a framework that combines organizational governance, continuous improvement, and digital integration to enhance performance in outpatient settings. It contributes to the advancement of interdisciplinary research and supports evidence-based decision-making in health service management.

Index Terms—Outpatient Management, Lean Healthcare, Digital Transformation, Governance, Data-Driven Health Systems.

I. INTRODUCTION

THE increasing complexity of healthcare systems demands the adoption of management models that foster operational efficiency, care quality, and sustainability. Specialized outpatient clinics play a crucial role within healthcare networks, serving as an intermediary between primary care and hospital services, particularly in the follow-up of patients with chronic and complex conditions. However, these services often face structural challenges that compromise their effectiveness, including the absence of standardized workflows, operational waste, and difficulties in incorporating technologies aimed at optimizing clinical and administrative processes [1].

Governance in healthcare emerges as a central element for ensuring transparency and quality in the management of specialized outpatient services. According to the Brazilian Institute of Corporate Governance [2], effective governance relies on clearly defined roles and responsibilities, the strengthening of organizational culture, and the implementation of control

mechanisms. Nonetheless, most of the existing literature on hospital governance focuses on large institutions, with limited studies addressing its application in specialized outpatient settings. Moreover, there is a gap in understanding how governance can facilitate the adoption of methodologies aimed at operational efficiency and digitalization of healthcare processes [3]. This view is supported by [4], who argue that the standardization and regulation of processes within healthcare institutions follow principles of institutional isomorphism, driven by sectoral norms and demands.

The application of Lean Healthcare, inspired by the Toyota Production System, has proven to be an effective strategy for reducing waste, reorganizing care flows, and optimizing resource utilization. Studies have shown that this methodology can improve the quality of outpatient services by reducing waiting times and increasing demand predictability [5], [6]. However, resistance among healthcare professionals to managerial changes and the need for ongoing training remain significant challenges to its widespread adoption [7]. This resistance may be linked to organizational cultural aspects [8] who emphasize how organizational culture can influence the adoption of new practices and the acceptance of managerial innovations.

At the same time, digital transformation has driven the modernization of healthcare services, enabling the automation of clinical and administrative processes. The use of electronic health records, artificial intelligence, and telemedicine has expanded as tools to improve demand forecasting, reduce operational errors, and increase patient safety. However, the lack of interoperability among health information systems, along with regulatory barriers and technological infrastructure limitations, has hindered the full implementation of these innovations [9]. Digitalization, when aligned with effective governance, can enhance integration among health units, improve data flow, and strengthen the longitudinal monitoring of patients [10].

The relationship between governance, Lean Healthcare, and digitalization in the management of specialized outpatient clinics remains underexplored in the literature. This underscores the need for studies that integrate these approaches and

This work was supported by CNPq.

assess their impact on operational efficiency and quality of care. In this context, the present study aims to conduct an integrative literature review to investigate how organizational governance, lean management, and digital transformation can be combined to improve the efficiency and sustainability of specialized outpatient services. The research seeks to consolidate theoretical knowledge, identify gaps in the literature [10], and offer recommendations to support managers in implementing more effective strategies for organizing these services.

II. CONCEPTS

A. Governance and Compliance in Specialized Outpatient Clinics

Governance in healthcare is a key element for the efficient management of care services, ensuring transparency, control, and a clear definition of responsibilities. In outpatient settings, its structuring aims to improve the quality of care, ensure financial sustainability, and align the interests of different stakeholders [2], [11].

Agency Theory contributes to the understanding of the relationship between managers and healthcare professionals, highlighting how information asymmetry can lead to conflicts and reduce service efficiency. To mitigate risks, the adoption of internal audits, performance indicators, and compliance guidelines is essential. Moreover, institutional isomorphism suggests that healthcare organizations tend to structure themselves similarly due to normative and cultural pressures. In Brazil, regulatory bodies such as the National Health Agency (ANS), the Federal Council of Medicine (CFM), and the National Accreditation Organization (ONA) establish standards that guide the organization of outpatient clinics, promoting legal security and strengthening institutional credibility [1], [4], [6], [12].

Compliance, in turn, functions as a critical mechanism for adherence to regulations and best practices, preventing operational risks and ensuring data privacy, especially in the context of the General Data Protection Law (LGPD – Law No. 13,709/2018). Its implementation involves audits, codes of conduct, training programs, and the use of technology to monitor processes. These measures help reduce fraud, increase financial predictability, and support decision-making [1], [3], [6], [10], [13].

Organizational culture directly influences adherence to standards and the effectiveness of institutional strategies. When structured around well-defined values, it contributes to team engagement and the reinforcement of governance. Thus, governance and compliance in specialized outpatient clinics are essential pillars for care quality, transparency, and institutional security. The adoption of control mechanisms and digital regulation enhances operational efficiency, reduces waste, and supports continuous improvement in care processes [6], [7], [14], [15].

B. Lean Healthcare and Operational Efficiency

The pursuit of operational efficiency has been a central concern in healthcare systems worldwide, especially in light of increasing demand for services and limited resources. In this context, Lean Healthcare, inspired by the Toyota Production System (TPS), has emerged as an effective approach to waste reduction, optimization of care flows, and enhancement of patient care quality. This methodology has been widely applied in hospitals and clinics to minimize operational costs, increase service predictability, and improve the patient experience [3], [14], [16].

The application of Lean Healthcare in specialized outpatient clinics aims to eliminate non-value-adding activities, reorganize clinical and administrative processes, and promote a culture of continuous improvement [17]. According to [18], TPS identifies seven types of waste, which can be adapted to the healthcare sector:

- Waiting – excessive waiting time for consultations, exams, and procedures.
- Excessive motion – unnecessary movement of patients and staff within the facility.
- Overprocessing – redundancy in medical records and documentation.
- Excess inventory – storage of supplies beyond what is necessary, leading to waste.
- Overproduction – exams and procedures performed without clinical justification.
- Defects – administrative and clinical errors, such as rework and failures in electronic health records.
- Underutilized talent – inadequate use of healthcare professionals' skills and competencies.

The application of Lean Healthcare in specialized outpatient clinics involves the use of various tools aimed at waste reduction and optimization of care flows. One of the most widely used techniques is Value Stream Mapping (VSM), which allows for visualization of the entire patient journey within the outpatient setting. This mapping helps identify bottlenecks and supports the implementation of improvements to make care delivery more efficient and responsive [19].

Another relevant approach is Kaizen, which promotes continuous improvement of care processes through incremental adjustments. Small daily changes, when accumulated over time, contribute to increased efficiency and waste elimination, fostering a more productive work environment aligned with patient needs [17].

The 5S methodology enhances workplace organization and standardization by eliminating waste and improving flow, following the principles of Seiri, Seiton, Seiso, Seiketsu, and Shitsuke [18]. The Pull System, in turn, aligns services with real demand rather than fixed staff schedules, reducing queues and wait times [20].

Applied in outpatient clinics, these tools help minimize waste, optimize resources, and enhance the patient experience by promoting a more efficient and responsive care environment.

C. Impacts of Lean Healthcare on Outpatient Efficiency

The application of these tools in specialized outpatient clinics has shown significant results in various studies. The implementation of Lean Healthcare in outpatient services can reduce average waiting times for consultations and diagnostic exams by up to 50%, in addition to increasing the productivity of healthcare professionals. Hospitals and clinics that have adopted this methodology have managed to reduce operational costs by 30% by eliminating waste related to supply management, rework, and unnecessary movements [17], [19].

Beyond waste reduction, Lean Healthcare also contributes to increased predictability in resource allocation. Data-driven models allow forecasting patient demand, optimizing the distribution of medical staff, and preventing both idle periods and work overload. This is particularly relevant for specialized outpatient clinics, where demand may vary significantly depending on seasonal factors and the complexity of the cases treated [1], [6].

Although Lean Healthcare offers significant benefits for the management of specialized outpatient services, its implementation faces several challenges. One of the main obstacles is resistance to change among healthcare professionals. There is often a tendency to maintain traditional processes, either due to unfamiliarity with Lean practices or difficulty adapting to a new organizational culture. A shift in mindset is essential for the successful adoption of the approach and requires efforts to raise awareness and engage healthcare teams [7].

Additionally, the lack of professional training represents a considerable barrier, as the effective application of Lean relies on the correct use of specific tools. These tools must be properly understood to generate meaningful improvements in care processes. Without adequate training, professionals may find it difficult to implement and sustain changes, which reduces the overall effectiveness of the methodology [17].

Another relevant challenge involves the integration of Lean practices with digital technologies, which are essential for optimizing patient flow and enhancing the allocation of available resources. However, the lack of interoperability between different management systems often hinders process automation and real-time data analysis. A poorly structured technological infrastructure can undermine the full adoption of Lean, thereby limiting expected efficiency gains [19], [21], [22].

To ensure the effective incorporation of Lean Healthcare into outpatient management, it is essential to invest in continuous professional development, foster an organizational culture committed to ongoing improvement, and guarantee the integration of appropriate technological solutions. Overcoming these barriers is necessary to achieve the expected outcomes and to consolidate a more efficient and sustainable healthcare delivery model.

D. Queueing Theory-Based Strategies for Care Capacity Management in Healthcare

The application of Queueing Theory in specialized outpatient clinics goes beyond the mathematical analysis of service flows. It requires the implementation of practical

strategies to optimize care capacity and reduce waiting times. Among these strategies, intelligent triage and prioritization stand out, directing patients to the appropriate level of care and avoiding overloading specialized services. The categorization of patient cases can be enhanced by predictive algorithms that analyze medical histories and symptoms to prioritize urgent cases, thereby ensuring a more efficient and equitable patient flow in outpatient care [23], [24].

Another critical aspect is the proper sizing of staff and resources, which can be supported by statistical models. These models help calculate the optimal number of healthcare professionals per shift, aligning service supply with actual demand and avoiding both staff overload and idle time. Flexible staff allocation based on data analysis contributes to better infrastructure utilization and improved service efficiency. In hospitals that implemented such capacity management models, a 30% reduction in waiting times for specialized consultations and an increase in appointment utilization rates were observed [25].

Controlled overbooking has also proven effective in minimizing the impact of patient no-shows. Forecasting absenteeism allows for scheduling a slightly higher number of patients than nominal capacity, reducing unused time slots and increasing the effective service rate. However, this approach must be used cautiously to avoid overcrowding and patient dissatisfaction. Studies indicate that optimizing service flows through these strategies also results in lower absenteeism and improved patient experience [23], [24], [26].

The use of demand forecasting technologies has transformed outpatient management. Artificial intelligence and machine learning tools analyze care patterns and anticipate demand peaks, assisting in resource planning. Furthermore, computerized systems enable real-time monitoring of wait times and room occupancy rates, enhancing predictability and operational efficiency [27], [28]. In Brazil, the Hospital Israelita Albert Einstein implemented demand forecasting and appointment optimization strategies, reducing idle times in medical schedules by 25%, thus demonstrating the effectiveness of digitalization in improving resource allocation and operational performance [29].

The integration of telemedicine has also emerged as a key differentiator in optimizing outpatient services. Virtual consultations can reduce the need for in-person visits in low-complexity cases, allowing on-site services to focus on patients requiring physical exams or specialized procedures. Initial triage via teleconsultation supports faster and more effective care, ensuring better demand distribution and operational efficiency [9].

Altogether, the implementation of these strategies in specialized outpatient care demonstrates tangible benefits in reducing waiting times, optimizing care flows, and enhancing patient experience. The combination of efficient triage, appropriate staff sizing, the use of artificial intelligence, and integration with telemedicine supports the development of a more predictable, accessible, and sustainable care model aligned with international best practices.

Despite the progress enabled by Queueing Theory in care capacity management, the implementation of these strategies in specialized outpatient clinics still faces significant challenges [23], [25]. A major barrier is the lack of integration between information systems, which hinders predictive demand analysis and undermines the optimization of service delivery. Limited interoperability across platforms restricts the effective use of data and impairs evidence-based strategic planning [9]. Resistance to adopting new methodologies also remains a critical obstacle. Many healthcare professionals still rely on traditional scheduling models, which limits the application of more flexible and predictive approaches. Adapting to technology and accepting organizational changes requires continuous efforts to train and engage teams, ensuring a successful transition to new care models [7].

Budgetary constraints present another limiting factor, particularly in public health units, where investment in technology, professional training, and process restructuring may be difficult. Without adequate financial support, implementing digital and predictive solutions becomes more challenging, directly impacting service efficiency [1], [6].

In light of these challenges, it is essential that care capacity management be grounded in evidence-based approaches to ensure better resource utilization and reduced waiting times. Queueing Theory provides a solid framework for demand analysis and service flow optimization, enabling strategic adjustments in medical staff allocation and service organization. International experience demonstrates that adopting these methodologies can significantly improve outpatient efficiency, making services more predictable, accessible, and sustainable [23], [26]. However, to effectively implement these strategies, it is crucial to invest in digital transformation, continuous team training, and the adaptation of organizational culture to new health management models, thus promoting a more agile and efficient care delivery system.

E. Digital Transformation and Technology Use in Outpatient Care Management

Digital transformation has driven the modernization of outpatient management, promoting greater operational efficiency, patient safety, and improvement in the overall care experience. Technologies such as Electronic Health Records (EHR), artificial intelligence, telemedicine, automation, and information security play a central role in optimizing care flows and integrating clinical data. However, the full adoption of these tools still faces challenges related to system interoperability, workforce training, and technological infrastructure [9].

The EHR allows for structured storage and sharing of clinical information, reducing medical errors and facilitating communication among healthcare professionals. Nonetheless, the lack of standardization hinders interoperability between public and private systems, limiting its full potential. Artificial intelligence, in turn, has been applied to demand forecasting, clinical decision support, and personalized care, enabling early disease detection and optimizing service capacity. Despite its benefits, implementation requires strict information

security measures and bias mitigation in algorithms to ensure greater reliability in medical decision-making [29].

Telemedicine has significantly expanded access to specialized services, reducing the burden on in-person appointments and enabling the remote monitoring of chronic patients. During the COVID-19 pandemic, it became an essential alternative for ensuring continuity of care and was regulated by the Federal Council of Medicine in Brazil. However, technological infrastructure and connectivity limitations continue to restrict its broader adoption in some regions. In parallel, the automation of clinical and administrative processes has improved service predictability, reduced waiting times, and organized patient flow using chatbots, automated triage, and real-time monitoring. The implementation of such tools requires ongoing investment, but the efficiency gains justify their long-term adoption [1], [6].

With the advancement of digitalization, information security has become essential for protecting patient data. The Brazilian LGPD establishes guidelines for the storage, sharing, and processing of personal health information, requiring healthcare institutions to adopt measures such as encryption, multifactor authentication, and regular audits. In addition to mitigating legal risks, compliance with these standards enhances institutional credibility and ensures the privacy of healthcare service users.

Therefore, digitalization in healthcare represents an irreversible path, bringing direct benefits to operational efficiency and the safety of care delivery. To maximize its impact, it is essential that institutions invest in modernizing digital infrastructure, provide continuous staff training, and ensure regulatory compliance—thus enabling greater predictability and sustainability for outpatient services.

III. METHOD

This study is an integrative literature review, a method that enables a comprehensive analysis of available evidence on a specific topic by synthesizing both scientific and theoretical findings. This approach allows for the identification of gaps in literature, emerging trends, and the construction of a theoretical framework grounded in governance, operational management, and the digitalization of specialized outpatient clinics.

The review was conducted through a systematic search of scientific articles, books, institutional reports, and international guidelines, prioritizing publications from the period between 2012 and 2024 to ensure the timeliness and relevance of the findings. The selection of studies followed a systematic and replicable process, ensuring transparency and methodological rigor. The literature search was conducted using academic databases recognized for their relevance in the fields of healthcare and hospital management: PubMed – for articles on governance, Lean Healthcare, digitalization, and operational efficiency in health; Scopus and Web of Science – for international studies on hospital management, optimization of care flows, and the impact of digitalization on health services; SciELO and Google Scholar – to include literature from the

Brazilian context on hospital governance and methodologies applied to the national healthcare setting.

The search was conducted using controlled descriptors (DeCS and MeSH), combined with Boolean operators to optimize the retrieval of relevant studies:

- “Governance in healthcare” AND “operational efficiency” AND “specialized outpatient clinics”.
- “Lean Healthcare” AND “outpatient management”.
- “Health digitalization” AND “electronic health records” AND “hospital efficiency”.
- “Queueing theory” AND “care capacity management”.

The inclusion criteria comprised studies published between 2012 and 2024, peer-reviewed articles from indexed scientific journals, and publications in Portuguese, English, or Spanish. Eligible studies addressed themes related to healthcare governance, Lean Healthcare, digitalization, and operational efficiency in specialized outpatient clinics. Additionally, reports and guidelines from regulatory bodies that contribute to the outpatient management model were considered. Conversely, the exclusion criteria ruled out duplicate articles across databases, opinion pieces or studies lacking explicit methodology, research exclusively focused on large hospitals without direct relevance to outpatient care and works that did not explicitly address the concepts of governance, operational efficiency, or digitalization in outpatient management.

The search and refinement process began with the initial identification of 1,452 studies. Titles were screened to assess their alignment with the research scope, resulting in 470 studies deemed relevant for the next phase. Abstracts of these 470 studies were then reviewed to evaluate their relevance to the research question and their methodological quality, after which 158 studies were retained for full-text reading. Following a comprehensive evaluation, 16 studies were selected to compose the integrative review, as they were considered the most relevant for building the theoretical framework and supporting the findings. The selected studies were then analyzed and categorized into four major thematic axes that structured the critical analysis of the results: (1) Governance and Compliance in Outpatient Management, which addressed guidelines for transparency, regulation, and organizational efficiency; (2) Lean Healthcare and Operational Efficiency, focusing on the application of lean methodologies to optimize care processes; (3) Care Capacity Management and Queueing Theory, involving strategies to optimize patient flow and the efficient use of resources; and (4) Digitalization and Use of Technology, highlighting the impact of electronic health records, artificial intelligence, and telemedicine in outpatient management.

IV. RESULTS

The selected articles were grouped into four thematic axes: (1) governance and compliance, (2) operational efficiency and Lean Healthcare, (3) capacity management and Queueing Theory, and (4) digitalization and health technologies. The intersection of these themes reveals that combining governance,

efficiency, and innovation improves outpatient clinic management.

Governance ensures transparency and process standardization (Tables 1–3), though it can increase administrative burden. [31] notes that fragmented governance leads to excessive bureaucracy. In Brazil, regionalized networks may mitigate this, but still face barriers [32].

Lean Healthcare is discussed as a strategy for improving efficiency. Authors [21] highlight the benefits of integrating Lean with digital tools. [33] shows that Lean and FRAM improve referral processes. In Brazil, Lean adoption has improved service organization and engagement [15], and increased competitiveness through efficient resource use [34]. A bibliometric study ([22]) indicates its growing relevance.

Tools such as VSM, Kaizen, and 5S have improved outpatient workflows ([35], [36]). Studies also show Lean Six Sigma in anesthesiology reduces procedural variability [13], [37], [38]. Its application in ICUs and ophthalmology improves discharge and patient flow [37], [38].

Accurate demand forecasting is critical. AI models can predict attendance and optimize schedules [14], [39]. MacLeod et al. (2020) [40] stress reducing cancer treatment wait times through staffing and scheduling. Blockchain can secure records and increase data integrity [41].

Finally, hospital-at-home digital services reduce hospitalization needs [16], and Lean use in UAE’s public sector shows that innovation and digitalization can improve hospital operations [42].

V. DISCUSSIONS

This integrative review examined how governance, operational efficiency, capacity management, and digitalization interact to improve quality and sustainability in specialized outpatient clinics. Despite their benefits, challenges such as system interoperability, resistance to change, and lack of professional training remain barriers to full implementation.

Governance plays a central role by promoting transparency and efficiency. It supports resource optimization and organizational culture [11], and helps mitigate risks [31], [43]. According to institutional theory [44], governance standardization reflects institutional isomorphism, though excessive bureaucracy can hinder agility [31].

When combined with efficiency models, governance enhances resource use and reduces waste. Lean Healthcare, based on the Toyota model, has improved workflows and productivity in healthcare [19], [20], [35]. Its integration with digital tools can reduce costs [45] and wait times [17], though cultural resistance remains a challenge [42].

Capacity management is critical for meeting outpatient demand. Queueing Theory aids in workforce planning [25], and AI helps reduce no-shows and increase efficiency [14].

Digitalization supports process automation, data integration, and decision-making through tools like electronic records, telemedicine, and AI [29], [9]. However, lack of interoperability still limits transformation [16].

Organizational resistance and training gaps are also key barriers. Cultural change is essential for innovation to succeed [7], and leadership engagement and continuous training are vital for implementation [15], [21].

Integrating governance, Lean, capacity models, and technology offers potential to transform outpatient clinic management. But success depends on strategic planning, technological investment, and evidence-based approaches—paving the way for more agile and patient-centered care aligned with digital health trends.

VI. CONCLUSION

This study analyzed the integration of governance, lean management, and digitalization in the management of specialized outpatient clinics, highlighting their importance for operational efficiency and the quality of care. The findings from the integrative review confirm that the adoption of good governance practices contributes to greater transparency, process standardization, and institutional control, thereby strengthening the safety and sustainability of services. Hospital governance, when combined with compliance mechanisms and auditing processes, reduces operational risks and ensures regulatory compliance—both of which are essential for the continuity and predictability of healthcare delivery.

Lean Healthcare has proven to be a key tool for optimizing care flows by eliminating waste and reorganizing clinical and administrative processes. The use of methodologies such as Value Stream Mapping (VSM), Kaizen, and 5S has shown positive impacts in reducing waiting times and improving the utilization of available infrastructure. However, challenges such as resistance to change among healthcare professionals and the need for ongoing training remain significant barriers to the full implementation of this approach.

Care capacity management, supported by Queueing Theory and artificial intelligence, has been shown to be essential for the efficient allocation of human and structural resources in specialized outpatient settings. Predictive models allow for

demand forecasting and more precise scheduling, helping to avoid staff overload and reduce patient no-shows. Nevertheless, the lack of interoperability among information systems hinders the implementation of data-driven solutions, limiting the optimization of care delivery flows.

Digitalization in healthcare has emerged as one of the main drivers of transformation in outpatient services, promoting operational efficiency, patient safety, and personalized care. The use of electronic health records, telemedicine, and automated management systems enables more accurate decision-making, reduces errors, and improves communication among professionals. However, challenges related to technological infrastructure, system standardization, and information security still need to be addressed to ensure the full adoption of these solutions.

Although the benefits of adopting an integrated management model are evident, the review identified structural and cultural challenges that must be addressed to ensure successful implementation. Resistance to change, system fragmentation, and the need for investment in technology and workforce training emerge as critical barriers. To overcome these obstacles, healthcare organizations must adopt a progressive approach, investing in team qualification and developing evidence-based strategies.

Thus, this study reinforces the need for an integrated framework for the management of specialized outpatient clinics, combining governance, operational efficiency, and technological innovation. By overcoming the identified challenges, these units will be able to evolve into a more sustainable, accessible, and patient-centered model of care, aligned with merging trends in digital health and international best practices. Furthermore, the study suggests that future research should further explore the empirical application of these approaches, investigating their feasibility across different institutional contexts and their adaptability to the specific realities of diverse health systems.

REFERENCES

- [1] E. A. Melo, G. G. Gomes, J. O. de Carvalho, P. H. B. Pereira, and K. P. de L. Guabiraba, "Access regulation to specialized outpatient care and the primary health care in national policies of sus," *Physis*, vol. 31, no. 1, 2021, doi: 10.1590/s0103-73312021310109.
- [2] IBGC - Instituto Brasileiro de Governança Corporativa, *Código das melhores práticas de governança corporativa*, 6th ed. São Paulo: IBGC, 2023.
- [3] V. R. de Santana and E. dos S. Santana, "ANÁLISE DO SISTEMA DE COMPLIANCE NA MITIGAÇÃO DE RISCOS CORPORATIVOS: ESTUDO DE CASO EM UMA EMPRESA DE ADMINISTRAÇÃO HOSPITALAR," *Revista Ibero-Americana de Humanidades, Ciências e Educação*, vol. 10, no. 5, pp. 5660–5677, May 2024, doi: 10.51891/rease.v10i5.14281.
- [4] P. J. Dimaggio and W. W. Powell, "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields," 1983.
- [5] J. Fassbinder Da Silva, F. De, L. Nunes, and P. M. Nunes, "Utilização De Mapeamento do Fluxo de Valor em uma Clínica Odontológica: Um Estudo de Caso Use of Value Stream Mapping in a Dental Clinic: A Case Study," 2022. [Online]. Available: <http://leansystem.ufsc.br/>
- [6] C. Melo, F. Berssaneti, G. Rampini, and I. Martinez, "Exploring Barriers and Facilitators to Lean Implementation in Healthcare Organizations."
- [7] Edgar H. Schein, *Schein, Edgar H. (1985): Organizational Culture and Leadership. San Francisco: Jossey-Bass Publishers*, 4th ed., vol. 1. The Jossey–Bass Business & Management Series, 2010. [Online]. Available: <https://www.researchgate.net/publication/303188862>
- [8] M. T. L. Fleury and R. M. Fischer, *Cultura e poder nas organizações*, 1st ed. São Paulo: Atlas, 1996.

- [9] Edward H. Shortliffe, *Biomedical Informatics*. in Health Informatics. New York, NY: Springer New York, 2014. doi: 10.1007/0-387-36278-9.
- [10] A. Angerer, J. Stahl, E. Krasniqi, and S. Banning, "The Management Perspective in Digital Health Literature: Systematic Review," *JMIR Mhealth Uhealth*, vol. 10, no. 11, p. e37624, Nov. 2022, doi: 10.2196/37624.
- [11] R. C. Andrade and J. P. Rossetti, *Governança Corporativa: Fundamentos, Desenvolvimento e Tendências*, 5th ed. São Paulo: Atlas, 1985.
- [12] Michael C. Jensen and William H. Meckling, "Teoria da firma: comportamento dos administradores, custos de agência e estrutura de propriedade teoria da firma: comportamento dos administradores, custos de agência e estrutura de propriedade," *Revista de Administração de Empresas - RAE*, no. 48, pp. 87–125, Apr. 2008.
- [13] G. D. S. Zimmermann, L. D. Siqueira, and E. Bohomol, "Lean Six Sigma methodology application in health care settings: an integrative review," 2020, *Associacao Brasileira de Enfermagem*. doi: 10.1590/0034-7167-2019-0861.
- [14] A. Nelson, D. Herron, G. Rees, and P. Nachev, "Predicting scheduled hospital attendance with artificial intelligence," *NPJ Digit Med*, vol. 2, no. 1, Dec. 2019, doi: 10.1038/s41746-019-0103-3.
- [15] T. O. da Silva, L. M. Vieira, T. S. Lemos, F. P. S. Anna, R. S. Sanches, and M. R. Martinez, "Hospital management and nursing management in the light of the lean philosophy in healthcare," *Cogitare Enfermagem*, vol. 24, 2019, doi: 10.5380/ce.v24i0.60003.
- [16] T. M. Isakov *et al.*, "From challenges to opportunities: Digital transformation in hospital-at-home care," *Int J Med Inform*, vol. 192, Dec. 2024, doi: 10.1016/j.ijmedinf.2024.105644.
- [17] H. I. Farag *et al.*, "One Health: Circadian Medicine Benefits Both Non-human Animals and Humans Alike," *J Biol Rhythms*, vol. 39, no. 3, pp. 237–269, Jun. 2024, doi: 10.1177/07487304241228021.
- [18] Taiichi Ohno, *O sistema Toyota de produção*, 1st ed. Bookman, 1997.
- [19] M. Graban, *Lean hospitals: improving quality, patient safety, and employee engagement*, 1st ed. Productivity Press, 2018.
- [20] James P. Womack, Daniel T. Jones, and Daniel Roos, *The Machine That Changed the World: The Story of Lean Production*, 1st ed., vol. 1. Simon and Schuster, 2007.
- [21] F. N. M. Brancalion, L. G. de Souza, S. Berger, and A. F. C. Lima, "Metodologia Lean: contribuições para melhoria dos processos de trabalho em saúde e enfermagem," *Rev Bras Enferm*, vol. 77, no. 2, 2024, doi: 10.1590/0034-7167-2023-0322pt.
- [22] L. C. N. Vieira, M. D. O. Menezes, C. A. Pimentel, and G. K. S. Juventino, "Lean healthcare no Brasil: uma revisão bibliométrica," *Revista de Gestão em Sistemas de Saúde*, vol. 9, no. 3, pp. 381–405, Oct. 2020, doi: 10.5585/rgss.v9i3.16882.
- [23] C. Y. Huang and T. Y. Kuo, "Queueing-theory-based models for software reliability analysis and management," *IEEE Trans Emerg Top Comput*, vol. 5, no. 4, pp. 540–550, Dec. 2017, doi: 10.1109/TETC.2014.2388454.
- [24] J. L. Wiler, E. Bolandifar, R. T. Griffey, R. F. Poirier, and T. Olsen, "An emergency department patient flow model based on queueing theory principles," *Academic Emergency Medicine*, vol. 20, no. 9, pp. 939–946, Sep. 2013, doi: 10.1111/acem.12215.
- [25] L. Green, "Queueing analysis in healthcare," in *International Series in Operations Research and Management Science*, vol. 91, Springer New York LLC, 2006, pp. 282–307. doi: 10.1007/978-0-387-33636-7_10.
- [26] E. Jafarnejad Ghomi, A. M. Rahmani, and N. N. Qader, "Applying queue theory for modeling of cloud computing: A systematic review," *Concurr Comput*, vol. 31, no. 17, Sep. 2019, doi: 10.1002/cpe.5186.
- [27] S. Mohammed Selim, S. Kularatna, H. E. Carter, N. G. Bohorquez, and S. M. McPhail, "Digital health solutions for reducing the impact of non-attendance: A scoping review," *Health Policy Technol*, vol. 12, no. 2, p. 100759, Jun. 2023, doi: 10.1016/j.hlpt.2023.100759.
- [28] A. S. F. Ribeiro, O. Husson, M. Reuvers, W. J. G. Oyen, C. Messiou, and W. T. A. van der Graaf, "Perspectives on access to imaging digital health records in oncology: A mixed methods systematic review," *Health Policy Technol*, vol. 13, no. 5, p. 100915, Nov. 2024, doi: 10.1016/j.hlpt.2024.100915.
- [29] A. Valente and D. Rezende, "Digital health and its impact on hospital management: A systematic review.," *Health Policy Technol*, vol. 10, no. 2, pp. 75–92, 2021.
- [30] J. Solla and A. Chioro, "Atenção ambulatorial especializada," in *Políticas e sistema de saúde no Brasil*, Editora FIOCRUZ, 2012, pp. 547–576. doi: 10.7476/9788575413494.0020.
- [31] P. P. T. Jeurissen, N. Klazinga, and L. Hagenaars, "Complex Governance Does Increase Both the Real and Perceived Registration Burden: The Case of the Netherlands Comment on 'Perceived Burden Due to Registrations for Quality Monitoring and Improvement in Hospitals: A Mixed Methods Study,'" *Int J Health Policy Manag*, vol. 11, no. 4, pp. 533–535, Apr. 2022, doi: 10.34172/ijhpm.2020.264.
- [32] A. M. dos Santos, *Redes regionalizadas de atenção à saúde: desafios à integração assistencial e à coordenação do cuidado*. EDUFBA, 2018. doi: 10.7476/9788523220266.
- [33] M. Safi, R. Clay-Williams, T. Ursin Grau, F. Brandt, and B. Ravnborg Thude, "FRAM and LEAN as

tools for describing and improving the referral process between outpatient clinics in a Danish Hospital: Complementary or conflicting?,” *Saf Sci*, vol. 166, Oct. 2023, doi: 10.1016/j.ssci.2023.106230.

[34] J. C. Prado-Prado, A. J. Fernández-González, M. Mosteiro-Añón, and J. García-Arca, “Increasing competitiveness through the implementation of lean management in healthcare,” *Int J Environ Res Public Health*, vol. 17, no. 14, pp. 1–26, Jul. 2020, doi: 10.3390/ijerph17144981.

[35] L. B. de Barros et al., “Lean Healthcare Tools for Processes Evaluation: An Integrative Review,” *Int J Environ Res Public Health*, vol. 18, no. 14, p. 7389, Jul. 2021, doi: 10.3390/ijerph18147389.

[36] V. Mishra, “Reducing Waiting Times in Diabetes Care: A Value Stream Mapping and OPA-Fuzzy-TOPSIS Analysis,” *Journal of Creating Value*, vol. 10, no. 2, pp. 266–285, Nov. 2024, doi: 10.1177/23949643241285719.

[37] A. W. Kam et al., “Using Lean Six Sigma techniques to improve efficiency in outpatient ophthalmology clinics,” *BMC Health Serv Res*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12913-020-06034-3.

[38] G. D. S. Zimmermann, L. D. Siqueira, and E. Bohomol, “Lean Six Sigma methodology application in health care settings: an integrative review,” 2020, *Associacao Brasileira de Enfermagem*. doi: 10.1590/0034-7167-2019-0861.

[39] F. N. M. Brancalion and A. F. C. Lima, “Process-based Management aimed at improving health care and financial results,” *Revista da Escola de Enfermagem da USP*, vol. 56, 2022, doi: 10.1590/1980-220x-reeusp-2021-0333en.

[40] A. Macleod, F. Campbell, D. Macrae, E. Gray, L. Miller, and M. Beat-tie, “Reducing wait time for administration of systemic anticancer treatment (SACT) in a hospital outpatient facility,” *BMJ Open Qual*, vol. 9, no. 4, Oct. 2020, doi: 10.1136/bmj-oq-2019-000904.

[41] B. Gayathri D and D. Sangeetha, “Enhancing security in digitized healthcare system using blockchain technology,” *Technology and Health Care*, pp. 1–23, Sep. 2024, doi: 10.3233/thc-240921.

[42] A. Abuhejleh, M. Dulaimi, and S. Ellahham, “Using lean management to leverage innovation in healthcare projects: Case study of a public hospital in the UAE,” *BMJ Innov*, vol. 2, no. 1, pp. 22–32, Jan. 2016, doi: 10.1136/bmjinnov-2015-000076.

[43] A. Duran, T. Chanturidze, A. Gheorghe, and A. Moreno, “Assessment of public hospital governance in romania: Lessons from 10 case studies,” *Int J Health Policy Manag*, vol. 8, no. 4, pp. 199–210, 2019, doi: 10.15171/IJHPM.2018.120.

[44] W. W. Powell, “The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields,” 2016. [Online]. Available: <http://www.jstor.org/stable/2095101>

[45] D. Tlapa et al., “Effects of Lean Interventions Supported by Digital Technologies on Healthcare Services: A Systematic Review,” Aug. 01, 2022, MDPI. doi: 10.3390/ijerph19159018.

TABLE 1. BENEFITS, CHALLENGES, AND STRATEGIES OF INTEGRATED MANAGEMENT IN SPECIALIZED OUTPATIENT CLINICS			
THEMATIC INTEGRATION	BENEFITS	CHALLENGES	PROPOSED STRATEGY
Governance + Lean Healthcare	Reduction of waste; optimization of operational efficiency; improved resource allocation.	Resistance to change among healthcare professionals; lack of specific training.	Continuous professional training; promotion of a culture of innovation and change management.
Governance + Digitalization	Greater control over administrative processes; increased transparency; regulatory compliance.	Low interoperability between information systems; need for technological investment.	Investment in integrated platforms; adoption of international interoperability standards.
Lean Healthcare + Care Capacity Management	Reduced waiting times; improved demand predictability; optimization of care delivery flows.	Lack of adherence to Lean practices by care teams; challenges in collecting and using predictive data.	Application of Queueing Theory; use of artificial intelligence for predictive analysis.
Digitalization + Care Capacity Management	Use of AI for demand forecasting; automation of care workflows; improved patient experience.	Risks related to information security; challenges in complying with data protection regulations (LGPD).	Implementation of digital security protocols, encryption, and multi-factor authentication.

TABLE 2. THE IMPACT OF DIGITAL TECHNOLOGIES ON OUTPATIENT MANAGEMENT			
TECHNOLOGY	DESCRIPTION	BENEFITS	IMPLEMENTATION CHALLENGES
Electronic Health Record (EHR)	Digital recording of consultations, exams, and medical prescriptions.	Reduction of errors, improved continuity of care, remote access to information.	Lack of standardization and interoperability between systems.
Artificial Intelligence (AI) and Machine Learning	Predictive demand analysis, clinical decision support, and process automation.	Better resource allocation, reduced waiting lines, increased care efficiency.	High initial cost, need for professional training.
Telemedicine	Remote consultations for patient monitoring and low-complexity care.	Reduced burden on in-person services, expanded accessibility.	Requirement for stable connectivity and specific regulation.
Blockchain for Data Security	Decentralized and immutable recording of medical information.	Greater reliability and protection of patient data.	Regulatory challenges and high computational resource consumption.

TABLE 3. STRATEGIC IMPACT ON PATIENT EXPERIENCE QUALITY			
IMPLEMENTED STRATEGY	DESCRIPTION	EXPECTED EFFECT ON PATIENT EXPERIENCE	KEY BENEFITS
Lean Healthcare and reorganization of care flows	Digital recording of consultations, exams, and medical prescriptions.	Reduced waiting times and optimization of the patient journey.	Greater satisfaction, reduced idle time, and faster processes.
Digitalization of care and automated scheduling	Predictive demand analysis, clinical decision support, and process automation.	Easier access to care and greater predictability.	Prevents delays, reduces no-shows, and improves communication.
Use of AI in triage and demand forecasting	Remote consultations for patient monitoring and low-complexity care.	Faster and more personalized care based on patient profile.	Decrease in unnecessary visits, optimization of care capacity.
Interoperability between systems and accessible clinical history	Decentralized and immutable recording of medical information.	Improved continuity of care and reduction in errors.	More accurate monitoring of the patient's health status.

Domain-as-Particle with PSO Methods for Neural-Network Feature Weighting

Fabio Berberi

University of Siena

Via Roma, 56, 53100 Siena, Italy

Leuphana University of Lüneburg

Universitätsallee 1, 21335 Lüneburg, Germany

ORCID: 0009-0004-8825-8707

Email: f.berberi@student.unisi.it

Paolo Mercorelli

Leuphana University of Lüneburg

Universitätsallee 1, 21335 Lüneburg, Germany

ORCID: 0000-0003-3288-5280

Email: paolo.mercorelli@leuphana.de

Abstract—We present a framework that integrates Particle Swarm Optimization (PSO), machine learning, K-Fold cross-validation, and surrogate modeling to identify optimal weight vectors for feature scaling in neural network training. In our approach, the n -dimensional weight space is partitioned into non-overlapping subdomains, each corresponding to a PSO particle. Particle movement is guided by a characteristic vector, which is determined by the best-performing candidates in each subdomain and by information exchanged with neighboring regions. To reduce evaluation costs, a surrogate model—trained on a uniformly sampled subset of candidates—pre-filters particles before full K-Fold validation. The top candidates then undergo comprehensive validation, updating the characteristic vectors for subsequent iterations. This domain-as-particle PSO framework enables efficient weight discovery, significantly reducing computational overhead while maintaining robust performance. The effectiveness of this approach is demonstrated on real-world datasets.

Index Terms—Particle Swarm Optimization, surrogate modeling, neural networks, K-Fold cross-validation, feature weighting, medical application.

I. INTRODUCTION

IDENTIFICATION methods play a pivotal role in the control and monitoring of dynamic systems. They form the backbone of modern engineering, seamlessly integrating concepts from various disciplines to achieve regulation, stabilization, and optimization of these systems. This multidisciplinary nature makes system identification and control a crucial area of both academic study and practical application, extending its impact across diverse fields such as engineering, computer science, mathematics, physics, biology, economics, and environmental engineering. Notable examples include contributions like [1] and [2], which demonstrate the successful application of identification and control techniques in environmental engineering and ecology. Similarly, [3] provides a practical case of identification methods applied within an economic and commercial context, specifically to prey-predator models.

Particle Swarm Optimization (PSO), first introduced in [4], is an evolutionary computation technique inspired by the social behaviors of bird flocking and fish schooling. Initially designed

for continuous optimization problems within computational intelligence, PSO has evolved into a versatile, interdisciplinary tool applied across numerous domains. In computer science, it is utilized for tasks such as feature selection in machine learning [5] and hyperparameter optimization in deep learning models. The biomedical sector has also adopted PSO for applications like image segmentation [6] and gene expression data analysis, illustrating its adaptability to high-dimensional and noisy data environments. PSO's popularity can be attributed to its simplicity, flexibility, and effectiveness in addressing nonlinear, non-differentiable, and multi-objective problems, making it a powerful tool for both theoretical research and real-world applications. A detailed and recent survey of PSO can be found in [7]. Optimizing feature weights enhances neural network training by scaling inputs to accelerate convergence and improve accuracy. Manual feature engineering or exhaustive hyperparameter search becomes infeasible in high-dimensional spaces. We propose a *domain-as-particle PSO framework*, where the n -dimensional weight space, defined as the *mother domain* $\mathcal{W}$, is partitioned into distinct regions. Each region acts as a particle that independently explores its section to search for optimal solutions.

Contribution and Structure of the Paper

This paper presents a new framework for neural-network feature weighting based on a domain-as-particle Particle Swarm Optimization (PSO) strategy. The method efficiently partitions and explores the weight space, combining surrogate modeling with adaptive domain updates to enable scalable optimization in high-dimensional scenarios. The effectiveness of this approach is demonstrated through experiments on real-world datasets and benchmark comparisons.

The remainder of the paper is organized as follows. Section II presents the proposed method in detail. Section III describes the dataset and initialization procedures. Section IV discusses the experimental results. Section V concludes the paper and outlines future research directions.

II. PROPOSED METHOD

The proposed domain-as-particle Particle Swarm Optimization (PSO) framework for neural-network feature weighting

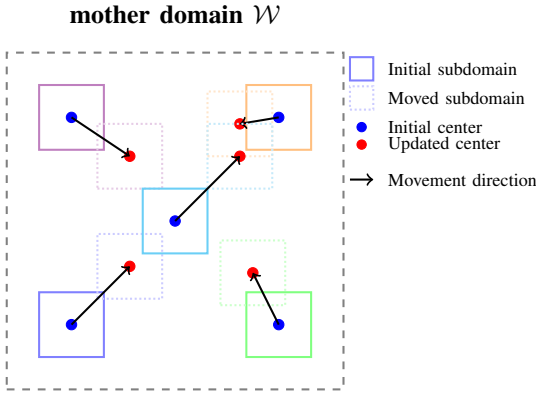


Fig. 1. Conceptual illustration of domain movement: five subdomains, initially uniformly distributed in $\mathcal{W}$, each move a short distance to new positions. The movement is driven by each subdomain's characteristic vector and neighbor influences.

operates through the following main steps:

- **partitioning:** The n -dimensional feature weight space is divided into multiple subdomains, each acting as a separate PSO particle.
- **Particle generation and surrogate modeling:** In each subdomain, a set of candidate weight vectors (particles) is generated. A subset is evaluated using neural network training and K-Fold cross-validation, and the results are used to train a surrogate model (ensemble of Random-Forest and GradientBoosting regressors) that predicts the quality of remaining candidates.
- **Selection and validation of top candidates:** The surrogate model identifies the most promising particles, which are then fully validated using K-Fold cross-validation.
- **Characteristic vector update:** The best candidates in each subdomain are combined into a characteristic vector (e.g., weighted average) to guide the next search step.
- **Domain update (PSO movement):** Each subdomain's center is shifted according to its characteristic vector and, optionally, information from neighboring subdomains.
- **Iteration and stopping criteria:** These steps repeat iteratively, with surrogate retraining and domain adaptation, until convergence or budget limits are met.

This approach allows efficient and scalable search for optimal feature weightings in high-dimensional problems.

A. Weight-Space Partitioning

The weight-space partitioning is the foundational step of the proposed algorithm (see Fig. 2). The global weight space $\mathcal{W}$ is defined as the n -dimensional hypercube $[\text{lb}, \text{ub}]^n$, where n is the number of features and lb, ub are the lower and upper bounds for each weight, respectively. This weight space $\mathcal{W}$ is partitioned into several uniformly distributed subdomains D_j , each with fixed side length L and centered at a grid point $\mathbf{d}_j$. Importantly, these subdomains D_j do not overlap and have deliberate spacing between them.

This spacing allows each subdomain D_j room to move and adapt during optimization, facilitating thorough and indepen-

dent exploration of the search space. As the algorithm iterates, each D_j shifts to investigate unexplored regions, ensuring efficient coverage of the entire main domain.

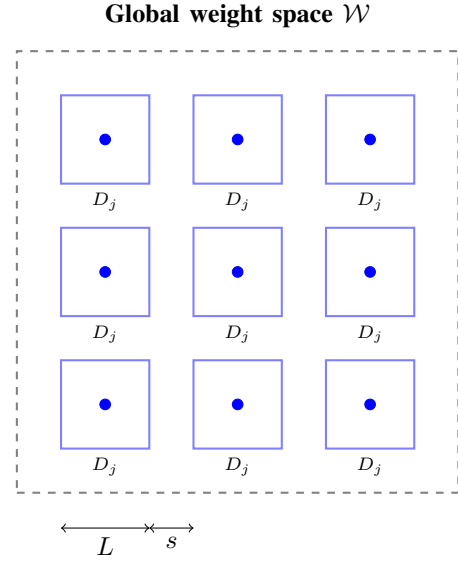


Fig. 2. Weight-space partitioning: $\mathcal{W}$ is partitioned into a 3×3 grid of subdomains D_j , each an n -dimensional cube of side length L , separated by spacing s , and offset by an internal margin M .

B. Particle Generation and Surrogate Modeling

After defining the subdomains (see Fig. 3), the next step is to generate candidate weight vectors inside each subdomain. For every D_j , both standard particles and those selected for surrogate modeling are **uniformly sampled** within the subdomain boundaries. This allows the algorithm to explore different weightings in each region of the search space.

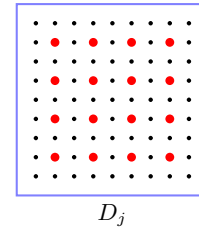


Fig. 3. Illustration of a single subdomain D_j with a uniform interior grid of black candidate particles (excluding the topmost row) and a uniformly spaced subset highlighted in red.

After generating both types of particles—the P full particles and the S surrogate particles—we begin the surrogate model construction as follows. For each surrogate particle, we estimate performance using K-Fold cross-validation: the dataset is split into K equal parts, with each part used once as the validation set and the others as the training set. The final performance metric is obtained by averaging results across all folds. This approach provides a robust estimate of the model's generalization ability.

With reference to the neural network (see Fig. 4): For each surrogate particle, the neural network is trained on the training folds and tested on the test fold, resulting in a composite value for each split. The composite value is defined as the arithmetic mean of three key classification metrics: accuracy, precision, and recall. This metric plays a central role in our algorithm as the primary objective to optimize, representing a holistic measure of the model's predictive power on the dataset.

Specifically, the composite value quantifies the overall quality of the candidate feature weighting by balancing multiple aspects of classification performance:

- **Accuracy** measures the proportion of correctly classified instances among all samples, providing a general assessment of model correctness.
- **Precision** evaluates the proportion of true positive predictions among all positive predictions made by the model, reflecting the model's ability to avoid false positives.
- **Recall** (also known as sensitivity) assesses the proportion of true positive cases correctly identified by the model, indicating its effectiveness at detecting positive instances.

Each of these metrics addresses a distinct aspect of classification performance, and combining them equally in the composite value ensures a balanced evaluation that mitigates biases inherent in any single metric. Maximizing the composite value during optimization directly corresponds to improving the model's ability to generalize and make accurate predictions, which is the ultimate goal of our feature weighting approach.

The hyperparameters and structural characteristics of the neural network used are not fixed but depend on factors such as the dataset size, the number of features, and other intrinsic properties of the data. Consequently, users applying this framework should perform preliminary experiments to identify a network configuration that produces high composite values from the earliest iterations, enhancing optimization efficiency.

In future work, we intend to augment this approach by leveraging Particle Swarm Optimization to automatically tune neural network hyperparameters, thus optimizing both feature weights and model architecture in a unified framework.

This procedure is repeated for all splits. At the end, the average performance is calculated for the model trained with the weight vector defined by the surrogate particle. This procedure is called full validation. This entire process is performed for every surrogate particle, so that for each surrogate particle, a composite value is generated. In other words, for each surrogate particle k in domain D_j , the K-Fold cross-validation returns a single composite value $\sigma_{j,k}$, representing the average performance of the neural network when trained and validated using the weight vector associated with that surrogate particle. At this stage, for each domain, the surrogate model is constructed by leveraging the composite values already computed for the surrogate particles. As illustrated in Fig. 5, the surrogate consists of a RandomForestRegressor and a GradientBoostingRegressor, both trained on (particle, composite value) pairs from K-Fold cross-validation.

```
{
  "MODEL_ARCHITECTURE": {
    "layer1_units": 124,
    "layer1_activation": "relu",
    "dropout1": 0.3,
    "layer2_units": 64,
    "layer2_activation": "relu",
    "dropout2": 0.3,
    "output_units": 1,
    "output_activation": "sigmoid"
  },
  "TRAINING": {
    "EPOCHS": 2,
    "batch_size": 16
  }
}
```

Fig. 4. Neural network configuration: two-layer feed-forward network with ReLU activation and dropout in the hidden layers, and a sigmoid activation in the output layer. The training is performed for 2 epochs with a batch size of 16. Both the architecture and all training hyperparameters are fully customizable and can be modified as needed for different applications.

The RandomForestRegressor operates by training a large number of independent decision trees, each built on different random subsets of the data and features. The final prediction is made by averaging the outputs of all the trees, which helps reduce overfitting and increases prediction stability.

The GradientBoostingRegressor, on the other hand, builds its decision trees sequentially. Each new tree is trained to correct the errors (residuals) of the combined predictions of the previous trees, gradually improving the overall predictive accuracy. This makes Gradient Boosting particularly effective at capturing complex patterns in the data.

For each candidate, the surrogate prediction is obtained by taking the mean of the outputs from the two regressors. By combining these two different regression techniques, the surrogate model provides fast, reliable, and robust estimates of candidate performance. This enables the algorithm to quickly identify and select the most promising candidates for expensive full validation with the neural network, significantly reducing computational costs.

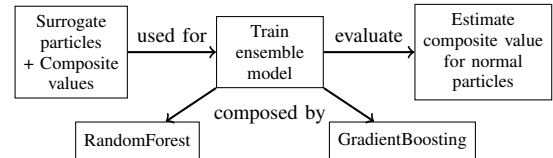


Fig. 5. Schematic of surrogate model creation: composite values from surrogate particles are used to train an ensemble consisting of RandomForest and GradientBoosting models. The predictions from both models are averaged to estimate the composite value for normal candidate particles.

C. Selection and Validation of Top Candidates

After the surrogate model predicts composite values for all candidate particles, a two-step selection process is applied,

as illustrated in Fig. 6. First, the top K particles with the highest surrogate-predicted scores are selected from the initial candidate set. Next, these selected candidates undergo full validation using K-Fold cross-validation with the actual neural network, in order to accurately assess their true performance.

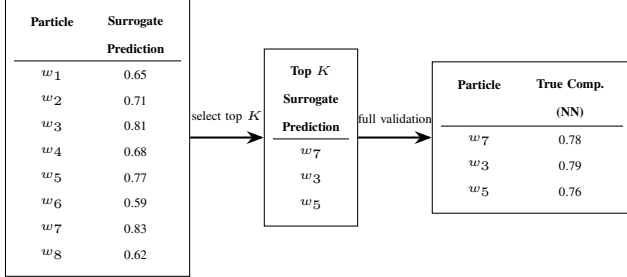


Fig. 6. Workflow of candidate selection: All candidate particles with surrogate predictions (left); selection of top K candidates (center); true composite values after full validation with the neural network (right).

D. Characteristic Vector Update

Once the top- K candidates $\{\mathbf{w}_{j,k}\}_{k=1}^K$ in each subdomain D_j have been fully validated, we aggregate their information into a single “characteristic” vector $\mathbf{c}_j$ that will guide the next movement of D_j . Let $\sigma_{j,k}$ denote the composite value assigned to candidate $\mathbf{w}_{j,k}$. We define:

$$\mathbf{c}_j = \frac{\sum_{k=1}^K \sigma_{j,k} \mathbf{w}_{j,k}}{\sum_{k=1}^K \sigma_{j,k}} \quad (1)$$

In other words, $\mathbf{c}_j$ is the weighted average of the K best weight vectors in D_j , with weights proportional to their observed performance. This choice has two desirable properties:

- *Exploit high-quality solutions*: vectors with larger $\sigma_{j,k}$ contribute more strongly, pulling $\mathbf{c}_j$ toward regions of higher performance.
- *Smooth update*: by averaging, we avoid abrupt jumps in domain center due to outlier candidates.

The updated characteristic vector $\mathbf{c}_j$ then serves as the *new position* (center) of subdomain D_j in the subsequent PSO iteration.

E. Domain Update (PSO Movement)

After defining the characteristic vector $\mathbf{c}_j$, each subdomain D_j is shifted by combining four components—*inertia*, *personal best*, *global best*, and, most critically, the *neighbor influence* that pulls D_j toward positions discovered by its adjacent domains. It is precisely this social component that enables our algorithm to operate in a PSO-like manner.

First, we introduce the three acceleration components:

$$\begin{aligned} \Delta_j^p &= \phi_p r_p (\mathbf{p}_j - \mathbf{c}_j^{(t-1)}), \\ \Delta_j^n &= \phi_n r_n (\mathbf{n}_j - \mathbf{c}_j^{(t-1)}), \\ \Delta_j^g &= \phi_g r_g (\mathbf{g} - \mathbf{c}_j^{(t-1)}). \end{aligned} \quad (2)$$

where:

- $\mathbf{p}_j$ is the personal best position of D_j ,
- $\mathbf{n}_j$ is the best position discovered by its neighbors,
- $\mathbf{g}$ is the global best across all subdomains,
- $r_p, r_n, r_g \sim U(0, 1)^n$ are independent uniform random vectors.

Next, the velocity is updated in PSO fashion:

$$\mathbf{v}_j^{(t)} = \underbrace{\omega \mathbf{v}_j^{(t-1)}}_{\text{inertia}} + \Delta_j^p + \Delta_j^n + \Delta_j^g, \quad (3)$$

and the subdomain center shifts according to

$$\mathbf{c}_j^{(t)} = \mathbf{c}_j^{(t-1)} + \mathbf{v}_j^{(t)}. \quad (4)$$

Fig. 7 illustrates how the characteristic vector (black arrow) and neighbor influences (red arrows) combine to drive each subdomain from its initial (blue solid) to its moved (blue dotted) position.

Interpretation of terms:

- $\omega \mathbf{v}_j^{(t-1)}$ (*inertia*): retains part of the previous velocity, smoothing motion.
- Δ_j^p (*cognitive/personal*): pulls the center toward $\mathbf{p}_j$, exploiting its own best.
- Δ_j^n (*social/neighbor*): pulls toward the neighbor best $\mathbf{n}_j$, enabling coordinated exploration.
- Δ_j^g (*global*): draws all subdomains toward the global best $\mathbf{g}$ for convergence.
- $r_p, r_n, r_g \sim U(0, 1)^n$ (*random vectors*): introduce stochastic variation—larger values favor exploration, smaller values fine-tune exploitation.
- The update $\mathbf{c}_j^{(t)} = \mathbf{c}_j^{(t-1)} + \mathbf{v}_j^{(t)}$ completes the PSO move for D_j .

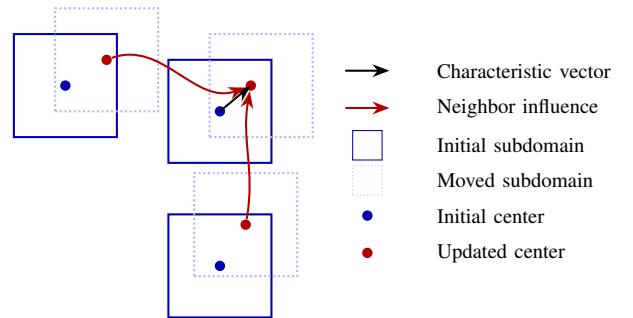


Fig. 7. Characteristic vector with neighbor influence: solid-line subdomains are initial, dotted are moved; centers shift from blue to red dots; black arrow shows characteristic vector; red arrows show neighbor influence.

Having detailed how each domain’s center is updated via its characteristic vector, we now describe how these updates are organized over time and when the algorithm terminates.

F. Iteration and Stopping Criteria

Our framework proceeds in discrete iterations $t = 1, 2, \dots$, mirroring the structure of classical PSO.

Per-iteration workflow:

- 1) *Domain movement*: each D_j is shifted by its updated velocity and characteristic vector.
- 2) *Surrogate set refresh*: the surrogate particles farthest from the new center are discarded and replaced by new samples drawn in the region just explored.
- 3) *Surrogate retraining*: the surrogate model for D_j is re-trained on the updated surrogate set and their composite values.

Stopping criteria:

- *Target composite value*: terminate as soon as any candidate reaches a predefined composite score.
- *Fixed budget*: run for a preset number of iterations or wall-time, then return the best vector seen across all domains and iterations.
- *Budget exhaustion*: stop when the allotted computational resources (e.g. number of network evaluations or CPU-hours) are consumed.

Upon termination, the algorithm returns the highest-scoring weight vector found.

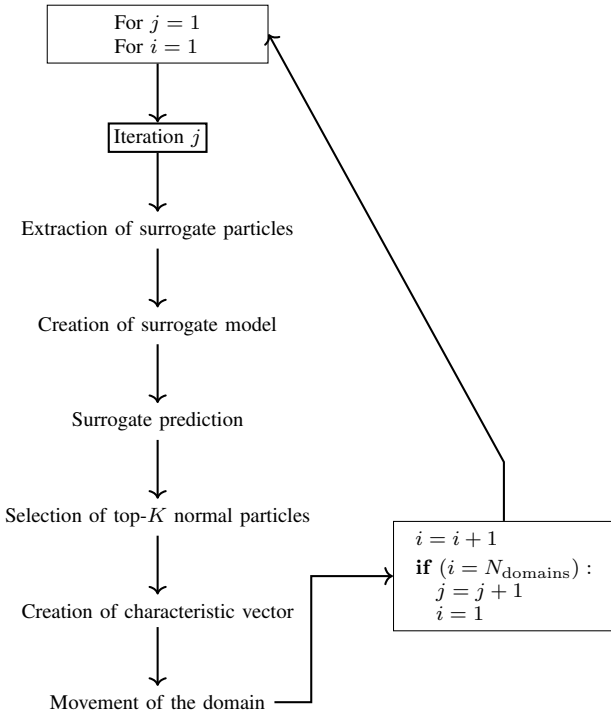


Fig. 8. Illustrative flow of the per-domain, per-iteration update loop: each iteration processes surrogate particle extraction, surrogate model training, prediction, top- K selection, characteristic-vector creation, and domain movement, followed by index updates. Here N_{domains} denotes the total number of subdomains within the parent domain.

III. DATASET AND INITIALIZATION PROCEDURES

The proposed framework was validated on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which includes $n = 30$ cellular features per sample and binary labels indicating malignant (M) or benign (B) tumors. The dataset consists

of $N = 569$ samples; the label column was converted to 1 (malignant) and 0 (benign) for classification.

All features were standardized using z-score normalization. The dataset was split into training and test sets with an 80-20 stratified split to preserve the original class balance. Categorical labels were mapped as follows: $M \rightarrow 1$, $B \rightarrow 0$.

The main phases of data preparation and the per-domain optimization cycle are illustrated in Fig. 8.

For class imbalance, balanced class weights were computed and applied during training.

The PSO-based optimization was initialized as follows:

- Global weight bounds: $lb = 1.0$, $ub = 10.0$
- Number of subdomains: $m = 10$
- Subdomain width: $L = 0.5$
- Particles per domain: $P = 100$
- Surrogate particles per domain: $S = 10$
- Precision: weights rounded to one decimal
- Domain update factor: $\gamma = 0.1$
- Number of global iterations: $T = 5$
- Early stopping threshold: $\epsilon = 0.001$
- Top $K = 20$ particles per domain revalidated

Within each subdomain, candidate weight vectors (particles) were generated using Latin Hypercube Sampling and sorted by the sum of their components. Surrogate models (Random Forest and Gradient Boosting Regressors) were trained on a uniform subset of particles, allowing efficient pre-selection before full neural network validation.

The neural network used was a feed-forward model with two hidden layers (124 and 64 units, ReLU activation, 0.2 dropout), trained using the Adam optimizer and binary cross-entropy loss, following established approaches that apply sensitivity and feature-importance analysis for medical diagnosis using neural networks [8]. Performance was assessed via 5-fold cross-validation, averaging accuracy, precision, and recall into a composite metric.

IV. EXPERIMENTAL RESULTS

To rigorously assess the performance of our domain-as-particle PSO framework, we conducted a direct comparison with the standard (classical) PSO approach on the widely studied Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Both methods were evaluated under challenging, high-dimensional conditions ($n = 30$), which are typical of real-world feature weighting problems.

Figure 9 displays the composite value trajectories for each subdomain during the optimization process using our proposed framework. In this experiment, we initialized 1000 particles per domain (i.e., per subdomain), adopted a batch size of 30 samples per neural network update, and used the neural network architecture described in Figure 4. The plot shows the evolution of composite values for all 10 domains distributed throughout the search space. Notably, subdomain 6 (highlighted in orange) achieved the highest composite value, indicating that this region of the weight space contained weight vectors that assigned the most effective feature scaling to the dataset. In contrast, the other domains, which do not

reach such high composite values, are evidently located in less favorable regions of the 30-dimensional search space, where the initialized weights do not lead to optimal model performance.

For the baseline comparison, Figure 10 presents results obtained using the classical PSO algorithm, without any subdomain partitioning or surrogate modeling. In this setting, we distributed 100,000 particles across the entire weight space to achieve comprehensive coverage of the 30-dimensional space, and used a batch size of 40 samples. Importantly, this approach does not leverage any of the domain-as-particle or surrogate-assisted strategies introduced in our framework. As a consequence, the algorithm must directly evaluate every candidate weight vector with full neural network training, leading to a substantial increase in computational burden and runtime.

From a computational perspective, the efficiency gain provided by our method is substantial. While the classical PSO required approximately hours to complete, our domain-as-particle approach achieved comparable or superior composite values in just 3 hours. This demonstrates not only a significant reduction in runtime, but also a clear advantage in terms of computational cost and scalability. The improvement in both efficiency and solution quality highlights the effectiveness of the proposed surrogate-assisted, multi-domain optimization strategy for neural-network feature weighting. To further benchmark our method, we implemented a simple attention-based classifier. The model consists of a self-attention layer with dimensionality equal to the number of input features (30), followed by two fully connected layers with ReLU activations and a sigmoid output layer for binary classification. The classifier was trained with the Adam optimizer at a learning rate of 1×10^{-3} , a batch size of 32, and for 50 epochs. This attention-based model achieved a composite value of approximately 0.94, which is slightly lower than the composite value obtained by our proposed domain-as-particle PSO framework, demonstrating that our method outperforms this baseline in terms of feature weighting effectiveness.

V. CONCLUSION AND FUTURE WORK

This paper introduced a domain-as-particle Particle Swarm Optimization (PSO) framework for neural network feature weighting, validated on medical data for breast cancer diagnosis. The method leverages surrogate models to efficiently guide the optimization process, achieving fast convergence to high-quality solutions with reduced computational cost. Our framework demonstrated robustness and scalability, making it well-suited for complex, high-dimensional feature weighting problems.

The algorithm has been tested on multiple datasets beyond the one presented here, consistently producing promising results that validate its general applicability and effectiveness. We plan to extend this line of research by applying the method

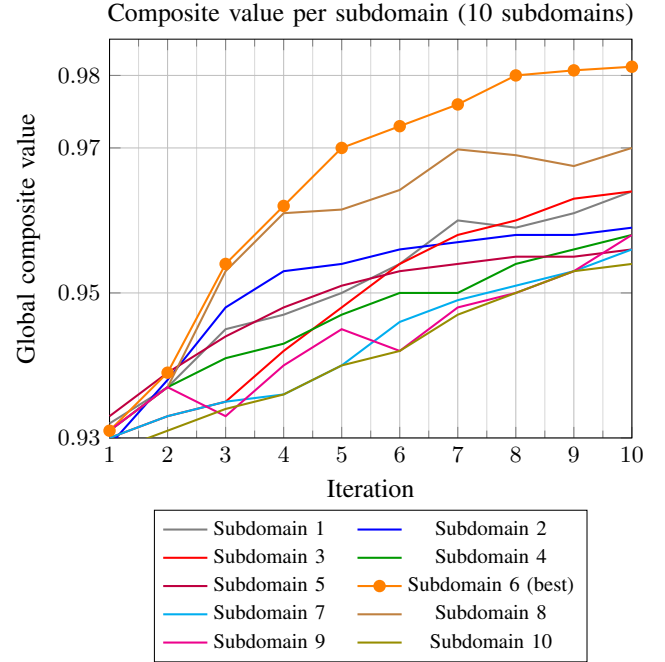


Fig. 9. Composite value per subdomain obtained using the domain-as-particle PSO method (multi-domain, surrogate-assisted).

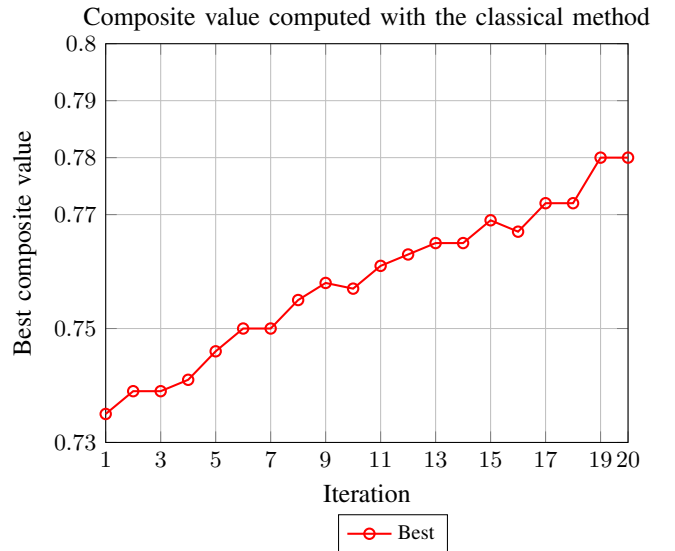


Fig. 10. Best composite value per iteration for the classical PSO approach (single-domain, no surrogate model).

to a wider variety of datasets from diverse domains, to further confirm its versatility and real-world impact.

As a significant avenue for future work, we propose an adaptive extension of the algorithm that incorporates automatic hyperparameter tuning using PSO itself. Specifically, this variant will treat the hyperparameters of the Domain-as-Particle algorithm as an n -dimensional search space, where n is the number of hyperparameters to optimize. PSO will then seek the vector of hyperparameter values that yields the best composite value from the very first iteration, effectively initializing the algorithm in a data-driven, intelligent manner. This approach will allow the algorithm to automatically adapt to the characteristics of each dataset, significantly reducing the time and effort required to manually tune hyperparameters and accelerating convergence toward optimal solutions.

Moreover, we plan to investigate the introduction of a *grandmother domain* concept, a hierarchical multi-level structure within which multiple mother domains operate. This layered framework will enable optimization across larger and more complex weight spaces by coordinating exploration at different granularities, potentially enhancing solution quality and search efficiency.

Finally, future developments will explore new surrogate modeling techniques and innovative exploration strategies to further improve performance and scalability. These advancements aim to broaden the applicability of our framework and address emerging challenges in neural network optimization for real-world applications.

ACKNOWLEDGMENT

This work was inspired by the lecture held by Prof. Paolo Mercorelli entitled: “Applied Algorithms in Estimation and in Control of Technical, Economical, and Biological Dynamical Systems” within the scope of the Complementary Studies Programme at Leuphana University of Lüneburg during the winter semester 2024–2025. In this framework, students can

explore other disciplinary and methodological approaches from the second semester onwards, focusing on additional aspects in parallel with their subjects and giving them the opportunity to sharpen skills across disciplines.

REFERENCES

- [1] K. Benz, C. Rech, and P. Mercorelli, “Sustainable management of marine fish stocks by means of sliding mode control,” in *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*, ser. FedCSIS 2019, vol. 18. IEEE, Sep. 2019. doi: 10.15439/2019f221. ISSN 2300-5963 p. 907–910. [Online]. Available: <http://dx.doi.org/10.15439/2019F221>
- [2] K. Benz, C. Rech, P. Mercorelli, and O. Sergiyenko, “Two cascaded and extended Kalman filters combined with sliding mode control for sustainable management of marine fish stocks,” *Journal of Automation, Mobile Robotics and Intelligent Systems*, p. 28–35, Jul. 2019. doi: 10.14313/jamris/3-2020/30. [Online]. Available: <http://dx.doi.org/10.14313/JAMRIS/3-2020/30>
- [3] D. Normatov and P. Mercorelli, “Parameters estimation of a Lotka-Volterra model in an application for market graphics processing units,” in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2022. doi: 10.15439/2022F61 pp. 935–938.
- [4] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, vol. 4, 1995. doi: 10.1109/ICNN.1995.488968 pp. 1942–1948 vol.4.
- [5] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016. doi: 10.1109/TEVC.2015.2504420
- [6] Y. Zhang, L. Wu, and S. Wang, “Magnetic resonance brain image classification by an improved artificial Bee Colony Algorithm,” *Progress In Electromagnetics Research*, vol. 116, p. 65–79, 2011. doi: 10.2528/pier11031709. [Online]. Available: <http://dx.doi.org/10.2528/PIER11031709>
- [7] T. M. Shami, A. A. El-Saleh, M. Alswaiti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, “Particle swarm optimization: A comprehensive survey,” *IEEE Access*, vol. 10, pp. 10 031–10 061, 2022. doi: 10.1109/ACCESS.2022.3142859
- [8] P. A. Kowalski and M. Kusy, “Determining the significance of features with the use of sobol’ method in probabilistic neural network classification tasks,” in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2017. doi: 10.15439/2017F225 pp. 39–48. [Online]. Available: https://annals-csis.org/Volume_11/drp/225.html

Tools for Implementing Social Innovation in the Circular Economy: Learnings from the CSS Boost project

Cristina Alejandra Barahona Cabrera
0009-0000-7791-2276
Social CRM Research Center
Grimmaische Straße 12
04109 Leipzig Sachsen
Email: cristina.barahona@scrc-leipzig.de

Olaf Reinhold
0000-0003-1977-1641
Duale Hochschule Sachsen
Rittergutsstraße 6
01591 Riesa Sachsen
Email: Olaf.reinhold@dhsn.de

Abstract—Social innovation is a valuable enabler of feedback and collaboration among citizens and stakeholders in the development and implementation of Circular Economy solutions. This article explores the characteristics of current social innovation toolkits, their integration into digital platforms, and their suitability for application in the circular economy, through a literature review and drawing insights from a case, the CSSBoost project. Most toolkits concentrate on business-model ideation and provide little advice for policy or community practice. Only two target circular economy contexts, and just one combines a knowledge repository with an interactive platform.

Index Terms—social innovation, Circular Economy, innovation toolkits, digital platforms.

I. INTRODUCTION

CIRCULAR economy is often positioned as a strategic response to resource depletion and climate risk. A strong interest exists in the circular economy (CE) aspects of governance, transition, and implementation, consumption, and consumer behavior; however, cultural, political, and local practices are often overlooked [1]. This represents an existing gap between the technical solutions promoted and the accompanying social actions needed to implement the solutions.

Social innovation offers an alternative to reduce this gap. Through key actor engagement and incorporation of user insights, it supports the institutional and behavioral change to transform linear supply chains into circular ecosystems.

Systematic knowledge about social innovation tools suitable for circular economy projects is scarce. This limits organizations in their selection of methods to design, pilot, and scale circular solutions on national and international levels.

Therefore, the research goal of the paper is to explore what the characteristics of current social innovation toolkits are, their integration into digital platforms, and their suitability for application in the circular economy. For answering this question, the paper is structured into six sections. Section II summarizes the theoretical background of circular economy and social innovations, section III introduces the CSSBoost project, and section IV outlines the methodological approach. Section V presents the results, finalizing with

section VI, implications and highlights for future work are offered.

II. THEORETICAL BACKGROUND

A. Circular Economy

In the traditional economic model, goods are manufactured, used, and then disposed of. The circular economy proposes an alternative consumption lifecycle, in which products are shared, leased, reused, refurbished, and recycled, resulting in the minimization of environmental impact[2], [3], [4].

The principles of circular economy advocate for waste reduction and countering of premature obsolescence through the improvement of durability, reusability, upgradability, and reparability of products. Circularity not only supports environmental sustainability but also offers companies an opportunity to develop long-term competitiveness. With the adoption of circular practices, savings in materials can be achieved throughout the improved value chains[5].

B. Social Innovation

First introduced as the impact of innovation on society in 1989 [6], social innovation (SI) has evolved into a cross-disciplinary subject, closely linked to creativity, social transformation, technological advancements, and entrepreneurship [7], [8], [9].

The purpose of social innovation can be described as the creation of social change through innovative solutions to communities' pressing, unfulfilled social needs and challenges [10], [11]. Social innovation is driven by a diverse array of stakeholders, including community-based organizations, government entities, social entrepreneurs, academic institutions, as well as nonprofit practitioners[12] [13] [14] [15], [16]. Together, these actors develop initiatives that promote social and environmental wellbeing; such solutions can manifest as products, services, or models [17], concepts, strategies, and tools[18], business models or market innovations [9].

An influential framework that explains how social innovation emerges and unfolds is Mulgan's [19] spiral of social innovation. Beginning with the identification and description

of unmet needs, it progresses through ideation of solutions, which are tested through prototypes and pilots. Successful innovations are then sustained by social innovators, slowly escalating to increase their impact, in some cases, achieving systemic change in concepts, mindsets, or even economic systems.

III. THE CSS BOOST PROJECT

A. About the Project

The European Union, under the Horizon Europe Programme, supports the development of circular systemic solutions (CSS), demonstration projects that implement circular economy principles at a city and regional scale [20]. Funded by the Horizon Programme, the CSSBoost project contributes to this objective through the implementation of five CSS pilots across Europe.

Each pilot targets a different sector, responding to country-specific priorities. In Crete, Greece, the focus is on agriculture and livestock. In Lisbon, Portugal, the actions address the transportation sector. For the North Black Forest in Germany, activities are related to plastics and recycling. In Marche, Italy, it is centered on water recovery. Over the course of the project, targeted interventions are planned to increase value chains' circularity, accompanied by a feasibility analysis for establishing a circular integrated international value chain. In parallel, the development of a regional platform for virtualization, monitoring, and analysis of the value chains is projected (Regional CE/CSS Platform) [21].

B. Potential of Social Innovation

Redesigning value chains to increase their circularity, exploring new business models, and motivating stakeholders to adopt circular practices are all integral to the development of circular systemic solutions. Achieving this goal requires an understanding of stakeholders' dynamics, needs, and motivations to facilitate the proposal of new forms of collaboration and involvement. Within this context, social innovation serves as a facilitator of the circular economy, engaging stakeholders in the development of sustainable solutions. In the CSSBoost Project, the social innovations manifest as improved value chains and strategies guided by Circular Economy principles.

Identified requirements for the implementation of social innovation in systemic circular solutions include promoting collaborative processes, engaging stakeholders and experts, providing specific methods and tools for social innovation, with corresponding training resources for both implementers and participants. Additionally, alignment with the digital Regional CE/CSS Platform is considered; it could facilitate stakeholder engagement, results, and training materials' dissemination.[21] Ultimately, these requirements resulted in a practical need to identify which social innovation tools are

available and are appropriate for circular economy environments.

IV. METHOD

A scoping literature review was conducted to identify existing social innovation tools relevant to the circular economy context. This approach was selected given the aim to map available tools for social innovation.

The search was carried out in Google Scholar, EU project repositories, and project websites hosting social innovation tools. Search terms combined "social innovation" with "tool", "toolkit", "method", and, where possible, "circular". Retrieved resources were screened for addressing social innovation, description of at least one tool or method, and availability of a complete English text. Those lacking citations, a description of purpose and procedural steps, or evidence of applicability were excluded. This process resulted in eight sources suitable for analysis.

The sources were then further examined. First, the intended social innovation outcome was screened, whether the tools were oriented towards creating new business models, services, or community and social projects. Second, completeness was assessed by looking for applied examples, ready-to-fill templates or forms, and sufficiently detailed explanations. Third, each item was checked for relevance to the CSSBoost project, including links to circular economy contexts, support of collaboration among stakeholders, and whether it was supported by a digital platform.

The relevant tools for CSSBoost were organized based on their intended purpose in social innovation: finding or framing social challenges, creating or strengthening stakeholder networks, jointly imagining and creating solutions, planning pathways for implementation and monitoring, and escalating successful solutions. This grouping, aligned to the spiral of social innovation, produced the first draft of the CSSBoost social innovation toolkit for use in circular economy applications.

V. RESULTS OVERVIEW

From the explorative review, eight sources met the inclusion criteria. They differ in their intended social innovation outcome, level of completeness, and degree of digital support (see Table 1). Most resources explain the process for achieving social innovation and encourage the involvement of internal and external stakeholders, yet only in [22] is possible to find a complete practical example.

Brest, Roumani, and Bade [22] offer a guide for decision makers on integrating analytic problem solving, strategic planning, and human-centered design to create social change through projects and services. Progressing similarly to the spiral of social innovation, it references social innovation tools accompanied by a descriptive example but provides no templates or digital platform support.

SEED2SCALE [23] project provides resources for policy-makers and social actors to run collaborative and inclusive social innovation processes across four stages: investigation, ideation, implementation, and evaluation. Materials are distributed through an online knowledge repository.

Komatsu, Deserti, and Rizzo [24] organize the social innovation process around a modified business model canvas: understanding social problems, identifying beneficiaries and supporters, specifying revenue and cost structures, defining channels and relationships, and identifying social impact measures. Worksheets and detailed explanations are provided, but there are no applied examples and no digital platform support.

Lackas et al. [25] present a guide for entrepreneurs and intrapreneurs where the innovation flow begins with a needs assessment, continues with stakeholder analysis, moves into business idea development and prototyping, and concludes with impact analysis, resembling the spiral of social innovation. Tools from design thinking are introduced; however, applied examples or digital platform support are not available.

The Social Impact Academy Toolbox [26] is a training resource to develop social business models. It does not provide applied examples, templates, or support from a digital platform.

The NetZeroCities Social Innovation Toolkit [27] offers tools for participatory workshops to help cities transition towards climate neutrality through business models, social projects, and services. Its sequence parallels the social innovation

spiral, moving through context analysis, problem reframing, envisioning alternatives, prototyping and experimentation, and the evaluation and scaling of promising approaches. A knowledge repository is publicly available, offering basic support from digital platforms.

In Development Impact & You [28], an innovation flow aligned with the spiral of social innovation to support business models, services, social projects, and strategy development is presented. It recommends participatory workshops with clear instructions and printable templates; some tools include illustrative examples. All materials are downloadable from the DIY Toolkit online knowledge repository.

Created for circular economy contexts, the Circular Toolbox [29] focuses on developing socially innovative business models. It offers guided templates and partial examples for mapping opportunities, understanding the customer and the market, prototyping, and improving business model examples. Users can access a knowledge repository and an interactive set of Miro boards that make remote collaboration possible.

Across the reviewed sources, the strongest suggested common topic is the development of social business models through collaborative workshops that broadly reflect the social innovation spiral. However, explicit guidance for circular economy applications and full practical examples remain scarce. Digital platform support emerged as an improvement opportunity, with only one source providing knowledge repositories and interactive templates.

TABLE I.
RESULTS SUMMARY OF THE LITERATURE REVIEW

Source	Intended social innovation outcome	Completeness			Relevance		
		Applied examples	Availability of templates	Detailed explanation of tool usage	Developed for CE context	Promotes collaborative approaches	Supported by digital platforms
Problem Solving, Human-Centered Design [22]	Social projects, services	Yes	No	Yes	No	Yes	No
Seed2Scale [23]	Business models	No	Yes	Yes	No	Yes	Knowledge repository
SI Business Toolbox [24]	Business models	No	Yes	Yes	No	Yes	No
SI Method Toolbox[25]	Business models	No	No	Partial	No	Yes	No
SI Academy Toolbox [26]	Social business model	No	No	No	No	Yes	No
Net Zero Cities SI Toolkit[27]	Business models, social projects, services	No	Yes	Yes	Partial	Yes	Knowledge repository
Development, Impact, and You [28]	Business models, social projects, services, strategies	Partial	Yes	Yes	No	Yes	Knowledge repository
The Circular Toolbox [29]	Business models	Partial	Yes	Yes	Yes	Yes	Interactive Miro Boards

VI. DISCUSSION

A. Theoretical Implications

Overall, social innovation follows an iterative multilateral process, where stakeholders develop solutions to unmet social

needs. These challenges are present in diverse areas, including healthcare, education, economic inequality, social inclusion, and environmental sustainability.

The results of the review suggest that social innovation toolkits are heavily focused on business model development, resulting in a shortage of guiding alternatives for other types of innovation, such as policy concepts, public services, and community practices. Most sources are organized around the process described by Mulgans' spiral of social innovation, along the stages of need discovery, idea generation, prototyping, implementation, scaling, and eventual system change. Common tools described include empathy mapping, persona creation, brainstorming sessions, centering around ideation, and user profiling. Comparatively few instruments offer support for scaling and achieving systemic transformation.

When it comes to social innovation toolkits in circular economic contexts supported by digital platforms, a gap is suggested, with only two sources developed specifically for CE, and among them, only one offers both a knowledge repository and an interactive online workspace.

These findings highlight the opportunity for researchers and organizations to provide targeted platform supported social innovation toolkits for the circular economy.

B. Practical implications

Redesigning value chains and material flows to increase their circularity depends as much on social dynamics as on technical solutions. Social innovation provides mechanisms to align stakeholders, influencing user conduct towards circular behaviors. Involving communities can surface barriers to the adoption of CE principles or products and help organizations move from linear supply chains to ecosystem thinking. Social innovation practices and tools can support the institutional change required for the scaling of CSS through stakeholder networks and targeted education. Strengthening the link between the two fields would offer the opportunity to mobilize stakeholders to address environmental challenges.

CSSBoost implementors face two challenges, first, improving selected value chains to increase their circularity for agriculture, mobility, plastics, and water recovery. Second, the assembly, testing, and dissemination of social innovation tools relevant to CE contexts that support the Regional CE CSS Platform across the four pilots. Once the project comes to an end, learnings about how regional contexts influence social innovation tools adoption, user satisfaction, and contribution to circular economy outcomes constitute a valuable input for the creation of a CSS social innovation guide.

Future studies can broaden the search to include non-English literature and additional grey sources, which will enrich the evidence base. Data from the ongoing CSSBoost pilots can be analyzed once field activities mature, providing empirical insight into how social innovation tools perform in diverse circular economy settings and how digital collaboration features shape stakeholder engagement and learning.

VII. SUMMARY

This paper explored the current publicly available toolkits for social innovation, their suitability for a Circular Economy context, as well as their integration with digital platforms. The

results suggest that most sources concentrate on business model ideation and provide limited guidance for policy, public service, or community innovations; only two address circular contexts, and just one offers interactive digital support.

VIII. ACKNOWLEDGEMENTS

The authors gratefully acknowledge that this research was financially possible thanks to the funding for the Boosting Circular Systemic Solutions through Virtual Regional Circular Economy Spaces (Project number 101135275) from the European Research Executive Agency. The authors also wish to thank the anonymous reviewers for their constructive suggestions.

REFERENCES

- [1] S. Zavos, T. Lehtokunnas, and O. Pyyhtinen, "The (missing) social aspect of the circular economy: a review of social scientific articles," *Sustain. Earth Rev.*, vol. 7, no. 1, p. 11, Apr. 2024, doi: 10.1186/s42055-024-00083-w.
- [2] European Commission, "Circular economy: definition, importance and benefits," Internal Market, Industry, Entrepreneurship and SMEs. Accessed: Jul. 17, 2024. [Online]. Available: https://single-market-economy.ec.europa.eu/industry/strategy/innovation/social_en
- [3] M. Geissdoerfer, P. Savaget, N. Bocken, and E. J. Hultink, "The Circular Economy – A New Sustainability Paradigm?," *J. Clean. Prod.*, 2017, doi: 10.1016/j.jclepro.2016.12.048.
- [4] U. Jakubelskas and V. Skvarciany, "Circular Economy Practices as a Tool for Sustainable Development in the Context of Renewable Energy: What Are the Opportunities for the EU?," *Oeconomia Copernic.*, 2023, doi: 10.24136/oc.2023.025.
- [5] European Commission, "Circular Economy Action Plan." European Union, 2020. [Online]. Available: <https://www.eu2020.de/resource/blob/2429166/156d2d98b66b2ff28b6990161eed91e9/12-17-kreislaufwirtschaftsaktionsplan-bericht-de-data.pdf>
- [6] N. Ayob, S. Teasdale, and K. Fagan, "How Social Innovation 'Came to Be': Tracing the Evolution of a Contested Concept," *J. Soc. Policy*, vol. 45, no. 4, pp. 635–653, Oct. 2016, doi: 10.1017/S004727941600009X.
- [7] S. Guha, S. Majumdar, and N. Marakkath, Eds., *Technology and Innovation for Social Change*, 1st ed. 2015. New Delhi: Springer India: Imprint: Springer, 2015. doi: 10.1007/978-81-322-2071-8.
- [8] R. P. van der Have and L. Rubalcaba, "Social innovation research: An emerging area of innovation studies?," *Res. Policy*, vol. 45, no. 9, pp. 1923–1935, Nov. 2016, doi: 10.1016/j.respol.2016.06.010.
- [9] A. Havas, "Social and business innovations: Linked in practice – but two worlds apart in theorising?," in *Atlas of Social Innovation*, C. Kaletka, A. Schröder, and M. Zingiebl, Eds., oekom verlag, 2019. doi: 10.14512/9783962386887.
- [10] N. Choi and S. Majumdar, "Social Innovation: Towards a Conceptualisation," in *Technology and Innovation for Social Change*, S. Majumdar, S. Guha, and N. Marakkath, Eds., New Delhi: Springer India, 2015, pp. 7–34. doi: 10.1007/978-81-322-2071-8_2.
- [11] G. Mulgan, S. Tucker, R. Ali, and B. Sanders, *Social Innovation: What It Is, Why It Matters and How It Can Be Accelerated*. 2007. [Online]. Available: https://www.researchgate.net/publication/277873357_Social_Innovation_What_It_Is_Why_It_Matters_and_How_It_Can_Be_Accelerated
- [12] D. Chalmers, "Social innovation: An exploration of the barriers faced by innovating organizations in the social economy," *Local Econ. J. Local Econ. Policy Unit*, vol. 28, no. 1, pp. 17–34, Feb. 2013, doi: 10.1177/0269094212463677.
- [13] R. Grimm, C. Fox, S. Baines, and K. Albertson, "Social innovation, an answer to contemporary societal challenges? Locating the

- concept in theory and practice,” *Innov. Eur. J. Soc. Sci. Res.*, vol. 26, no. 4, pp. 436–455, Dec. 2013, doi: 10.1080/13511610.2013.848163.
- [14] H. Jiao, “A conceptual model for social entrepreneurship directed toward social impact on society,” *Soc. Enterp. J.*, vol. 7, no. 2, pp. 130–149, Aug. 2011, doi: 10.1108/17508611111156600.
- [15] P. G. Svensson, T. Q. Mahoney, and M. E. Hambrick, “What Does Innovation Mean to Nonprofit Practitioners? International Insights From Development and Peace-Building Nonprofits,” *Nonprofit Volunt. Sect. Q.*, vol. 49, no. 2, pp. 380–398, Apr. 2020, doi: 10.1177/0899764019872009.
- [16] L. Dryjańska, J. Košťálová, and D. Vidović, “Higher Education Practices for Social Innovation and Sustainable Development,” *Innov. Technol. Knowl. Manag.*, 2022, doi: 10.1007/978-3-030-84044-0_6.
- [17] Louise Pulford and Filippo Addarii, “This is European social innovation,” *Enterprise & Industry online magazine*, 2010. doi: 10.2769/825.
- [18] P. Dawson and L. Daniel, “Understanding social innovation: a provisional framework,” *Int. J. Technol. Manag.*, vol. 51, no. 1, p. 9, 2010, doi: 10.1504/IJTM.2010.033125.
- [19] R. Murray, J. Caulier-Grice, and G. Mulgan, *The Open Book of Social Innovation*. in Social innovator series: ways to design, develop and grow social innovation. 2010. [Online]. Available: http://temp.uefisc-di.ro/edigiregion_v2/the_open_book_of_social_innovationNESTA.pdf
- [20] European Commission, “Circular Cities and Regions Initiative (CCRI)’s circular systemic solutions.” Accessed: May 22, 2025. [Online]. Available: <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl6-2023-circbio-02-1-two-stage;callCode=null;freeTextSearchKeyword=CircBio;matchWholeText=true;typeCodes=0,1,2,8;statusCodes=31094501,31094502;programmePeriod=null;programCcm2Id=null;programDivisionCode=null;focusAreaCode=null;destinationGroup=null;missionGroup=null;geographicalZonesCode=null;programmeDivisionProspect=null;startDateLte=null;startDateGte=null;crossCuttingPriorityCode=null;cpvCode=null;performanceOfDelivery=null;sortQuery=sortStatus;orderBy=asc;onlyTenders=false;topicListKey=topicSearchTablePageState>
- [21] CSSBoost Project, “Boosting Circular Systemic Solutions through Virtual Regional Circular Economy Spaces.” Accessed: May 23, 2025. [Online]. Available: <https://cssboost-project.eu/>
- [22] P. Brest, N. Roumani, and J. Bade, “Problem Solving, Human-Centered Design, and Strategic Processes.” [Online]. Available: <https://pacscenter.stanford.edu/wp-content/uploads/2015/09/Download-the-full-article-here.pdf>
- [23] “SI Toolbox,” Seed2Scale. [Online]. Available: <https://www.seede-uproject.eu/learning-repository/toolbox/>
- [24] T. Komatsu, A. Deserti, and F. Rizzo, “Social Innovation Business Toolbox,” 2016. [Online]. Available: https://www.simpact-project.eu/tools/toolbox_business_web.pdf
- [25] M. Lackas, T. Zebahl, J. Thielen, M. Nägelsbach, P. Meyer, and A. Schröer, “Social Innovation Method Toolbox (for entrepreneurs and intrapreneurs in the field social service provision),” Center for Social Investment, Heidelberg University, Heidelberg. [Online]. Available: https://dtp.interreg-danube.eu/uploads/media/approved_project_public/0001/49/bc8b15f28fa323e973acb7dc279cd3404b750ded.pdf
- [26] Interreg, “Social Innovation Academy: How to innovate together for a better tomorrow.” 2020. [Online]. Available: https://socialimpact.eu/fileadmin/user_upload/social_impact/themen/international/Dateien/S_i_M_D.C.5.1_toolbox_ENG.pdf
- [27] T. Komatsu Cipriani, “Social Innovation Toolkit.” 2023. [Online]. Available: <https://netzerocities.app/resource-3121>
- [28] Theo Keane *et al.*, *DIY - development impact & you: practical tools to trigger & support social innovation*. London, 2014. [Online]. Available: <https://media.nesta.org.uk/documents/diy-toolkit-full-download-a4-size.pdf>
- [29] Circle Economy, “The Circular Toolbox,” The Circular Toolbox. [Online]. Available: <https://www.thecirculartoolbox.com/>

Modeling and optimizing flow networks with several constraints using sequential dynamical systems

Jens Dörpinghaus^{*†}, Michael Tiemann^{*†}, Robert Helmrich^{*‡},

^{*} 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>

[‡] University of Bonn, Bonn, Germany

Abstract—This paper introduces a novel framework for modeling and optimizing flow networks with multiple constraints using Sequential Dynamical Systems (SDS). We have extended the Sequential Flow Network (SFN) definition to encompass both Sequential Flow Networks (SFN) and Bounded Sequential Flow Networks (bSFN), which incorporate directional constraints and weighted transitions. These models can be utilized to simulate intricate real-world applications, such as educational pathways and labor market issues. In order to optimize local flow to specific nodes with minimal global impact, we propose three novel approaches: a linear programming formulation and two greedy heuristics. The evaluation metrics employed are defined as a means to balance local improvement and global disturbance. The efficacy of these methods is evaluated through experimentation on both artificial and real-world-inspired random networks. Notwithstanding the encouraging results and observations yielded by the experimental analysis of random graphs, suggestions for further research will be made in order to overcome the limitations of the present study.

I. INTRODUCTION

A CONSIDERABLE number of real-world problems manifest as phenomena or dynamics over networks and graphs. For instance, urban traffic and transportation networks can be represented as graphs, as illustrated in [1]). As another example, the spread of disease on social contact graphs is naturally represented in graphs [2]. Other examples include packet flow in information and engineering networks, such as cell phone communication, gene annotation and gene regulatory network (GRN), or optimization of SDS schedules, see [3]. Thus, SDSs use networks for modeling, simulation, and analysis. Networks are also widely used in other context for data modeling and analysis, see for example [4].

The generic class of Graph Dynamical Systems (GDS) is distinct from other dynamical systems. These systems operate on discrete time and may utilize a finite number of states. Consequently, classical dynamical systems theory and tools frequently prove to be inapplicable. While GDS are rooted in discrete mathematics, algebra, combinatorics, graph theory, and probability theory, they are primarily utilized within the context of computer simulation.

The central research question guiding this study is concerned with the utilization of SDSs in the modeling and optimization with several constraints of flow networks on natural numbers, characterized by linear transition functions. The primary emphasis of this study will be on real-world

problems derived from labor market research. A comprehensive literature review will precede a concise discussion of the methods, tools, and theory for validation and theoretical insights of SDSs that are necessary for modeling flow networks with SDSs. In this study, we will propose three methodologies for addressing the aforementioned problem. The first is an approach based on linear programming, and the second and third are greedy heuristics. Subsequent to this discussion, we will present and analyze a series of experimental results. The study's conclusions and outlooks are articulated in the final section.

II. LITERATURE REVIEW

Research on GDS and SDS remains limited. For a comprehensive introduction to SDSs, see [5] and [3]. A close relationship exists between these models and Generalized Cellular Automata with parallel update schemes, see [6]. SDSs with sequential update schemes were introduced between 1999 and 2001 by Barrett et al, see [7]. Another related concept is that of stochastic graph dynamical systems, see [8].

The examination of flow networks in graphs is not a novel concept; see [9], [10]. However, to the best of our knowledge, there is a lack of literature on modeling flow in graphs with dynamical systems. The optimization of flow, whether local or global, is a subject of study within the framework of classical dynamical systems theory, see [11]. This concept has also been explored in the context of distributed systems [12], chemical systems [13], and traffic networks [14]. However, these issues are frequently addressed through the implementation of optimization methodologies or, in certain instances, artificial intelligence algorithms, see [15], [16]

In summary, the field of SDSs is not generally associated with flow networks. The objective of this study is to examine the feasibility of leveraging methodologies from linear programming to enhance the operational efficiency of flow networks in SDSs, which are characterized by multiple constraints.

III. METHOD

In this section, we will first introduce Sequential Dynamical Systems (SDS) and then develop the novel concept of flow networks modeled with SDS. Subsequently, the issue of local optimization of these flow networks will be presented.

The following three approaches will be introduced: a linear programming (LP) approach and two greedy approaches.

A. Sequential Dynamical System

An SDS consists of the following parts, which we will illustrate with a continuous example introduced by [3]. However, we will not strictly follow their notation and will add other remarks that focus on our research. First, we need a **Graph** $G = (V, E)$ with vertices V and edges E .

Example III.1. For example we may use a circular graph on four nodes: $V = \{v_0, \dots, v_3\}$ and $E = \{(v_0, v_1), (v_1, v_2), (v_2, v_3), (v_3, v_0)\}$.

Each node has a particular **vertex state** x_i from a state set K , for example $K = \mathbb{F}_2 = \{0, 1\}$. This results in a **system state**, which is also called the *configuration* of an SDS. For G in the example III.1 the system state contains four vertex states:

$$x = (x_0, x_1, x_2, x_3).$$

Next, we use a G -**local function** $F_i : K^n \rightarrow K^n$, which is also called a *local transition function*. It takes the system state as input.

For each vertex $v_i \in V$ $f_{v_i} : K^{d(v_i)+1} \rightarrow K$ is called the *vertex function*. Here, $d(v)$ denotes the degree of vertex v and $N(v)$ its closed neighborhood. The input set is vertex state of the node v_i and the vertex state of all its neighbors, denoted by $x[N(v_i)]$. We can define the local function F_v node-wise as by

$$F_{v_i} = F_i = (x_0, \dots, x_{i-1}, f_{v_i}(x[N(v_i)]), x_{i+1}, \dots, x_n)$$

Example III.2. Continuing Example III.1 we may set

$$F_0(x_0, x_1, x_2, x_3) = (\text{nor}_3(x_0, x_1, x_3), x_1, x_2, x_3)$$

$$F_1(x_0, x_1, x_2, x_3) = (x_0, \text{nor}_3(x_0, x_1, x_2), x_2, x_3)$$

$$F_2(x_0, x_1, x_2, x_3) = (x_0, x_1, \text{nor}_3(x_1, x_2, x_3), x_3)$$

$$F_3(x_0, x_1, x_2, x_3) = (x_0, x_1, x_2, \text{nor}_3(x_2, x_3, x_0))$$

It is important to note that this function can only change the state of vertex i . We apply the updates sequentially and therefore need to define a **order**, e.g. $\pi = (0, 1, 2, 3)$.

To start the system, we define a **initial state** $x^0 = (x_0^0, \dots, x_n^0)$. Typically, the context provides sufficient clarity regarding the intended state, thereby obviating the necessity for dual indexes.

Example III.3. Continuing example III.1, we may set

$$x^0 = (x_0, x_1, x_2, x_3) = (1, 1, 0, 0)$$

By applying the maps we get $(1, 1, 0, 0) \xrightarrow{F_0} (0, 1, 0, 0) \xrightarrow{F_1} (0, 0, 0, 0) \xrightarrow{F_2} (0, 0, 1, 0) \xrightarrow{F_3} (0, 0, 1, 0)$.

Effectively we have applied the composed map $F_3 \circ F_2 \circ F_1 \circ F_0$. In a more algorithmic perspective, each step of an SDS involves n substeps:

Algorithm 1 System Update

```

1: for  $i = 1$  to  $n$  do
2:    $x_{\pi(i)} = f_{\pi(i)}(x[N(v_{\pi(i)})])$ 
3: end for

```

In summary, A SDS is thus defined by a graph G , $\mathbf{F}_G = (F_v)_{v \in V}$, which is the vertex-indexed family of vertex functions and π :

Definition III.4 (Sequential Dynamical System, SDS, see [3]). Let $G = (V, E)$ be a graph, let $(f_v)_{v \in V}$ be a family of vertex functions, and let $\pi = (v_{\pi(1)}, v_{\pi(2)}, \dots, v_{\pi(n)})$ be a permutation of the vertices of G . The sequential dynamical system (SDS) is the triple

$$(G, (F_v)_{v \in V}, \pi).$$

Its associated SDS-map is $[\mathbf{F}_G, \pi] : K^n \rightarrow K^n$ defined by

$$[\mathbf{F}_G, \pi] = F_{\pi_n} \circ F_{\pi_{n-1}} \circ \dots \circ F_{\pi_1}.$$

Often, scenarios are considered where G is undirected. Thus, if not specified, we will assume that G is undirected. The application of the G -local map F_v is the *update of vertex v* , and the application of $[\mathbf{F}_G, \pi]$ is a *system update*, see Algorithm 1.

To visualize the behavior of an SDS we may use a table representing all node states in a table, where each column represents a particular node state and each row represents a system update. In the case of $K = \mathbb{F}_2$, a vertex state that is zero is represented as a white square and a vertex state that is one is represented as a black square. Consider the previous example:

$x^0 =$	(1,	0,	1,	0)
x^0	■		■	
x^1				■
x^2		■		
x^3			■	
x^4	■			
x^5		■		■
x^6			■	
x^7	■		■	

This table represents the so-called *forward orbit* of $x = (1, 0, 1, 0)$:

Definition III.5 (Forward Orbit, see [3]). Let x be a system state of a SDS with system update function $[\mathbf{F}_G, \pi]$. The forward orbit is given by

$$O^+(x) = (x, [\mathbf{F}_G, \pi](x), [\mathbf{F}_G, \pi]^2(x), [\mathbf{F}_G, \pi]^3(x), \dots).$$

However, this only represents the *forward orbit* of an initial state. To visualize a complete SDS we may use Phase spaces, see [3] for more details. It is obvious that we can only visualize a finite state of system states with these approaches. Thus, it is common to analyze the dynamics of sequential dynamical systems defined using classical Boolean functions. They have several nice properties, including symmetry:

Definition III.6. We say that a function $f(x_1, \dots, x_n)$ is symmetric if the order in which we describe its inputs does not change the output: i.e. if $f(x_1, \dots, x_n) = f(x_{\pi(1)}, \dots, x_{\pi(n)})$, for any permutation π .

Some functions like nor, nand, or and and are all symmetric. Other functions may not, especially the linear functions which we will consider in the next section are generally not symmetric.

B. Flow Networks

We will now consider flow networks. Usually they are studied as another problem in graph theory where shortest paths can be useful [9], [10]. In fact, the Maximum Flow Problem is e so useful that we can apply them to many practical problems. Suppose we want to transport a good from one point to another, for example water, natural gas, oil, or electricity. In these networks edges refer to some kind of pipe or route to transport these good, not only pipes, but also roads, or railways. Here, the question is: How can we send as much as possible? Or in other words: How can we maximize the flow?

However, when simulating flow with dynamical systems, the question is usually rather: How can we manipulate the flow so that particular nodes get more goods, for example without affecting the whole other nodes? So for example, how can one particular factory get more resources without the need that other factories decrease their production.

In this case, a flow network $G = (V, E)$ is a directed graph comprising, here every node $v \in V$ may be a source node and a target node. We set $K = \mathbb{N}$. Each edge $e_i \in E$ has a weight $w(e_i) \in \mathbb{R}$ which identifies the share of how many goods from the source node “go” to the target node. Attention: This is not the capacity, but a real share. When we define $E^-(v)$ as the set of incoming edges for node $v \in V$, we can define the the generic class of SDS flow networks:

Definition III.7 (SDS Flow Network (SFN)). Let $G = (V, E)$ be a directed weighted graph with edge weights $w : E \rightarrow \mathbb{R}$. Let $(f_v)_{v \in V}$ be a family of vertex functions with $f_v : K^{d(v_i)+1} \times E^-(v) \rightarrow K$, and let $\pi = (v_{\pi(1)}, v_{\pi(2)}, \dots, v_{\pi(n)})$ be a permutation of the vertices of G . The sequential dynamical system (SDS) defined by $(G, (F_v)_v, \pi)$ is called a sequential flow network (SFN).

In summary, the basic difference to an SDS is the usage of weights in the vertex functions. The difference to generic flow networks is that every node is a source and target node at the same time. However, these networks may have further constraints and we will develop some of them by starting with a very simple example and extend it.

We have n nodes $v_1, \dots, v_n$ and we may group all other nodes in q groups $V_1, \dots, V_q$ to keep track of the update order. This step is not technically necessary but allows to define constraints.

We can model a small network with educational pathways, as described in Figure 1. Here, $V = \{ \underbrace{V_0}_{v_0}, \underbrace{V_1}_{v_1, v_2, v_3}, \underbrace{V_2}_{v_4, v_5} \}$,

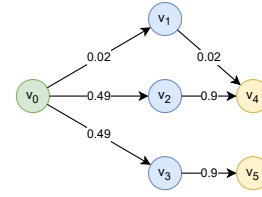


Fig. 1: An example of educational pathways. Green nodes are in V_0 , blue nodes in V_1 and yellow nodes in V_2

and we have six nodes with a certain weight, see Figure 1. Here, the group V_0 represents the ‘incoming’ actors on the labor market, for example school-leavers, V_1 represents school-leaving certificates and V_3 vocational or academical education. Since education and qualification is usually not ‘lost’, we can assume that a particular share of people with a given qualification achieve another qualification during a given time, for example a year. Here, $w(e_i) \in [0, 1]$. In our artificial and simplified example, 49% of all people entering the labor market receive school leaving certificate v_3 and 90% of them get a vocational degree v_5 . Note, that in this example we only allow edges between nodes u, v with $u \in V_i$ and $v \in V_{i+1}$.

With this, We can now define the the G-local function as

$$F_i(x_0, \dots, x_n) = (x_1, \dots, x_i + \underbrace{\sum_{j \in N^-(v_i)} w((v_j, v_i))}_{\text{position } i} x_j, \dots, x_n).$$

This function sums over all incoming edges and adds the share defined as edge weight of the system state of the incoming node.

By using the reverse order $\pi = (n, n-1, \dots, 2, 1, 0)$ we can now define an SDS $(G, (F_v)_v, \pi)$ and compute the forward orbit for a given initial state, for example

$$x = (10000, 2000, 80000, 40000, 130000, 110000).$$

Thus, we will apply the composed map $F_5 \circ F_4 \circ F_3 \circ F_2 \circ F_1 \circ F_0$ step by step as follows:

$$\begin{aligned} F_5 &= (10000, \dots, \underbrace{110000}_{=x_5} + (\underbrace{0.9}_{=w((v_3, v_5))} \cdot \underbrace{40000}_{=x_3})) \\ &\quad \underbrace{\hspace{10em}}_{=146,000} \\ F_4 &= (\dots, \underbrace{130000}_{=x_4} + (\underbrace{0.02}_{=w((v_1, v_4))} \cdot \underbrace{2000}_{=x_1}) + (\underbrace{0.9}_{=w((v_2, v_4))} \cdot \underbrace{80000}_{=x_2}), \dots) \\ &\quad \underbrace{\hspace{10em}}_{=202,040} \\ F_3 &= (\dots, \underbrace{40000}_{=x_3} + (\underbrace{0.49}_{=w((v_0, v_3))} \cdot \underbrace{10000}_{=v_0}), \dots) \\ &\quad \underbrace{\hspace{10em}}_{=44,900} \end{aligned}$$

Here, the colors refer to the coloring in Figure 1. So we get the following forward orbit:

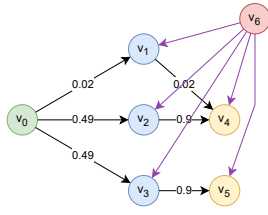


Fig. 2: An example of educational pathways. Green nodes are in V_0 , blue nodes in V_1 and yellow nodes in V_2 , red nodes in V_3

	v_0	v_1	v_2	v_3	v_4	v_5
x^0	10,000	2,000	80,000	40,000	130,000	110,000
x^1	10,000	2,200	84,900	44,900	203,040	146,000

As we can see, the number of people with a certain education raises every time step. This is not very natural, and we can thus extend the model by adding negative weights $w(e_i) \in [-1, 1]$ and adjusting the G-local function with

$$u(x_i, x_j) = \begin{cases} e((v_j, v_i))x_j & e((v_j, v_i)) \geq 0 \\ e((v_j, v_i))x_i & e((v_j, v_i)) < 0 \end{cases}$$

$$F_i(x_0, \dots, x_n) = (x_1, \dots, x_i + \underbrace{\sum_{j \in N^-(v_i)} u(x_i, x_j)}_{\text{position } i}, \dots). \quad (1)$$

With this, we can also model an increasing flow with backward edges. Extending our previous example with another node in V_3 and backward edges with a small share to all other nodes, see Figure 2, we can model the share of people leaving the labor market, for example because of death or retirement. Here, $e(v, v_6) = -0.01 \forall v \in V \setminus \{v_0, v_6\}$.

Again, we can compute the forward orbit. Let

$$x = (10000, 2000, 80000, 40000, 130000, 110000, 0)$$

be the initial state. Then

$$F_6 = (10,000, \dots, 0)$$

$$F_5 = (10,000, \dots, \underbrace{110000 - (0.01 \cdot 110000) + (0.9 \cdot 40000)}_{=144,900}, 0)$$

$$F_4 = (\dots, \underbrace{130000 - (0.01 \cdot 130000) + (0.9 \cdot 80000) + (0.02 \cdot 2000)}_{=200,740}, \dots)$$

$$F_3 = (\dots, \underbrace{40000 - (0.01 \cdot 40000) + (0.49 \cdot 10000)}_{=44,500}, \dots)$$

$$F_2 = (\dots, \underbrace{80000 - (0.01 \cdot 80000) + (0.49 \cdot 10000)}_{=84,100}, \dots)$$

$$F_1 = (10,000, \underbrace{2000 - (0.01 \cdot 2000) + (0.02 \cdot 10000)}_{=2,180}, \dots)$$

$$F_0 = (10,000, \dots, 0)$$

	v_0	v_1	v_2	v_3	v_4	v_5	v_6
x^0	10,000	2,000	80,000	40,000	130,000	110,000	0
x^1	10,000	2,180	84,100	44,500	200,740	144,900	0

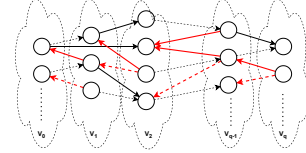


Fig. 3: Illustration of the concept of Bounded SDS Flow Networks: The positive flow goes from left to right, the negative flow from right to left

Which leads to the following forward orbit: We can make two observations: First, the vertex state of v_6 never changes. However, if we add ‘backward’-edges, e.g. (v_5, v_6) with $w(v_5, v_6) = -w(v_6, v_5)$, this simulates the desired behavior. The second observation is that the newly added negative edges conflict with our assumption that edges are only between two nodes in groups V_i and V_{i+1} . This rule is not realistic, especially since some education may rely on other education in the same group, for example elementary school as dependency for higher school degrees. Thus, we define a specific subset of SDS flow networks:

Definition III.8 (Bounded SDS Flow Network (bSFN)). *Let $G = (V, E)$ be a directed weighted graph with edge weights $w : E \rightarrow \mathbb{R}$. Let $(f_v)_{v \in V}$ be a family of vertex functions with $f_v : K^{d(v_i)+1} \times E^-(v) \rightarrow K$, and let $\pi = (v_{\pi(1)}, v_{\pi(2)}, \dots, v_{\pi(n)})$ be a permutation of the vertices of G .*

All n nodes $v_0, \dots, v_n$ are grouped in q groups $V_0, \dots, V_q$. We allow edges $e = (u, v)$ with $w((u, v)) \geq 0$ only between nodes $u \in V_i, v \in V_j$ with $i \leq j$ or edges $e = (u, v)$ with $w((u, v)) < 0$ only between nodes $u \in V_i, v \in V_j$ with $j \leq i$.

The sequential dynamical system (SDS) defined by $(G, (F_v)_{v \in V}, \pi)$ is called a bounded flow network (bSFN).

If not otherwise mentioned, we will assume that π is the reverse order $(n, \dots, 1, 0)$. This means that the flow in this network is bounded by V_0 and V_q , see Figure 3 for an illustration. We can now study scenarios where we apply methods to optimize the local flow in these networks.

Local Optimization of Flow Networks

In general, the local optimization of flow networks refers to the maximization of the incoming flow to one particular node. So let $v_i \in V_i$ with $0 < i < q$ and x an initial system state and $x' = [F_G, \pi](x)$. How can we change G so that we maximize x'_i whereas we want to keep $\delta_j = |x_j - x'_j| \forall j \in [0, \dots, n], j \neq i$ minimal. In other words, after the system update the system state for v_i should be maximized whereas ideally all other nodes have the same system state or change minimally.

While generally we can use or add new positive nodes from all nodes in V_i and all V_j with $j < i$ and use or add new

negative nodes from all nodes in V_i and all V_j with $j > i$, we may restrict the candidate set for positive edges to $C^+ \subseteq E$ and the candidate set for negative edges to $C^- \subseteq E$.

We set $c_{ij}^+ = w((v_i, v_j))$ for $v_i, v_j \in V$, $i \leq j$ if $(v_i, v_j) \in E$ and else $c_{ij}^+ = 0$. Similarly, we set $c_{ij}^- = w((v_i, v_j))$ $v_i, v_j \in V$, $i \geq j$ if $(v_i, v_j) \in E$ and else $c_{ij}^- = 0$. In Figure 4 we describe the adjacency matrix representation, where each entry holds the weight of an edge or 0 if no edge exists. The lower triangular part (green) represents values in C^+ , the upper triangular part the values in C^- :

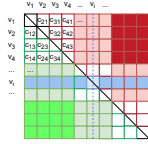


Fig. 4: Adjacency matrix representation

It is easy to see that the updated value of v_ι can be computed by considering the ι -th row of the matrix. So we want to maximize

$$x'_\iota = x_\iota + \sum_{v_i \in V; (i, \iota) \in E} c_{i\iota}^+ x_i - \sum_{v_i \in V; (\iota, i) \in E} c_{i\iota}^- x_i.$$

However, only those values of $c_{i\iota}$ that are in C^+ or C^- are subject to optimization. All other variables are fixed. In addition, this formula does not update those vertex states which are updated before $\pi^{-1}(\iota)$, leading to an approximation of the optimal solution.

However, if we restrict the update schema to $\pi = (n - 1, \dots, 1, 0)$, the update of v_i does not influence any other nodes in the current system update:

Lemma III.9. *Let $\pi = (n - 1, \dots, 1, 0)$ be the update schema for a Bounded SDS Flow Network on n nodes. Then, the vertex state of x_i is not included in the vertex update function of any other node v_j with $j < i$.*

Proof. Let v_j be a node with a vertex update function that includes x_i with $i > j$. Then, according to Formula 1, and edge (v_i, v_j) with positive weight would exist. This is a contradiction to Definition III.8. $\square$

More generic, changing the value of x'_ι influences all other summands for other update values x'_j , $j \neq \iota$ which is represented by the ι -th column in the adjacency matrix. If π is the ordering $(0, 1, \dots, n)$, this only affects the lower triangular part, if $\pi = (n, \dots, 1, 0)$ it only affects the upper triangular part.

From now on, we will assume the update schema is $\pi = (n - 1, \dots, 1, 0)$. Since C^+ and C^- includes edges not connected to v_ι – in the first system update only x_ι is changed, we will compare $[\mathbf{F}_G, \pi]^2(x)$ to $[\mathbf{F}_{G'}, \pi]^2(x)$ where G' is the graph with adjusted weights. We can then compute

the distance between the expected and modified vertex state of all other nodes as

$$\delta = |[\mathbf{F}_G, \pi]^2(x) - [\mathbf{F}_{G'}, \pi]^2(x)|$$

In summary, we can model this as optimization problem using a linear program:

$$\max \quad z = x'_\iota - \delta \quad (2)$$

$$\text{s.t. } x'_\iota = x_\iota + \sum_{v_i \in C^+} c_{i\iota}^+ x_i - \sum_{v_i \in C^-} c_{i\iota}^- x_i \quad (3)$$

$$\delta = \sum_{j=0, \dots, n-1} |([\mathbf{F}_G, \pi]^2(x))_j - ([\mathbf{F}_{G'}, \pi]^2(x))_j| \quad (4)$$

$$c_{ij}^+ \in [0, 1], c_{ij}^- \in [-1, 0] \text{ for all } c_{ij}^+ \in C^+, c_{ij}^- \in C^- \quad (5)$$

Here, line 3 gives the objective function for the updated system state, omitting the update of other node's state. Line 4 returns the influence of changing edge weights, using the modified state of v_ι .

Example III.10. *Coming back to our initial example in the last section, and let $C^+ = \{e_{1,4}, e_{2,4}, e_{3,4}\}$, the optimization approach is as follows:*

$$\max \quad z = 2000c_{14} + 80000c_{24} + 40000c_{34} - \delta + 128700.0 \quad (6)$$

$$\text{s.t. } \delta = -83.2 + 4160c_{14} + 163300c_{24} + 84100c_{34} \quad (7)$$

The optimal solution in this trivial case is $c_{14} = c_{24} = c_{34} = 1$.

However it is also possible to use a greedy approach to maximize the local flow. The simplest approach would just set the value for all incoming edges to node v_ι to the maximum positive value. This means, if we restrict the edge weights to the range $[-1, 1]$ to set all edge weights in C^+ to 1 and in C^- to zero:

Algorithm 2 Greedy 1

Ensure: bSDS-FN $G = (V, E)$, edge weights restricted to $[a, b]$, allowed edge sets C^+ and C^- , and a node $v_\iota \in V$

- 1: **for** $(v, v_\iota) \in C^+$ **do**
- 2: $w((v, v_\iota)) = b$
- 3: **end for**
- 4: **for** $(v, v_\iota) \in C^-$ **do**
- 5: $w((v, v_\iota)) = \max(0, a)$
- 6: **end for**

However, this brute-force approach will significantly influence parts of the network. The changes are limited to those nodes, which are reachable from node v_ι . Considering Example 2, we see that changing x_4 or x_5 has no influence on other nodes. We can measure this influence using the betweenness centrality measure [17], [18], which is the sum of the fraction of all-pairs shortest paths that pass this particular

node; it was first introduced by [19]. Given a node v , it calculates all shortest paths in a network $P_v(k, j)$ for all beginning and ending nodes $k, j \in V$. If $P(k, j)$ denotes the total number of paths between k and j , the importance of v is given by the ratio of both values. Thus, the betweenness centrality according to [20] is given by

$$bc(v) = \sum_{k \neq j, v \neq k, v \neq j} \frac{P_v(k, j)}{P(k, j)} \cdot \frac{2}{(n-1)(n-2)},$$

Coming back to Example 2, we see that the bc -value of all nodes is zero, except $bc(v_1) = bc(v_2) \approx 0.01$, and $bc(v_3) \approx 0.08$.

Algorithm 3 Greedy 2

Ensure: bSDS-FN $G = (V, E)$, edge weights restricted to $[a, b]$, allowed edge sets C^+ and C^- , and a node $v_l \in V$

- 1: **for** $(v, v_l) \in C^+$ **do**
- 2: $w((v, v_l)) = \min\{\max\{w((v, v_l)) \cdot bc(v_l), bc(v_l)\}, b\}$
- 3: **end for**
- 4: **for** $(v, v_l) \in C^-$ **do**
- 5: $w((v, v_l)) = \max\{\min\{w((v, v_l)) \cdot bc(v_l), -bc(v_l)\}, a\}$
- 6: **end for**

D. Evaluation Metrics

As we have discussed in the last section, the optimization approach has two foci: First, it tries to optimize the flow towards one particular node, second, it tries to keep the influence on the whole network flow minimal. Thus, while we want to maximize

$$\delta_k(x_l) = x_l^k - x_l^{k-1},$$

we want to minimize

$$\Delta_k(X) = \sum_{i \neq l} |x_i^k - x_i^{k-1}|.$$

Thus, our evaluation will be based on both metrics.

IV. EXPERIMENTAL RESULTS

We used Python 3.11.2 with Pulp and NetworkX for creating random instances and implement the greedy heuristic as well as the Linear Program. We used GLPK (GNU Linear Programming Kit) 5.01 to solve the linear program. Random instances are build for a particular number of nodes n and a given probability $p \in [0, 1]$ that two nodes are connected when not violating the conditions defined previously. Edge weights are randomly chosen from the same interval. We define C^+ and C^- with all possible direct edges. For all instances, we used 40 iterations to compare the results with Δ_k and δ_k with $k = 2$.

A. Small networks

In Table I we present the average, minimal and maximal error rates for $n = 200$ and $n = 400$ and different values for p . For a visualization of the corresponding values we refer to Figures 5 and 6.

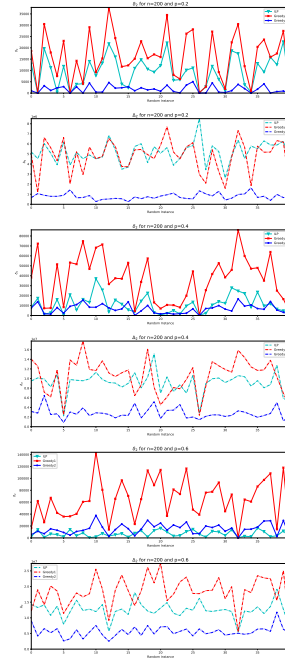


Fig. 5: Measures for different values of p and $n = 200$

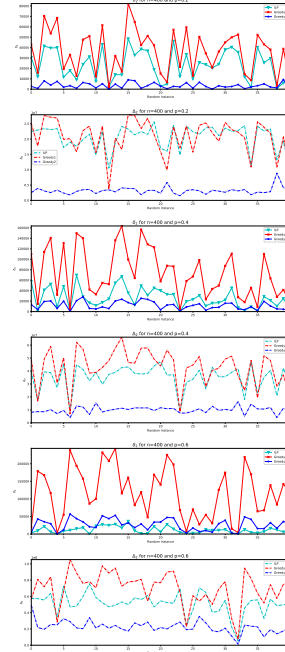


Fig. 6: Measures for different values of p and $n = 400$

For flow networks which are not so dense ($p = 0.2$) we see that the LP-approach outperforms Greedy1 with both metrics. However, for flow networks with more edges, Greedy1 produces better local improvements δ , but also much higher global errors Δ . Greedy2 generally produces best results, while keeping the global error low – but also with lower local

improvement. However, comparing to the IP solutions tend to be at least comparable and for larger p IP even outperforms Greedy2. The overall behavior is similar for larger instances, see the results for $n = 400$.

B. Real-world inspired random networks

The German Labor Market Ontology (GLMO) was developed to facilitate the modeling of labor market flows. The ontology in question was developed principally on the basis of two sources: the multilingual European ontology for occupations and skills, known as ESCO, and the German classification of occupations, denoted as KldB. In addition to the comprehensive classification system of KldB occupations, the GLMO encompasses sets of skills, tools, and educational training relevant to the German labor market. These competencies are delineated by the BA and will be designated as BA skills, tools, and educational training in the subsequent discussion. In the ontology, the concepts are organized in a hierarchical structure and mapped to KldB-occupational unit groups [21], [22]. In [23], the ontology was expanded by incorporating data from BERUFENET, an online portal that provides information on KldB occupational titles. BERUFENET organizes these occupational titles into distinct study fields, activity fields, and activity areas. In addition, this encompasses mappings to associated or alternative occupations, supplementary qualifications, and other CVET categories from KURSNET, along with information fields comprising extensive additional information (e.g., competencies, abilities, knowledge, and skills, cf. [24], [25]). It is possible to create a comprehensive model of the German labor market that incorporates educational pathways, in conjunction with additional data, such as information regarding individuals with specific training.

In order to analyze the efficacy of our approach for these networks, we created an additional set of larger random instances. Here, 2,000 nodes were utilized, and the probability p set at 0.4. The results are presented in Figure 7 and Table II. It is evident that Greedy2 demonstrates superior performance in comparison to IP for larger instances. In order to determine the most efficacious course of action, a rigorous examination of the pertinent real-world applications is imperative. This involves a judicious evaluation of the relative merits of a modest yet substantial increase, characterized by a limited global impact, as compared to a more pronounced increase accompanied by a concomitant global error (IP). Therefore, given the consideration of these three approaches, it can be concluded that a universal solution does not exist.

V. CONCLUSIONS AND OUTLOOK

In this study, a novel approach for modeling and optimizing flow networks with multiple constraints was introduced. This approach was developed using the framework of Sequential Dynamical Systems (SDS). The extension of the system dynamics (SDS) framework to encompass sequential flow networks (SFN) and a subclass of bounded SDS flow networks (BSFN) has yielded a highly versatile structure for the simulation and analysis of complex networked systems. In

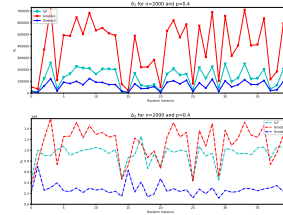


Fig. 7: Measures for different values of p and $n = 2000$

these systems, each node possesses the capacity to function as both a source and a sink, thereby enhancing their operational flexibility.

We hereby present three approaches for local flow optimization within these networks: a linear programming formulation and two greedy heuristics. Through extensive experimentation on both random and real-world inspired random network scenarios, it was found that the linear programming method generally yields superior local improvements in node-specific flow while maintaining control over global impact. However, for large-scale networks or instances with high edge densities, the Faster Greedy2 heuristic offers a compelling balance between performance and computational efficiency.

The findings of this study underscore the efficacy of SDS-based flow modeling in facilitating precise manipulation of local flows within the confines of bounded and interpretable constraints. This attribute renders it particularly well-suited for applications in educational pathways, labor market modeling, and other socioeconomic systems. The incorporation of negative weights facilitates the incorporation of realistic dynamics, such as attrition or regression in states, thereby enhancing the expressiveness of traditional flow models.

A number of promising avenues emerge for future research. Firstly, extending the model to accommodate stochastic or time-dependent edge weights has the potential to enhance realism, particularly in dynamic environments such as transportation or information networks. Secondly, the incorporation of machine learning algorithms to adaptively learn optimal edge weights from historical data has the potential to enhance the system's predictive capabilities and adaptability.

Additionally, the exploration of other update schemes or hybrid SDS models has the potential to enhance scalability and model expressiveness. Another area of interest involves the formalization of stability and convergence properties of flow dynamics for BSFN, especially in feedback-rich or cyclic networks. The potential practical utility of this methodology can be demonstrated by its application to real-world datasets, such as longitudinal educational or employment data. Such an application could also drive domain-specific innovation.

REFERENCES

- [1] J. Hackl and B. T. Adey, "Estimation of traffic flow changes using networks in networks approaches," *Applied Network Science*, vol. 4, no. 1, p. 28, 2019.

TABLE I: Minimum, average and maximum for the two measures δ_k and Δ_k

			δ_2			Δ_2		
			min	avg	max	min	avg	max
$n = 200$	$p = 0.2$	LP	2621090.37	5152700.64	8508910.71	0.00	9615.49	22671.63
		Greedy1	1195140.97	4750200.37	7694436.14	0.00	16071.90	37858.45
		Greedy2	249131.21	781855.68	1633661.57	0.00	1713.43	5039.86
	$p = 0.4$	LP	2233706.46	9096606.18	15119007.11	314.91	11148.71	37028.85
		Greedy1	2376102.21	10494505.55	6504125.29	408.99	36787.32	16679.52
		Greedy2	885100.59	2801082.14	17792735.64	1.14	6697.86	86151.74
	$p = 0.6$	LP	5465072.63	12447128.95	19132120.24	274.29	6722.21	16558.73
		Greedy1	5502528.56	18335531.33	27356699.78	439.05	64364.35	143897.45
		Greedy2	2474992.97	5625483.43	11857886.62	0.87	15242.24	37814.45

			δ_2			Δ_2		
			min	avg	max	min	avg	max
$n = 400$	$p = 0.2$	LP	9849239.34	20549663.28	25666389.85	729.96	23253.02	48495.46
		Greedy1	3744006.19	20687091.99	27944082.81	2113.98	36346.70	82402.39
		Greedy2	1522814.45	3193905.60	8906831.00	3.26	3446.25	9280.32
	$p = 0.4$	LP	8724475.30	35006775.26	47207071.83	314.54	26304.02	69638.76
		Greedy1	6788012.04	42726164.84	66155919.59	3927.77	73750.12	163657.35
		Greedy2	3963835.51	9984506.68	16366638.62	392.56	11682.35	28340.86
	$p = 0.6$	LP	5318927.62	51140901.87	75497480.76	38.78	11219.22	36393.41
		Greedy1	7516038.04	69934679.46	105264267.88	2970.70	118048.94	243847.50
		Greedy2	1209262.25	22555721.87	49223081.37	294.02	26522.69	56786.72

TABLE II: Minimum, average and maximum for the two measures δ_k and Δ_k

			δ_2			Δ_2		
			min	avg	max	min	avg	max
$n = 2000$	$p = 0.4$	LP	440188141.23	921787895.81	1255290402.09	8280.66	135184.65	255070.07
		Greedy1	376467304.46	1130567882.79	1585149455.70	23580.63	405392.55	730495.57
		Greedy2	111937948.72	279409397.33	699888936.51	3068.38	67414.42	123757.16

- [2] R. Shirzadkhani, S. Huang, A. Leung, and R. Rabbany, "Static graph approximations of dynamic contact networks for epidemic forecasting," *Scientific Reports*, vol. 14, no. 1, p. 11696, 2024.
- [3] H. Mortveit and C. Reidys, *An introduction to sequential dynamical systems*. Springer Science & Business Media, 2007.
- [4] J. Dörpinghaus, S. Schaaf, and M. Jacobs, "Soft document clustering using a novel graph covering approach," *BioData mining*, vol. 11, pp. 1–20, 2018.
- [5] M. A. Porter and J. P. Gleeson, "Dynamical systems on networks," *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, vol. 4, p. 29, 2016.
- [6] C. L. Barrett and C. M. Reidys, "Elements of a theory of computer simulation i: sequential ca over random graphs," *Applied Mathematics and Computation*, vol. 98, no. 2-3, pp. 241–259, 1999.
- [7] C. L. Barrett, H. S. Mortveit, and C. M. Reidys, "Elements of a theory of simulation ii: sequential dynamical systems," *Applied Mathematics and Computation*, vol. 107, no. 2-3, pp. 121–136, 2000.
- [8] C. Barrett, H. B. Hunt III, M. V. Marathe, S. Ravi, D. J. Rosenkrantz, and R. E. Stearns, "Modeling and analyzing social network dynamics using stochastic discrete graphical dynamical systems," *Theoretical Computer Science*, vol. 412, no. 30, pp. 3932–3946, 2011.
- [9] R. Diestel, *Graphentheorie*, 3rd ed. Berlin: Springer-Verlag, 2006.
- [10] S. O. Krumke and H. Noltemeier, *Graphentheoretische Konzepte und Algorithmen*, 2nd ed. Wiesbaden: Vieweg + Teubner, 2009.
- [11] H. Ronellenfitsch and E. Katifori, "Global optimization, local adaptation, and the role of growth in distribution networks," *Physical review letters*, vol. 117, no. 13, p. 138301, 2016.
- [12] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel, "Distributed optimization with local domains: Applications in mpc and network flows," *IEEE Transactions on Automatic Control*, vol. 60, no. 7, pp. 2004–2009, 2014.
- [13] B. Grimstad, B. Foss, R. Heddle, and M. Woodman, "Global optimization of multiphase flow networks using spline surrogate models," *Computers & Chemical Engineering*, vol. 84, pp. 237–254, 2016.
- [14] S. Scellato, L. Fortuna, M. Frasca, J. Gómez-Gardenes, and V. Latora, "Traffic optimization in transport networks based on local routing," *The European Physical Journal B*, vol. 73, pp. 303–308, 2010.
- [15] G. V. Puskorius and L. A. Feldkamp, "Neurocontrol of nonlinear dynamical systems with kalman filter trained recurrent networks," *IEEE Transactions on neural networks*, vol. 5, no. 2, pp. 279–297, 1994.
- [16] O. San, R. Maulik, and M. Ahmed, "An artificial neural network framework for reduced order modeling of transient flows," *Communications in Nonlinear Science and Numerical Simulation*, vol. 77, pp. 271–287, 2019.
- [17] J. Dörpinghaus, V. Weil, C. Düing, and M. W. Sommer, "Centrality measures in multi-layer knowledge graphs," *Annals of Computer Science and Information Systems*, 2022.
- [18] J. Dörpinghaus, V. Weil, and M. W. Sommer, "Towards modeling and analysis of longitudinal social networks," *Applied Network Science*, vol. 9, no. 1, p. 52, 2024.
- [19] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.
- [20] M. O. Jackson, *Social and Economic Networks*. Princeton: University Press, 2010.
- [21] J. Dörpinghaus, J. Binnewitt, S. Winnige, K. Hein, and K. Krüger, "Towards a german labor market ontology: Challenges and applications," *Applied Ontology*, no. 18(4), pp. 1–23, 2023.
- [22] J. Dörpinghaus and M. Tiemann, "Vocational education and training data in twitter: Making german twitter data interoperable," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 946–948, 2023.
- [23] D. Martić, A. Fischer, and J. Dörpinghaus, "Extending the german labor market ontology with online data," in *INFORMATIK 2024*. Gesellschaft für Informatik eV, 2024, pp. 2019–2030.
- [24] 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.
- [25] J. Dörpinghaus, D. Samray, and R. Helmrich, "Challenges of automated identification of access to education and training in germany," *Information*, vol. 14, no. 10, p. 524, 2023.

Ontological support for integration computer tools in digital humanities research

Iwona Grabska-Gradzińska
0000-0002-5799-5438

Institute of Applied Computer Science,
Faculty of Physics, Astronomy
and Applied Computer Science,
Jagiellonian University
ul. Łojasiewicza 11, 30-348 Kraków, Poland
Email: iwona.grabska@uj.edu.pl

Barbara Strug, Grażyna Ślusarczyk
0000-0002-2204-507X
0000-0003-1032-1644

Institute of Applied Computer Science,
Faculty of Physics, Astronomy and Applied Computer Science,
Jagiellonian University
ul. Łojasiewicza 11, 30-348 Kraków, Poland
Email: {barbara.strug,grazyna.slusarczyk}@uj.edu.pl

Abstract—As knowledge embedded in literary texts is often expressed in informal and implicit ways, in recent years the field of digital humanities has witnessed the emergence of many tools designed to support text analysis and creation of digital editions. However, these tools operate on heterogeneous, differently structured data coming from various sources. The integration of existing tools, which would allow for creating an effective system dedicated to international cooperation in the field of literary research, can be supported by the ontological approach. Therefore in this paper the concept of the ontology dedicated for reasoning about literary text properties, is presented. The formal representation of knowledge embedded in text together with its syntactic and semantic schema facilitates heterogeneous data integration and helps to bridge the semantic gap across various editorial projects.

I. INTRODUCTION

LITERARY works, especially historical texts, whether preserved in print or manuscript form, are cultural artifacts, bearing witness to the development of literary expression, editorial practice, and printing craftsmanship. They possess varying degrees of aesthetic and stylistic merit and serve as a reflection of their epochs, encapsulating the intellectual climate and the needs of its contemporary readership[1]. Moreover, such texts are repositories of a wide range of data that are relevant not only to literary studies but also to historical and archival research [2].

In recent years, the field of digital humanities has witnessed the consolidation of data and metadata description standards, alongside the emergence of a wide array of tools designed to support the creation of digital editions[3]. Among these, the Text Encoding Initiative (TEI) [4], an XML-based markup language, has become the de facto standard for encoding source texts in scholarly editions. TEI enables the detailed representation of textual structure, editorial interventions, and semantic annotation, making it particularly suitable for the complex requirements of philological and historical scholarship [5].

Complementing TEI, the Dublin Core metadata standard serves as a widely adopted framework for the description of bibliographic and cataloging metadata. Its simplicity and interoperability have made it a preferred choice in diverse

digital contexts, ranging from library systems and digital repositories to content management platforms. Importantly, Dublin Core plays a crucial role in computational ontologies and the semantic web: its metadata elements are formally defined as RDF properties, allowing for their integration into knowledge graphs, linked data infrastructures, and ontology-driven information systems. As such, Dublin Core facilitates machine-readable, semantically enriched descriptions of resources that can be processed, queried, and related across heterogeneous data environments.

However, the above mentioned tools operate on structured or unstructured data coming from different sources and organized in different ways. The integration of existing tools would allow for creating an effective system dedicated to international cooperation in the field of literary research[6]. Such integration can be supported by the ontological approach, as ontologies are well known to be an effective way of tackling the problem of interoperability among data [7].

In this paper the concept of the ontology for digital text edition, which facilitates reasoning about literary text properties, is presented. The formal representation of knowledge embedded in text together with its syntactic and semantic schema facilitates heterogeneous data integration and helps to bridge the semantic gap across various editorial projects. Moreover it gives the possibility of hiding the technical aspects of defining SPARQL queries behind an ontology-based user interface.

II. COMPUTER TOOLS IN TEXT ANALYSIS

As knowledge conveyed in historical and literary texts is often expressed in informal and implicit ways it requires readers and/or editors to engage in processes of inference, association, allusion recognition, and narrative reconstruction. There are many tools available to support such endeavours and facilitate text analysis and specialized research.

A wide range of tools has been developed to leverage the aforementioned standards, supporting various stages of the digital editorial workflow. These tools facilitate the preparation of critical editions, the indexing and semantic search of metadata, the alignment of descriptive metadata with encoded

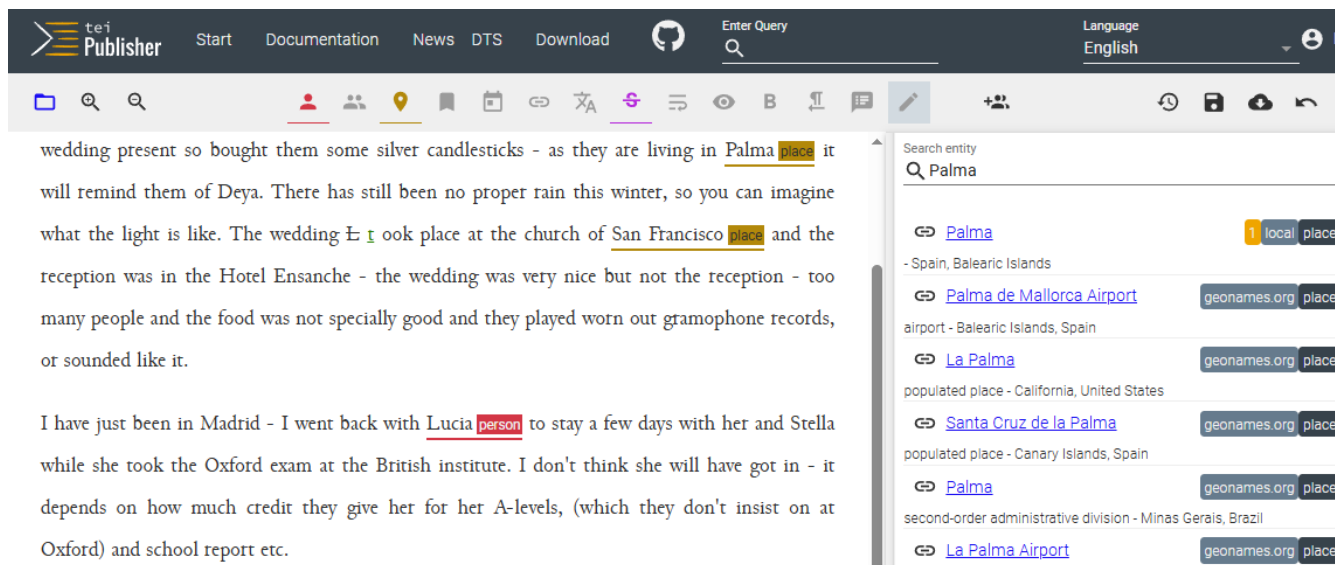


Fig. 3. *Semi-automated entity annotation in TEI Publisher.* The interface displays a TEI-encoded text undergoing semantic enrichment. The editor selects the first occurrence of an entity (e.g., a person, place, or organization), prompting the system to suggest subsequent instances for validation. On the right, external knowledge sources—such as Wikidata or GeoNames—provide contextual metadata and authoritative identifiers for the highlighted entity, aiding disambiguation and promoting interoperability with linked data frameworks.

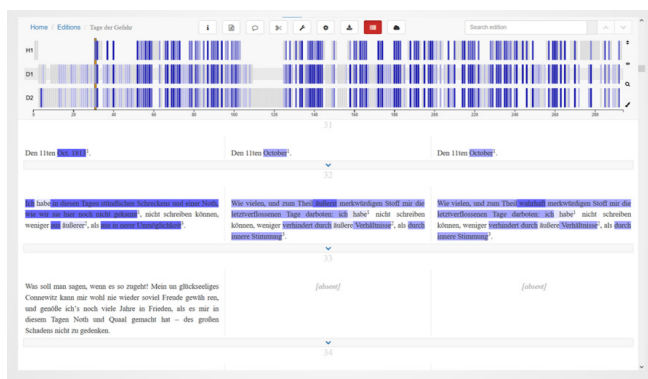


Fig. 4. LERA digital tool for inspecting similarities and differences between multiple versions of a text.

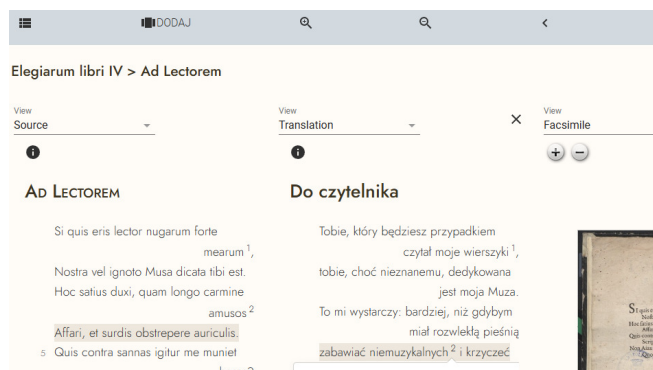


Fig. 5. TEI Publisher edition with interactive alignment of original and translated verses, along with dynamic display of footnotes.

are displayed in parallel to facilitate critical comparison, translation analysis, and reader comprehension.

Fig. 5 illustrates an example of such an edition, showcasing side-by-side rendering of an original text and its translation, structured to highlight the correspondence between semantically or syntactically related segments. This approach is particularly valuable in philological, literary, and translation studies, as it supports inquiries into translation strategies, interpretive choices, and textual equivalence.

B. Integration with external data services and metadata aggregation systems

The analysis and contextualization of texts are further enhanced by the availability of publicly accessible knowledge bases and data services that provide structured information across a wide range of domains. These include both state-

maintained resources such as bibliographic and authority databases (e.g., the Integrated Authority File in Germany [10] or the National Library of Poland's catalogs [11]) and collaboratively curated general-purpose knowledge graphs, most notably Wikidata [12]. Additionally, there are domain-specific platforms that offer structured datasets focused on particular types of information or scholarly needs, such as GeoNames [13] for geographical entities, or Trismegistos [14], which aggregates metadata related to the ancient world.

By linking digital texts to these external repositories through unique identifiers and semantic relationships, researchers can enrich their editions with authoritative context, facilitate interoperability across systems, and enable advanced forms of cross-referencing, data integration, and computational analysis. These connections are foundational to the construction of a semantic ecosystem in the digital humanities, where texts

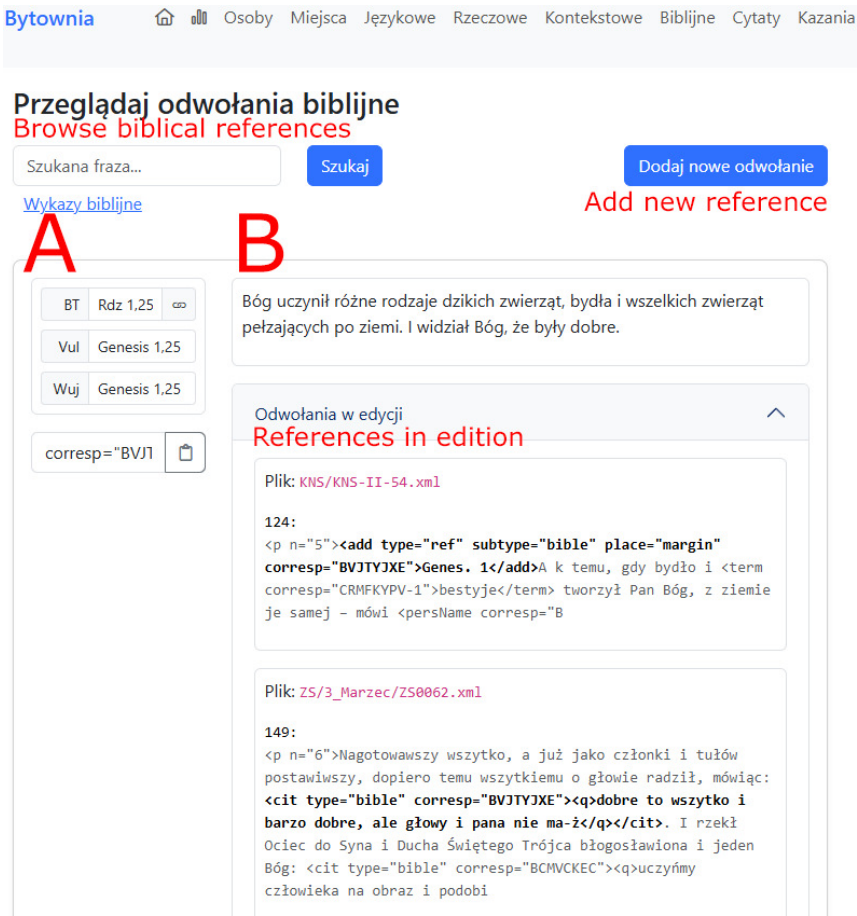


Fig. 6. Bytownia – the tool, which allows for tracing searched phrases in the text, is presented. On the left-hand side the searched biblical references are shown (A), while on the right-hand side the occurrences of these references within the analyzed text are marked (B).

are not isolated artifacts but interconnected components of a broader web of knowledge.

In Fig. 3 an additional aid for the editor in the form of the integration of external data sources (displayed on the right-hand side of the interface) can be seen. It provides encyclopedic or authority-based information about the selected entity. These linked data services enrich the annotation process by offering context from established knowledge bases, thereby supporting disambiguation and encouraging the use of persistent identifiers (e.g. Wikidata, GeoNames). This functionality not only supports the creation of richly encoded digital editions, but also fosters semantic interoperability with broader data ecosystems.

In the context of more specialized references and scholarly-oriented editions, it can be particularly beneficial to develop a service that aggregates data from publicly available external knowledge bases while incorporating detailed contextual information added during the editorial process.

Such a mechanism has been implemented in the emerging digital edition of the works of Piotr Skarga at the Jagiellonian University. The nature of Skarga’s biblical references, which

include direct quotations, marginal annotations, and allusions to biblical figures, necessitated the creation of a custom service designed to streamline the insertion and management of biblical citations. This tool enables editors to enrich the source text with canonical references (sigla) drawn from multiple biblical translations simultaneously. Moreover, it supports dynamic tracking of the function and rhetorical significance of each referenced passage within Skarga’s homiletic practice.

The interface presented in Fig. 6 allows editors to view and manage a comprehensive list of already linked biblical references, facilitating an analytical perspective on intertextuality and the exegetical structure of Skarga’s sermons. This approach exemplifies how domain-specific annotation services, integrated with external authoritative sources, can enhance both editorial precision and scholarly interpretation in digital critical editions.

This service not only enables the enrichment of textual objects by drawing data from selected external APIs (see Fig. 7), but it also exposes its own API, allowing the aggregated metadata to be reused across multiple editorial projects.

More importantly, by providing structured access to its inter-

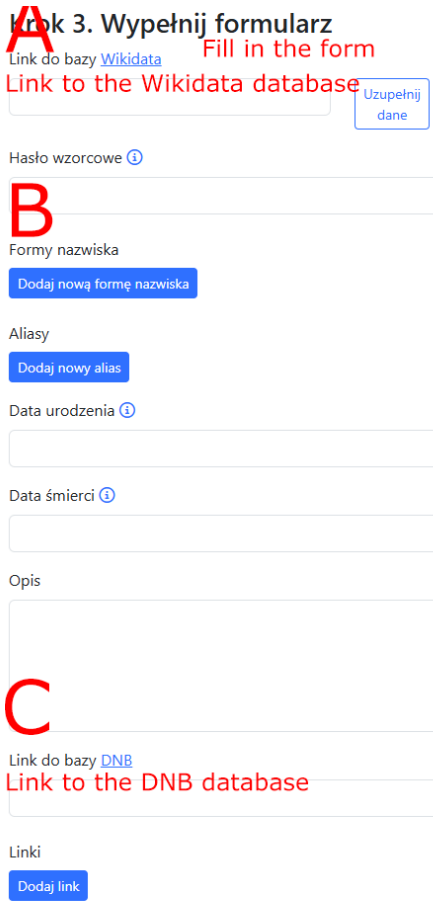


Fig. 7. Mechanism for synchronizing external knowledge services data (A, B)— enriching them with edition-specific aliases and notes (C).

nal data, the service facilitates the integration of the metadata repository with a computational ontology. This interoperability enables the implementation of reasoning tools that can operate over the entire body of accumulated knowledge. Through such semantic integration, the system supports advanced inferential functions, such as detecting implicit relationships, classifying textual phenomena, or identifying patterns in citation practices, thereby significantly enhancing the interpretive potential of the edition.

In this way, the service functions as both a dynamic editorial aid and a knowledge infrastructure, supporting not only the manual annotation process but also machine-assisted scholarly inquiry through semantic enrichment, linked data principles, and ontology-driven reasoning.

C. The heterogeneity of solutions adopted across individual editions

Print editors historically used typographic and layout conventions to signal textual structures and distinctions between content types. Contemporary digital editions take this further by employing markup systems to define the structure and semantics of texts in an unambiguous way. As a result, digital editions are not only readable artifacts but also structured repositories of knowledge organized according to standards that facilitate computational analysis and interoperability.

Knowledge encoding in textual corpora demands interpretive judgment, contextual awareness, and a nuanced understanding of textual semantics, making it inherently complex and variable. The conventions for formatting metadata and embedding semantic information within texts are themselves products of editorial practice. Over the years, as digital editions have evolved, these practices have developed organically through the cumulative experience of individual editors, resulting in a rich but highly heterogeneous ecosystem of encoding approaches. The diversity is particularly evident when comparing editions of varying genres and document types including poetry, prose, archival materials, historical records, and derivative or auxiliary texts.

Fortunately, despite this conceptual and methodological variety, the underlying data format in most digital editions is standardized through the TEI standard, based on XML, ensures a hierarchical and semantically rich data structure. This allows for a precise modeling of internal relationships within a text, preserving its logical organization while enabling efficient search, processing, and analytical operations using digital tools. The high granularity of TEI markup enhances editorial transparency and supports the comparison of variant textual versions across editions.

However, the implementation of TEI is far from monolithic. The manner in which individual elements and tags are defined and applied varies significantly between projects. These differences often stem not only from innovative or experimental approaches to the encoding process, but also from the intrinsic diversity of the source materials being edited. The needs of a critical edition of 17th-century poetry, for instance, differ considerably from those of a diplomatic transcription of

archival correspondence or a scholarly apparatus for historical charters.

As a result, while the use of TEI fosters a degree of interoperability and standardization, the full standardization of semantic structures across digital editions remains a challenging and, to some extent, aspirational goal. It requires ongoing efforts in community coordination, the development of best practices, and the possible alignment with formal ontologies and controlled vocabularies that can help bridge semantic gaps across editorial projects.

As many knowledge relationships embedded in a corpus of source texts remain opaque to the reader, it is valuable to support text analysis with tools grounded in formal representations of knowledge. Viewing a digital edition as a knowledge base accessible via APIs allows external systems to query, extract, and infer information from the encoded content. Formal knowledge representation frameworks enable the imposition of an additional conceptual layer over the edition, allowing textual content to be interpreted in terms of concepts and their interrelations. This, in turn, facilitates the application of computational ontologies and graph-based knowledge systems.

III. ONTOLOGY-BASED DATA ORGANIZATION

Ontologies can be used to unambiguously describe elements such as textual variant types, individual roles, and historical-cultural contexts. When integrated with rule-based systems or inference engines, they enable the generation of new knowledge from existing data. Ultimately, this supports the creation of shared conceptual models that facilitate the comparison and integration of diverse digital editions.

Ontology-based data organization can be described as a methodological approach for structuring and managing information using ontology, that is formal specification of concepts and their interrelationships within a specified domain. This approach offers a semantically rich alternative to traditional data management systems by enabling machine-readable representations of knowledge and facilitating better data integration, retrieval, and analysis [15]. In contrast to taxonomies or thesauri, ontologies not only define hierarchical relationships but also allow for defining complex relations, adding specific constraints, and rules that reflect domain knowledge requirements.

In recent decades, the implementation of ontology-based systems has gained acceptance in various disciplines, including biomedical, chemical or geographic information systems, and more recently, the digital humanities. The ontological approach allows for supporting the creation of metadata standards and fostering cross-disciplinary data sharing. Important examples include the CIDOC Conceptual Reference Model (CIDOC-CRM), which provides a semantic framework for cultural heritage information [16], [17], and the Europeana Data Model (EDM), which integrates diverse museum and library collections across Europe.

In digital humanities, ontology-based data organization plays a critical role as there is a need for modeling complex

cultural and historical data. These domains often involve heterogeneous sources such as manuscripts, artifacts, maps, or even recorded oral documents, which often require specialized and contextualized interpretation. Ontologies can facilitate the semantic annotation and linking of these sources, enabling potential users to query datasets using concept-based rather than keyword-based search.

Moreover, ontology-driven tools support the development of digital editions, scientific annotation platforms, and visualizations that provide researchers with help in carrying out humanities research. By enabling options such as disambiguation, provenance tracking, or reasoning from cultural data, ontologies enhance the transparency and reproducibility of research work [18]. The reusability and sustainability of use is further supported by the fact that ontological models usually follow FAIR data principles (Findable, Accessible, Interoperable, Reusable) [19], [20]. The principles focus on so called machine-actionability, which means the capacity of computational systems to find, access, interoperate, and reuse data, and do it with none or minimal user intervention. As the users more and more depend on computational support to deal with data, especially large amount of data with complex interrelations, such an approach can support data accessibility for non-technical users.

Despite its advantages, mentioned above, ontology-based data organization poses some challenges, particularly in the digital humanities context. They include the need for domain-specific modeling expertise, and the problems of understanding between domain experts and data specialists. Another issue come from the possible conflict between formal logic, required by data modeling, and interpretive flexibility common in humanities research. In addition there is a problem of accepting common shared vocabularies among scholars with diverse theoretical backgrounds. Nevertheless, ongoing collaborative efforts and community-driven ontology development, such as the TEI [20] and the OntoMedia [21] ontology, continue to bridge these gaps. Thus ontology-based data organization presents an approach to managing and analyzing complex data in the digital humanities that can open new possibilities in data sharing and using. By adding rich semantic structures to raw data, such an approach can empower researchers to find new insights by following linked data and advance collaborative research by opening access to previously inaccessible sources.

IV. ONTOLOGY FOR DIGITAL TEXT EDITION

Integration of data and metadata related to literary texts coming from different, often big and heterogeneous sources, like library or museum data bases and digital archived, requires harmonisation into a unified view. Such integration is facilitated by developing a domain-specific ontology.

Many methodologies for building ontologies have been proposed [22], [23]. Modern agile methodologies like eXtreme Design (XD) [24], Modular Ontology modeling (MOM) [25], Simplified Agile Methodology for Ontology Development (SAMOD) [26], are typically based on ontology requirements

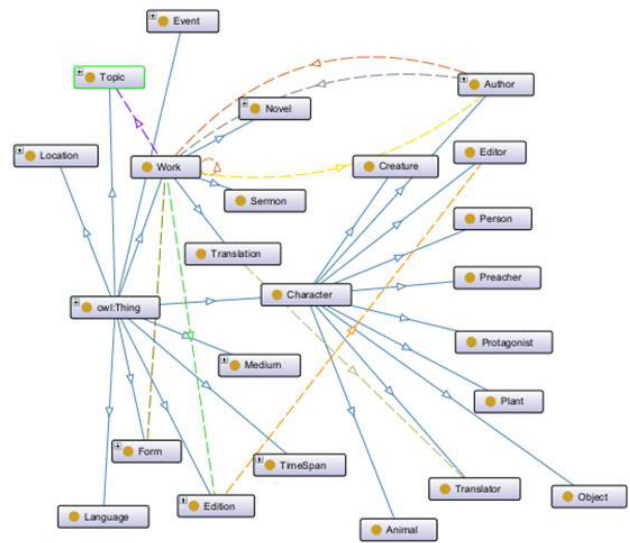


Fig. 8. A fragment of TEON ontology

or ontology design patterns. A common way to express ontology requirements, which specify the intended task of the resulting ontology, is through Competency Questions (CQ) [27] illustrating the typical information needs that one would require the ontology to respond to. These questions can be then formalized as queries over some test set of data A_{in} in order to test the created ontology [28].

Our prototype ontology has been built upon a set of competency questions developed by philologist experts. An example of a competency question is "What literary motifs are discussed in the positivist novel?". These questions refer to data related to literary texts and constrain the scope of knowledge to be represented in an ontology. On this basis, classes, properties and relations of the Text Encoding Ontology (TEON) were defined. Such questions can be then used to test if the ontology contains enough information to answer them.

The most important classes available in the current version of the ontology are *Work*, *Character*, *Style*, *Topic*, *Location*, *TimeSpan*, *Edition*, *Form*, *Medium*, *Institution*, *Event*, *Language* and *Gender*. Class *Work* contains data related to a literary text, and has subclasses specifying work genre as *Novel*, *Letter*, *Essay*, *Poem*, *Sermon*, *Diary* etc. Class *Character* contains subclasses *Author*, *Editor*, *Person*, *Translator*, *Protagonist*, *Preacher*, *Animal*, *Plant*, *Creature* and *Object*, which represent possible variants of characters related to works. For example the class *Author* provide access to information about the person who is the author of at least one work, while the class *Creature* describes a fantastic creature appearing in a literary work. In Fig. 8 a snapshot of the fragment of the ontology is depicted.

We assume that the input data concerning literary text are stored in the form of the csv files. A fragment of the file storing data about literature of positivism, converted to .xls format for clarity of presentation, is shown in Fig. 9. In this figure several

E2									
powieść									
	A	B	C	D	E	F	G	H	I
1	Nr	Tytuł	Autor	Rok wydania	Typ	Styl	Miejsce akcji	Czas akcji	Motywy
2	1.	Nad Niemnem	Eliza Orzeszkowa	1888	powieść	pozytywizm	Korczyn,	VI-VIII 1886	Powstanie styczniowe,
3							Bohaterowicze		mezalians
4	2.	Lalka	Bolesław Prus	1890	powieść	pozytywizm	Warszawa,	1878-1879	Powstanie styczniowe,
5							Zasławek		asymilacja żydów

Fig. 9. A fragment of the csv test file containing information on works in the positivism style

entries related to work title, author, year of issue, type, style, place of action, time of action and motives can be seen.

These data are transformed into RDF triples using R2RML (Relational to Resource Description Framework Mapping Language) language [29], [30]. This language provides capabilities to connect the structure of the csv data to the ontology vocabulary. The obtained triples become instances which populate the ontology and form a data graph also referred to as a RDF triplestore. This store is then used as a data sources for SPARQL queries that support implementation of the competency questions. In Fig. 10 a data graph with individual instances coming from the entries shown in Fig. 9.

The considered triplestore supports SPARQL queries related to data concerning literary texts. In Fig. 11 a query related to authors who refer to the January Uprising (pol. Powstanie Styczniowe) in their works is presented. The result obtained using this query contains two author names: Eliza Orzeszkowa

and Bolesław Prus.

V. CONCLUSION

This paper presents a current version of the ontology that will provide a semantic schema for the computer system making different resources accessible for researchers in different areas of humanities research.

The described methodology is a part of an ongoing research within the framework of the Jagiellonian University Flagship Programmes "Digital Humanities Lab" and "European Heritage in the Jagiellonian Library: Digital Authoring of the Berlin Collections". It aims at providing a shared ontology-based tool supporting the creation of a knowledge graph storing information from existing datasets in different, and usually incompatible and non-interoperable data formats, scattered among different places. It will also offer a unified way of querying available literary resources and their metadata.

REFERENCES

- [1] J. Gruchała, "Editing: a Knowledge and a Skill" (in polish). *Wielogłos*, no. 3 (13), 2012, pp. 157–164. DOI: 10.4467/2084395XW12.012.0867. ISSN 1897-1962.
- [2] J. Gruchała, "The Virtual Publisher and the User of a Digital Edition" (in polish). In: Elżbieta Wichrowska (ed.), *The European Literary Canon*, pp. 282–288. ISBN 978-83-235-0834-2, 2012.
- [3] E.F. Cavanaugh and J.E. Stertz, "Building Accessibility: Platforms and Methods for the Development of Digital Editions and Projects", in *Digital Editing and Publishing in the Twenty-First Century*, eds. J. O'Sullivan, M. Pidd, S. Whittle, B. Wessels, M. Kurzmeier, and Ó. Murphy, Scottish Universities Press, 2025. DOI: 10.62637/sup.GHST9020.5. ISBN 978-1-917341-04-2.
- [4] TEI Consortium. "TEI P5: Guidelines for Electronic Text Encoding and Interchange". Version 4.9.0, last updated 24 January 2025. <https://www.tei-c.org/Guidelines/P5/>
- [5] E. Spadini and J.L. Palenzuela, "Re-using Data from Editions", in *Digital Editing and Publishing in the Twenty-First Century*, eds. J. O'Sullivan, M. Pidd, S. Whittle, B. Wessels, M. Kurzmeier, and Ó. Murphy, Scottish Universities Press, 2025. DOI: 10.62637/sup.GHST9020.8. ISBN 978-1-917341-04-2.
- [6] G. Franzini, M. Kestemont, G. Rotari, M.Jander, J. Ochab, E. Franzini, J. Byszuk, and J. Rybicki. "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm". *Frontiers in Digital Humanities*, vol. 5, 2018. DOI: 10.3389/fdigh.2018.00004.
- [7] R. Kishore, R. Sharman and R. Ramesh (2004). "Computational Ontologies and Information Systems I: Foundations", *Communications of the Association for Information Systems*. 14. 158-183. 10.17705/ICAIS.01408.
- [8] TEI Publisher. "An environment for publishing TEI documents". Version 9.0.9, maintained by the eXist Solutions team, 2025. <https://teipublisher.com>

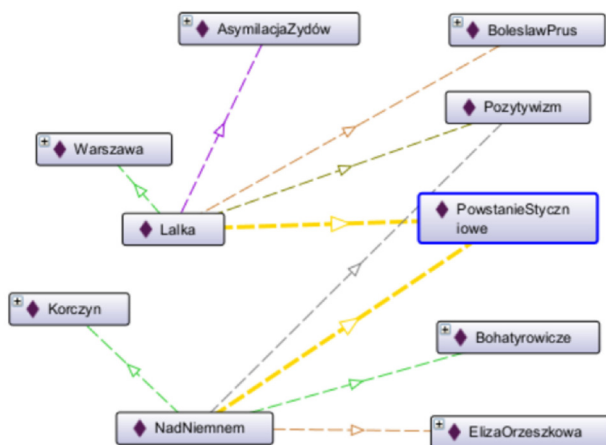


Fig. 10. A data graph with selected instances

```

PREFIX uj:
<http://www.semanticweb.org/al2020/ontologies/2025/2/ontoDigit
alHum#>

SELECT ?autor WHERE {
  ?dzieto uj:refersToEvent uj:PowstanieStyczn iowe .
  ?dzieto uj:hasAuthor ?autor .
}

```

Fig. 11. An example of a sparql query

- [9] M. Pöckelmann, A. Medek, J. Ritter, and P. Molitor, "LERA - An interactive platform for synoptical representations of multiple text witnesses" In: *Digital Scholarship in the Humanities (DSH)*. Oxford University Press 2022. DOI: 10.1093/llc/fqac021
- [10] German National Library. *Gemeinsame Normdatei (GND)*. Available at: <https://www.dnb.de/EN/gnd> [Accessed: 20.05.2025].
- [11] Biblioteka Narodowa. *National Library of Poland's catalogues*. Available at: <https://katalogi.bn.org.pl/> [Accessed: 20.05.2025].
- [12] Wikimedia Foundation. *Wikidata: A free knowledge base that anyone can edit*. Available at: <https://www.wikidata.org/> [Accessed: 20.05.2025].
- [13] GeoNames. *GeoNames geographical database*. Available at: <https://www.geonames.org/> [Accessed: 20.05.2025].
- [14] Trismegistos. *Trismegistos: An interdisciplinary portal of the ancient world*. Available at: <https://www.trismegistos.org/> [Accessed: 20.05.2025].
- [15] T.R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, 5, pp 199–220, 1993. Doi: 10.1006/knac.1993.100
- [16] M. Doerr, "The CIDOC Conceptual Reference Model: An Ontological Approach to Semantic Interoperability of Metadata," *AI Magazine*, 24, pp 75–92, 2003. DOI: 10.1609/aimag.v24i3.1720
- [17] CIDOC-CIDOC "Conceptual Reference Mode", <https://cidoc-crm.org/>
- [18] S. Schreibman, R. Siemens, and J. Unsworth, (Eds.), *A New Companion to Digital Humanities*, John Wiley and Sons; 2016. DOI:10.1002/9781118680605
- [19] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., "The FAIR Guiding Principles for scientific data management and stewardship", *Sci Data* 3, 160018, 2016. DOI: 10.1038/sdata.2016.18
- [20] J. Tello, M. Göbel, U. Veentjer, S. Funk, N. Rißler-Pipka, and K. Du, . "FAIR Derived Data in TEI and Its Publication in the TextGrid Repository," *Journal of the Text Encoding Initiative* 18, 2024 DOI: 10.4000/13llz.
- [21] F. Lawrence, M. Tuffield, M. Jewell, A. Prugel-Bennett, D. Millard, M. Nixon, M.C. Schraefel, and N. Shadbolt, "OntoMedia - Creating an Ontology for Marking Up the Contents of Heterogeneous Media", *Ontology Patterns for the Semantic Web ISWC-05 Workshop*, Galway, Ireland, 2005.
- [22] N. Noy and D. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," *Knowledge Systems Laboratory*, 32, 2001.
- [23] H.S. Pinto, C. Tempich and S. Staab, "Ontology Engineering and Evolution in a Distributed World Using DILIGENT," In: *Staab, S., Studer, R. (eds) Handbook on Ontologies. International Handbooks on Information Systems*, Springer, Berlin, Heidelberg, pp. 153–176, 2009. DOI:10.1007/978-3-540-92673-3_7
- [24] E. Blomqvist, K. Hammar and V. Presutti, "Engineering ontologies with patterns— the eXtreme design methodology," In: Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, and Valentina Presutti, Eds., *Ontology Engineering with Ontology Design Patterns*, vol. 25 of *Studies on the Semantic Web*. IOS Press, 2016. DOI: 10.3233/978-1-61499-676-7-23
- [25] P. Hitzler and A. Krisnadhi "A Tutorial on Modular Ontology Modeling with Ontology Design Patterns: The Cooking Recipes Ontology," *CoRR*, 2018. DOI:10.48550/arXiv.1808.08433
- [26] S. Peroni, "A Simplified Agile Methodology for Ontology Development," In: *Dragoni, M., Poveda-Villalón, M., Jimenez-Ruiz, E. (eds) OWL: Experiences and Directions – Reasoner Evaluation. OWLED ORE 2016*, Lecture Notes in Computer Science, vol. 10161, 2017. Doi:10.1007/978-3-319-54627-8
- [27] M. Grüninger and M.S. Fox, "The role of competency questions in enterprise engineering," In *Asbjorn Rølstadas, Ed., Benchmarking—Theory and Practice*, Springer, pp. 22–31, 1995. DOI:10.1007/978-0-387-34847-6_3
- [28] C.M. Keet and A. Ławrynowicz, "Test-driven development of ontologies," In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, Eds., *The Semantic Web. Latest Advances and New Domains—13th International Conference, ESWC, Heraklion, Crete, LNCS*, vol. 9678, pp. 642–657, 2016, DOI: 10.48550/arXiv.1512.06211
- [29] J. F. Sequeda, "On the Semantics of R2RML and its Relationship with the Direct Mapping," In *ISWC (Posters and Demos)*, pp. 193–196, 2013.,
- [30] M. Rodriguez-Muro and M. Rezk, "Efficient SPARQL-to-SQL with R2RML mappings," *Journal of Web Semantics*, 33, pp. 141–169, 2015, DOI: 10.1016/j.websem.2015.03.001.

3D Brain Extraction from Magnetic Resonance Imaging Using Knowledge Distillation

Kali Gurkahraman
0000-0002-0697-125X

Department of Computer Engineering
Sivas Cumhuriyet University
Sivas, Turkey
kgurkahraman@cumhuriyet.edu.tr

Ahmet Firat Yelkuvan
0000-0003-4148-1923

Department of Computer Engineering
Sivas Cumhuriyet University
Sivas, Turkey
aftyelkuvan@cumhuriyet.edu.tr

Rukiye Karakis
0000-0002-1797-3461

Department of Software Engineering
Sivas Cumhuriyet University
Sivas, Turkey
rkarakis@cumhuriyet.edu.tr

Abstract—Brain extraction, or skull stripping, is a crucial preprocessing step in magnetic resonance imaging (MRI), isolating brain tissue from surrounding structures like the skull and scalp. However, existing methods have limitations, such as parameter sensitivity in traditional approaches and computational complexity in advanced deep learning architectures. This study proposes a knowledge distillation framework utilizing two UNet++ models—a high-capacity teacher network and an efficient student network—for 3D brain extraction tasks. The teacher network generates detailed grayscale brain predictions, capturing subtle intensity transitions and anatomical boundaries. The student network learns to produce precise binary segmentation masks from the teacher’s feature representations, guided by a hybrid loss function combining Dice, Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE). Evaluations conducted on T1-weighted, T2-weighted, and proton-density weighted MRI images from the IXI dataset demonstrated the student model’s superior performance, achieving a Dice coefficient of 0.97857. These findings suggest that the proposed framework may offer a practical and accurate solution for brain extraction in diverse medical imaging scenarios.

Index Terms—brain extraction, deep learning, hybrid loss, knowledge distillation, UNet++

I. INTRODUCTION

BRAIN extraction (or skull stripping) is a key Magnetic Resonance Imaging (MRI) preprocessing step that isolates brain tissue from non-brain structures like the skull and scalp. Its accuracy significantly impacts downstream tasks such as volumetric analysis, image registration, lesion segmentation, anatomical delineation, cortical thickness estimation, motor function prediction, and neurosurgical planning [1-3].

Brain extraction enhances the accuracy of downstream analyses by isolating the brain region, thereby eliminating extraneous signals and noise from surrounding non-brain tissues. This results in more precise and consistent outcomes. Although manual segmentation remains the gold standard, it is labor-intensive, time-consuming, and subject to inter-operator variability, affecting reproducibility [4]. Inadequate brain extraction, if not manually corrected, can introduce significant errors in further neuroimaging analyses. To overcome these challenges, many (semi-)automated brain extraction methods have been proposed and refined in recent years.

In the literature, many methods have been proposed to

separate the brain—considered the region of interest—from non-brain tissues. These include basic image processing steps like erosion, dilation, thresholding, and edge detection [4–7]. One well-known approach is the Brain Surface Extraction (BSE) method developed by Shattuck et al. [6], which is used in the BrainSuite software [7]. BSE combines anisotropic diffusion filtering, edge detection, and morphological operations to extract the brain. However, these types of methods often need manual tuning of parameters for each image, which makes them less practical and more time-consuming for large datasets.

To improve brain extraction, deformable models such as active contours (snakes) and level sets have been used [4]. These methods iteratively adjust a shape to match the brain’s boundaries by minimizing energy. A well-known example is the Brain Extraction Tool (BET) [8] and its improved version BET2 [9], both part of the FMRIB Software Library (FSL) [10]. They begin with a spherical model at the head’s center and deform it to fit the brain. These models handle intensity variations well and produce smooth results, but their accuracy depends on initial settings and may be limited by irregular brain shapes or pathologies [4].

Another widely used approach involves atlas-based methods, which use pre-segmented brain atlases to guide the segmentation of new MRI scans. These approaches typically involve registering the atlas to the subject’s image and transferring labels to identify brain regions. For example, Dale et al. [11] introduced a skull-stripping method within the FreeSurfer framework [12], which normalizes image intensities and deforms a tessellated ellipsoidal template to fit the inner skull surface. Another well-known method, BEaST [13], performs patch-based segmentation with linear registration to the ICBM152 template using multi-resolution images from both healthy and Alzheimer’s patients. Although atlas-based methods can achieve high accuracy—especially when the atlas closely represents the target population—their performance strongly depends on the accuracy of image registration and the quality of the chosen template.

In recent years, hybrid methods have been proposed to build on the strengths of atlas-based and other traditional techniques. These methods aim to improve both robustness and accuracy

by combining multiple approaches. For instance, thresholding can be enhanced with machine learning, or deformable models can be integrated with atlas-based strategies. Souza et al. [14] used eight segmentation methods to generate brain masks, which were then fused using the STAPLE algorithm [15] to produce a consensus result. Another example, the Hybrid Watershed Algorithm (HWA) [16], combines watershed segmentation with deformable surface modeling based solely on intensity information. While hybrid approaches often yield better results than individual methods, their performance still depends on factors such as parameter tuning, atlas/template selection, and registration quality.

Artificial neural networks (ANNs), especially deep learning (DL) models, have shown strong performance in medical image analysis tasks [3, 17-18]. In brain extraction, several DL-based methods have been developed to improve segmentation accuracy. One of the earliest was proposed by Kleesiek et al. [19], who introduced a convolutional neural network (CNN) for brain extraction. Their model achieved competitive Dice scores on T1-weighted, T2-weighted, and FLAIR MRI scans and showed better specificity compared to traditional methods. Building on early CNN-based work, researchers have adapted 3D-UNet architectures from their 2D versions, using encoder-decoder blocks to improve performance [4, 19-23]. Hwang et al. [20] applied a modified 3D-UNet to T1w MRIs, and Isensee et al. [4] introduced HD-BET, which performed well across different MRI sequences and scanners. Zhang et al. [21] proposed FRNET with residual connections and a boundary loss function, showing strong results on infant MRIs, though it has not been tested on adult data. Other studies have also enhanced 3D-UNet models using residual features [22], or by combining real and synthetic images through GANs, as seen in the work by Hoopes et al. [23].

The UNet model has been adapted in many ways to handle complex medical imaging tasks better, especially those involving MRI. In recent studies, several ensemble and cascaded versions of UNet have been developed to improve segmentation accuracy. Cascaded UNet architectures, in particular, have shown strong performance in capturing complex anatomical structures and resolving ambiguous boundaries. These models often work in multiple stages, where an initial network produces a rough prediction and a second network refines it. This setup leads to more stable and detailed results, especially around brain edges [24-26].

However, despite their benefits, cascaded UNet models can be computationally expensive and may overfit when trained on limited or highly variable datasets. To address these issues, this study proposes a knowledge distillation framework where a high-capacity teacher model guides a lightweight student model for 3D brain extraction from MRI scans. The teacher network produces a detailed grayscale brain prediction, which helps the student model learn to generate an accurate binary brain mask. Both models are designed for volumetric data, and a hybrid loss function—combining Dice, structural similarity index measure (SSIM), and mean squared error (MSE)—is used to improve boundary accuracy while handling

class imbalance. This approach achieves strong segmentation performance on the tested dataset and shows potential for broader application in similar medical imaging tasks.

II. METHODOLOGY

This study proposes a knowledge distillation framework that integrates two UNet++ models, each trained with a distinct objective to balance anatomical precision and segmentation efficiency. The teacher model is trained to perform grayscale brain extraction using input from three common MRI modalities: T1-weighted (T1w), which highlights anatomical structure; T2-weighted (T2w), which is sensitive to fluid and pathology; and proton density-weighted (PDw) imaging, which emphasizes tissue contrast based on hydrogen concentration. These scans were obtained from the IXI dataset and selected to ensure diversity in anatomical and contrast information. This design enables the teacher to capture fine-grained intensity patterns and structural boundaries. In contrast, the student model receives the same MRI input but is trained to produce a binary brain mask. By learning from the feature representations of the teacher, the student model acquires anatomical awareness while remaining optimized for efficient binary segmentation.

The proposed framework incorporates a modified DL architecture (UNet++), a hybrid loss function combining Dice, SSIM, and MSE, and a knowledge transfer strategy to enhance generalization. Details of the model structure, the training objectives for both teacher and student networks, the knowledge distillation process, and the evaluation metrics used for performance assessment are described in the following sections.

A. UNET++

In this study, the UNet++ architecture, shown in Figure 1, was employed for 3D brain extraction from MRI. UNet++ was selected due to its enhanced capability to capture fine structural details and improve segmentation accuracy—particularly at object boundaries—through its nested and densely connected design. As an advanced variant of the original U-Net, UNet++ replaces simple skip connections with intermediate convolutional blocks that help reduce the semantic gap between encoder and decoder features. In the architecture diagram, solid black arrows indicate down-sampling in the encoder, while dashed black arrows represent up-sampling in the decoder. Dashed green arrows show lateral dense skip connections at the same resolution, and dashed blue arrows highlight up-sampled features passed into intermediate convolution blocks. The light blue modules represent these nested blocks, which contribute to progressive feature refinement. Additionally, the architecture supports deep supervision at multiple decoder depths, facilitating efficient training and making it well-suited for complex volumetric segmentation tasks such as brain extraction [27].

B. Knowledge Distillation

Knowledge distillation (KD) is a technique where a larger teacher model guides a smaller student model by transferring

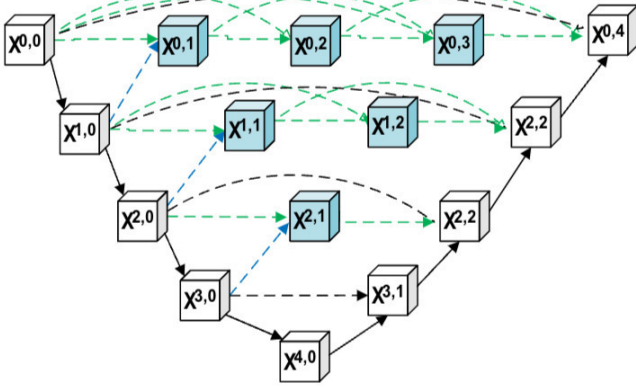


Fig. 1. UNet++ is structured with an encoder and decoder linked by a series of nested, densely connected convolutional blocks.

learned representations. Instead of learning only from ground truth labels, the student also learns from the teacher's soft outputs, capturing richer structural information. In this study, the teacher produces grayscale brain predictions, and the student learns to generate binary brain masks, supervised by both segmentation and distillation losses.

To transfer knowledge from the teacher network to the student network in a medical image segmentation context, we employed a custom distillation loss formulation that integrates both task-specific segmentation loss and feature-level guidance from the teacher model[28].

During training, the student model is supervised by two objectives:

$$L_{total} = (1 - \alpha) \times L_{seg} + \alpha \times L_{distill} \quad (1)$$

Where L_{seg} is the segmentation loss computed between the predicted mask from the student and the ground truth binary mask. This ensures that the student learns the final segmentation task correctly.

$L_{distill}$ is the distillation loss, calculated between the student's predicted mask and the gray-level soft prediction output of the teacher. This guides the student to imitate the spatial structure and internal representation captured by the teacher.

$\alpha \in [0, 1]$ is a weighting coefficient that balances the contribution of the segmentation loss and the distillation loss. In this study α , which is set as 0.3, emphasizes the importance of teacher guidance during training.

C. Loss Function

This study employed a hybrid loss denoted as L_{seg} uses a hybrid loss function combining MSE, SSIM loss, and Dice loss, while $L_{distill}$ typically employs MSE to align the student's output with the soft gray-level guidance provided by the teacher network. The overall hybrid loss function is as in (2).

$$L_{seg} = L_{mse} + L_{ssim} + L_{dice} \quad (2)$$

L_{mse} is used to penalize pixel-wise intensity differences and ensures that the predicted output closely matches the ground truth in terms of raw voxel intensities.

L_{ssim} captures perceptual differences by focusing on luminance, contrast, and structure, thus preserving anatomical consistency in the predicted images.

L_{dice} promotes spatial alignment between binary structures in the ground truth and prediction, which is crucial for accurate segmentation performance.

L_{ssim} and L_{dice} losses are calculated as in (3) and (4).

$$L_{ssim} = 1 - SSIM(P, R) \quad (3)$$

$$L_{dice} = 1 - DICE(P, R) \quad (4)$$

SSIM and DICE between predicted (P) and real (R) images are computed as in (5) and (6).

$$SSIM(P, R) = \frac{(2\mu_P\mu_R + C_1)(2\sigma_{PR} + C_2)}{(\mu_P^2 + \mu_R^2 + C_1)(\sigma_P^2 + \sigma_R^2 + C_2)} \quad (5)$$

$$DICE(P, R) = \frac{2|P \cap R|}{|P| + |R|} \quad (6)$$

The proposed KD framework employs two UNet++ models with distinct training objectives tailored to optimize both anatomical precision and segmentation efficiency. The teacher model is trained to learn a grayscale brain extraction task, where the input is the original T1w, T2w, or PDw MRI image, and the output is a grayscale brain-only image. This approach encourages the teacher network to capture subtle intensity transitions and detailed structural boundaries of the brain tissue. On the other hand, the student model is trained using the same original brain image as input, but its target output is a binary brain mask that delineates the brain region. By learning from the teacher's feature representations via knowledge distillation, the student model gains anatomical awareness while being optimized for efficient binary segmentation.

D. Dataset

The IXI dataset includes multiple MRI modalities. In this study, T1w, T2w, and proton PDw MRI images from the IXI dataset [23] were selected to ensure sufficient anatomical detail and to introduce modality diversity in the brain extraction experiments (Table 1).

TABLE I
SUMMARY OF IXI DATASET MODALITIES, Voxel SIZE, AND DATASET

Modality	Voxel Size (mm ³)	Images
T1w MRI	0.9×0.9×1.2	50
T2w MRI	0.9×0.9×1.2	50
PDw MRI	0.9×0.9×1.2	50

E. Performance Metrics

The segmentation performance of the UNet++ model was assessed using three standard metrics: Dice coefficient, sensitivity, and specificity. As shown in equation (6), the Dice coefficient is computed as the ratio of twice the overlap between the predicted and ground truth masks to the total area covered by both masks.

Sensitivity, also known as recall, quantifies the model's ability to correctly identify brain tissue within the segmentation, as defined in equation (8). In contrast, specificity measures the effectiveness of the model in correctly excluding non-brain tissue and is computed using the formulation provided in equation (9).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed DL model was implemented using the Keras library (Python 3.9) with TensorFlow. Experiments were conducted on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB), an Intel i9-12900KS processor (3.40 GHz), and 64 GB of RAM. The model was trained using the Adam optimizer with dropout and L2 regularization to reduce overfitting. Key training parameters were set as follows: a learning rate of 0.001, momentum of 0.8, and weight decay of 0.00001. Based on empirical testing, training was performed for 100 epochs with a batch size of 4 to accommodate the memory demands of 3D volumes. An 80/20 train-test split was used, and model performance was evaluated using five-fold repeated random subsampling validation (RSV).

Figure 2 presents representative axial, coronal, and sagittal slices from a T1w MRI scan in the IXI dataset. Figure 2a shows the original input image, while Figure 2b displays the corresponding grayscale brain image used as ground truth for training the teacher model. Figure 2c illustrates the output predicted by the UNet++ teacher network. The predicted brain images show strong visual similarity to the reference images, preserving anatomical structures and intensity gradients across views. Quantitatively, the teacher model achieved a Mean Absolute Error (MAE) of 0.0249, a PSNR of 51.46 decibel (dB), and an SSIM of 0.9490, demonstrating high reconstruction accuracy and perceptual quality. These results confirm the effectiveness of the proposed approach in generating anatomically faithful grayscale brain extractions.

Figure 3 shows representative binary segmentation results from axial, coronal, and sagittal MRI slices. Figure 3a displays the ground truth binary brain masks used as references. Figure 3b illustrates predictions generated by a standard UNet++ model trained without knowledge distillation, while Figure 3c presents outputs from the UNet++ student model trained with knowledge distillation. The student model predictions

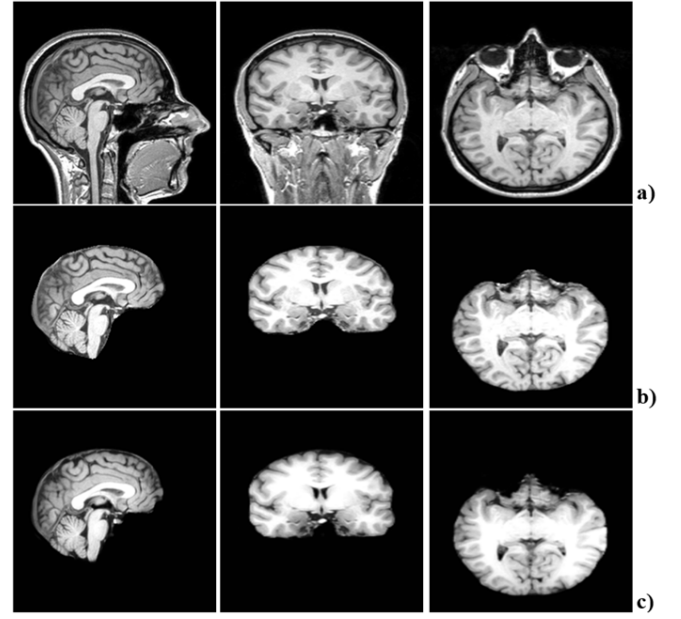


Fig. 2. Sagittal, coronal, and axial slices from a T1w MRI image in the IXI dataset. (a) Original 3D input volume, (b) ground truth grayscale brain image, (c) grayscale brain output predicted by the UNet++ teacher model.

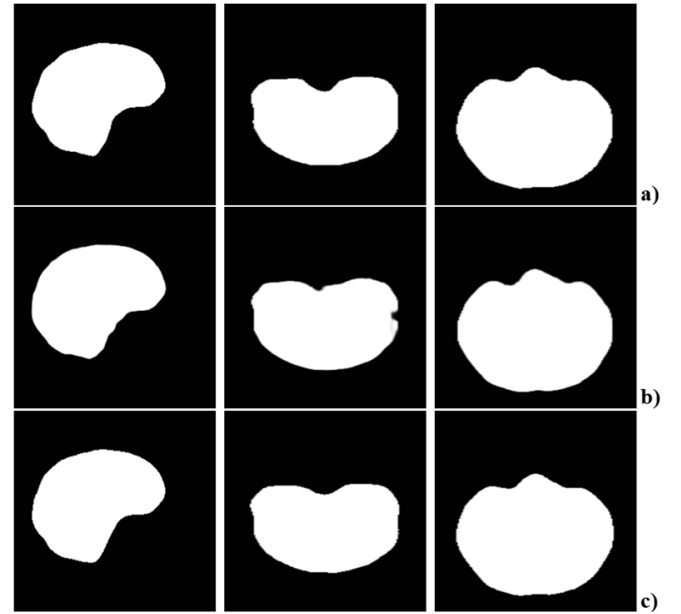


Fig. 3. Binary brain masks from sagittal, coronal, and axial views: (a) Ground truth brain masks, (b) UNet++ model predictions without knowledge distillation, (c) UNet++ student model predictions trained via knowledge distillation.

(Figure 3c) closely match the ground truth masks, demonstrating smoother boundaries and fewer segmentation inaccuracies compared to the standard UNet++ outputs (Figure 3b).

Table 2 summarizes these observations quantitatively by comparing the segmentation performance of the standard UNet++ model, the proposed UNet++ student model, and a

TABLE II
SEGMENTATION PERFORMANCE COMPARISON OF UNET++ MODELS.

DL Model	Dice	Sensitivity	Specificity
UNet++	0.95997	0.95766	0.99827
UNet++ Student	0.97857	0.97478	0.99899
Hoopes et al.[23]	0.96700	-	-

recent method in terms of Dice, sensitivity, and specificity metrics [23]. The UNet++ student model achieved superior performance, with a Dice score of 0.97857, sensitivity of 0.97478, and specificity of 0.99899. These qualitative and quantitative results collectively confirm that the knowledge distillation approach significantly enhances segmentation accuracy, especially in improving boundary delineation and detection sensitivity for 3D brain extraction tasks.

This study demonstrates strong segmentation performance using a knowledge distillation framework with UNet++ models trained on T1w, T2w, and PDw MRI from the IXI dataset. Accurate brain extraction is critical in MRI, enabling both automated analysis and clinical interpretation to focus on relevant regions, particularly in noisy images or when detecting subtle lesions. Unlike traditional methods that require manual parameter tuning and are sensitive to variability, the proposed approach is robust and parameter-free. Knowledge distillation further allows a lightweight student model to achieve high accuracy at a lower computational cost, making it well-suited for practical use. Nevertheless, certain limitations should be considered, as discussed below.

Although multiple MRI contrasts were utilized, experiments were confined to a single dataset; therefore, further validation across different datasets is necessary to establish broader applicability and robustness. Additionally, the relatively large 3D input dimensions (256×288×288) used in this study might present challenges in terms of computational resource demands and practical deployment, particularly in clinical scenarios. The relatively small size of the dataset may also limit the statistical power and generalizability of the results. Future studies may address this by augmenting the dataset, for example, through synthetic image generation or advanced data augmentation techniques, to improve robustness. Furthermore, the proposed framework is inherently flexible and could be adapted to other MRI sequences, such as diffusion-weighted or FLAIR imaging, as well as to different neuroimaging modalities, including PET and functional MRI, by adjusting to their specific characteristics and spatial resolutions (e.g., 1 mm or 2 mm isotropic), thereby extending its clinical applicability. Nevertheless, given the modular nature of the proposed knowledge distillation framework, it is reasonable to suggest that extending this approach to other UNet-based architectures may also yield improved segmentation performance.

IV. CONCLUSION

In this study, we proposed a knowledge distillation-based DL model using UNet++ models for 3D brain extraction from MRI scans. Our results demonstrate that knowledge distillation effectively enhances segmentation accuracy, particularly at

brain boundaries, by transferring detailed anatomical knowledge from a high-capacity teacher model to a more efficient student model. The student model achieved superior Dice, sensitivity, and specificity scores compared to both a standard UNet++ model and recent literature results. These findings highlight the capability of knowledge distillation to improve segmentation performance while maintaining computational efficiency.

Despite the promising outcomes, further validation using additional MRI datasets is necessary to confirm the generalizability of the proposed approach. Future studies could explore integrating attention mechanisms or evaluating other UNet-based architectures within the proposed knowledge distillation framework to further enhance segmentation performance.

ACKNOWLEDGMENT

This work is supported by the Scientific Research Project Fund of Sivas Cumhuriyet University.

REFERENCES

- [1] P. Kalavathi and V. S. Prasath, "Methods on skull stripping of MRI head scan images—a review," *J. Digit. Imaging*, vol. 29, pp. 365–379, 2016.
- [2] R. de Boer *et al.*, "Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods," *NeuroImage*, vol. 51, no. 3, pp. 1047–1056, 2010.
- [3] R. Karakis, K. Gurkahrman, G. D. Mitsis, and M. H. Boudrias, "Deep learning prediction of motor performance in stroke individuals using neuroimaging data," *J. Biomed. Inform.*, vol. 141, Art. no. 104357, 2023.
- [4] F. Isensee *et al.*, "Automated brain extraction of multisequence MRI using artificial neural networks," *Hum. Brain Mapp.*, vol. 40, no. 17, pp. 4952–4964, 2019.
- [5] M. E. Brummer *et al.*, "Automatic detection of brain contours in MRI datasets," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 153–166, 1993.
- [6] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, 2001.
- [7] D. W. Shattuck and R. M. Leahy, "BrainSuite: An automated cortical surface identification tool," *Med. Image Anal.*, vol. 6, no. 2, pp. 129–142, 2002.
- [8] S. M. Smith, "Fast robust automated brain extraction," *Hum. Brain Mapp.*, vol. 17, pp. 143–155, 2002.
- [9] M. Jenkinson, M. Pechaud, and S. Smith, "BET2 - MR-based estimation of brain, skull and scalp surfaces," Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Oxford, 2005.
- [10] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.
- [11] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. Segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.
- [12] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.
- [13] S. F. Eskildsen *et al.*, "BEaST: Brain extraction based on non-local segmentation technique," *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.
- [14] R. Souza *et al.*, "An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement," *NeuroImage*, vol. 170, pp. 482–494, 2018.
- [15] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, 2004.
- [16] F. Segonne *et al.*, "A hybrid approach to the skull stripping problem in MRI," *NeuroImage*, vol. 22, no. 3, pp. 1060–1075, 2004.
- [17] M. Yapici, R. Karakis, and K. Gurkahrman, "Improving brain tumor classification with deep learning using synthetic data," *Comput. Mater. Continua*, vol. 74, no. 3, pp. 5049–5067, 2023.
- [18] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

- [19] J. Kleesiek *et al.*, “Deep MRI brain extraction: A 3D convolutional neural network for skull stripping,” *NeuroImage*, vol. 129, pp. 460–469, 2016.
- [20] H. Hwang, H. Z. U. Rehman, and S. Lee, “3D U-Net for skull stripping in brain MRI,” *Appl. Sci.*, vol. 9, no. 3, Art. no. 569, 2019.
- [21] Q. Zhang *et al.*, “FRNET: Flattened residual network for infant MRI skull stripping,” *arXiv preprint arXiv:1904.05578*, 2019.
- [22] K. Gurkahraman and C. Dasgin, “Brain extraction from magnetic resonance images using UNet modified with residual and dense layers,” *Türk Doga ve Fen Dergisi*, unpublished.
- [23] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, “SynthStrip: Skull-stripping for any brain image,” *NeuroImage*, vol. 260, Art. no. 119474, 2022.
- [24] X. Feng, C. Wang, S. Cheng, and L. Guo, “Automatic liver and tumor segmentation of CT based on cascaded U-Net,” in *Proc. Chin. Intell. Syst. Conf.*, 2019, pp. 155–164.
- [25] Y. Ding *et al.*, “A stacked multi-connection simple reducing net for brain tumor segmentation,” *IEEE Access*, vol. 7, pp. 104011–104024, 2019.
- [26] S. K. R. Chinnam, V. Sistla, and V. K. K. Kolli, “Multimodal attention-gated cascaded U-Net model for automatic brain tumor detection and segmentation,” *Biomed. Signal Process. Control*, vol. 78, Art. no. 103907, 2022.
- [27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, vol. 11045, Springer, 2018, pp. 3–11.
- [28] G. Li, K. Wang, P. Lv, *et al.*, “Multistage feature fusion knowledge distillation,” *Scientific Reports*, vol. 14, article no. 13373, 2024.

Enhanced GI Tract Cancer Diagnosis Using CNNs and Machine Learning Models

Abdul Haseeb
0009-0002-6069-1467
Capital University of Science and
Technology, Islamabad
Expressway,
Kahuta Road, Zone-V Islamabad.
Email: abdul.haseeb@cust.edu.pk

Faheem Shehzad
0009-0003-7204-183X
Department of Electrical
Engineering and Information
Technology (DIETI),
University of Naples “Federico II”,
Naples, Italy.
Email: faheem.shehzad@unina.it

Sidra Naseem
Department of Computer Science,
COMSATS University Islamabad,
Vehari Campus, Mailsi Road,
Off Multan Road,
Vehari, Punjab, Pakistan
Email:
sidrafaheem228@gmail.com

Abstract—Gastrointestinal cancer exhibits the greatest mortality rate among all cancers, at 35.4%. Endoscopy is one of the few methods for obtaining visuals of gastrointestinal tract lesions. Manual cancer detection is arduous. Deep learning can autonomously diagnose gastrointestinal tract lesions. Automation produces erroneous detection results. This study used the challenging Hyper-Kvasir dataset for training and validation purposes. The dataset undergoes first preprocessing with Brightness Preserving Histogram Equalization. Furthermore, processed datasets comprise training and validation sets. For segmentation, pretrained backbone-based U-Net architecture is used. The U-Net backbones include EfficientNet-B0, EfficientNet-B7, and DenseNet201. The pre-trained models utilize ImageNet, so Hyper-Kvasir is employed for the fine-tuning of gastrointestinal tract segmentation. The optimal Intersection over Union (IoU) is 85.2% for the EfficientNet-B7 backbone inside the U-Net design. A custom convolutional neural network is employed to classify the hyper kvasir dataset. The suggested network derives profound features for classification using artificial neural networks. The proposed methodology surpassed state-of-the-art (SOTA) methods.

Index Terms—Gastrointestinal Cancer, U-Net, CNNs, Deep learning, Classification, Segmentation.

I. INTRODUCTION

GASTROINTESTINAL (GI) cancer occurs when malignant cells grow inside GI tract [1]. In the past fifty years, GI cancer has had the second highest mortality rates among different types of cancers [2]. According to global cancer index, GI cancer has the mortality rate of 35.4% whereas 26.3% of all the cancers are diagnosed as GI tract cancers. Over the past few years, the performance of artificial intelligence-driven computer-aided diagnosis (CAD) tools in various medical fields has been greatly improved by deep learning algorithms [3], particularly artificial neural networks (ANNs) [4]. Identifying gastrointestinal (GI) illnesses subjectively takes time and professional competence. By automating the detection and categorization of GI illnesses, computer-assisted diagnosis (CAD) technology may reduce these diagnostic obstacles. Such technologies might help doctors detect and cure serious medical diseases early

on. Medical practitioners benefit from CAD technology's precise diagnosis and appropriate action [5]. Deep learning (DL) [6] are statistical based methodologies that authorize computer systems to sovereignly identify patterns and properties from unprocessed data inputs, including structured data, images, text, and audio. The substantial progress in artificial intelligence (AI) based on DL has had a profound impact on numerous domains within clinical practice [7, 8].

In semantic segmentation, every pixel of an object is assigned a specific label corresponding to its class. This process involves categorizing each pixel in an image into predetermined classes. Semantic segmentation relies on the concept of a mask that incorporates edge detection, which helps identify the connected regions in an image that belong to the same class [9]. For the purposes of semantic segmentation, multiple architectures are being used by researchers. Recent studies shows that one of few most effective frameworks for image segmentation in medical domain is U-Net [10].

However, using state-of-the-art pretrained models like U-Net for semantic segmentation may provide considerable results. The design has two partitions: contraction (Encoder) and dilation (Decoder). To get image context, convolutional and pooling layers are used. While the latter half spreads the picture utilizing skip connections and anti-convolution (transpose convolution). The segmented image is the outcome. U-Net is a semantic segmentation benchmark. It improves outcomes with numerous fundamental architectural modifications. Several pretrained networks compose the U-Net architecture's backbone, improving performance. One of the best semantic segmentation designs is U-Net. This design is largely utilized in medical image segmentation.

This study introduces a deep learning methodology for segmenting gastrointestinal (GI) tract lesions in endoscopic images, employing a U-Net architecture with several pretrained models (EfficientNet-B0, EfficientNet-B7, and DenseNet201) utilized as fixed encoders. The decoding component employs the encoded information for precise segmentation. Brightness Preserving Histogram Equalization

(BPHE) is utilized as a preprocessing technique to improve image quality. A unique convolutional neural network is employed for feature extraction and classification utilizing the Hyper-Kvasir dataset. The efficacy of each model is assessed by several criteria and juxtaposed with one another and contemporary state-of-the-art techniques. The document is organized to encompass background, relevant research, methods, findings, and conclusions with prospective directions.

II. LITERATURE REVIEW

This section describes the recent advancements in GI tract segmentation. Researchers achieved significant results for GI cancer segmentation. In [11], authors used U-Net model with depth of five to perform semantic segmentation on hyper Kvasir dataset. Moreover, researchers used the image size of 96×96 at the start of the model. Additionally, in the encoder part the gradual decrease in size is observed till the size of the image becomes 6×6 . Furthermore, convolutional layers with filter size of 3×3 is used whereas filter size for pooling layers is 2×2 . Authors used loss score as evaluation parameter for results and obtained value of 0.69 for loss. A new pipeline for unsupervised domain adaptation (UDA) for the purpose of semantic segmentation is introduced in [12], that combines feature-level adaptation with image-level adaption. To address domain shifts at the image-level, the proposed approach includes a global photometric alignment and global texture alignment modules that are used to align images from the source domains and target domains based on their image-level properties. A global manifold alignment approach is used for feature-level domain change by mapping pixel features from the two domains onto the source domain's feature manifold. Additionally, desired domain consistency regularization is carried out on enhanced target domain images, and category centers in the source domain

are regularized using a class-oriented triplet loss. On hyper kvasir dataset authors achieved 81.5% mean IoU.

As described by Nguyen Thanh Duc et al. [13], polyps in the colorectal region can be detected rapidly and accurately using novel deep learning algorithms. To recognize lesions in colonoscopy images, the authors proposed Colon-Former, a deep learning architecture employing an encoder-decoder architecture. The encoder is a lightweight and efficient modeling framework for multi-scale global semantic connections. The decoder is a representation of visual data generated by a hierarchical network that has been enhanced. Five distinct reference datasets were used to construct the proposed system. This paradigm describes multiscale functionalities via transformers and convolutional neural networks (CNNs). It only supports one architecture and utilizes data from five unique collections. In addition, we observed that it produced the finest results compared to other methods, leading us to conclude that it is a cutting-edge method.

Experts in the field of colonoscopy image analysis emphasized the importance of image segmentation for the detection of lesions caused by colorectal cancer [14]. For image segmentation, regional dense-pixel classification and boundary-based polygon algorithms have previously been established. Using a graphical neural network (GNN) that is based on a deep neural network, the authors developed a novel polyp detection methodology. This technique identifies the polyp area using an attention-enhancement module (AEM). Using the AEM, border and area characteristics of polyps can be extracted. Each plot is data-driven, so the GNN, which functions as a weighted link between the nodes of various domains, preserves the global and local connections between the nodes. It focuses on the demographics and geographic boundaries of the region. The GNN outperforms competing methods and accurately detects malignant lesions in colonoscopy-acquired biomedical images. However, the complexity of the system makes it challenging to precisely

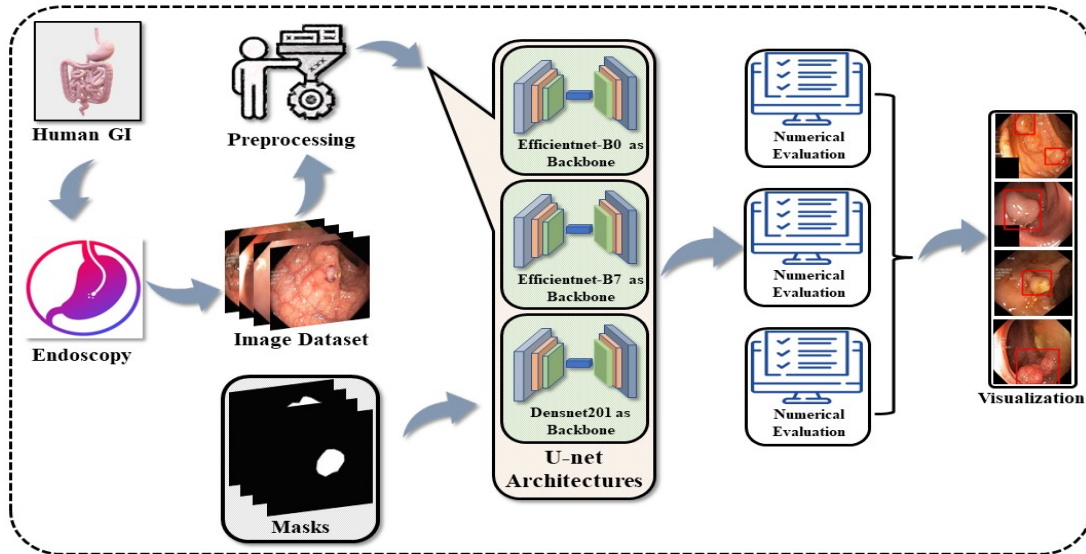


Figure 1: System Model for semantic segmentation of GI tract lesions.

identify the polyp region. Based on comparisons to other innovative methods, this GNN system is the superior model. The polyps could only be identified using specialized equipment.

III. METHODOLOGY

The proposed methodology comprised of multiple steps. In the first step the dataset is pre-processed to enhance the image quality. To achieve this purpose, histogram equalization technique is used to improve the spatial quality of image. The second step of the methodology is to feed these images with their corresponding masks to multiple U-Net based on different backbones. Evaluation of the system is completed using different evaluation parameters. In the last step the system is tested and obtained the visual representation of segmented region of GI tract cancer. Fig. 1 shows the methodology for the segmentation of GI tract lesion.

A. Dataset

Segmentation Dataset: This study explores with hyper kvasir segmentation dataset [19]. The collection comprises 1000 endoscopic images and masks from various GI tract locations from numerous people. Due of the dataset's inconsistent image sizes, images and masks are resized. The final size is 256x256 after resizing. Additionally, the dataset has two subgroups. Subset one has training images and subset two testing images. The training subset includes 80% of the data and the testing subset 20%. Randomization is considered while splitting data. Fig. 2 shows dataset examples of images and masks.

Classification Dataset: Additionally, hyper Kvasir classification dataset with 23 classes is utilized for classification. Due to class imbalance in dataset, dataset is enhanced with images to address this issue. These methods modify the spatial features of dataset images without changing their orientation. Dataset has training, validation, and testing subsets.

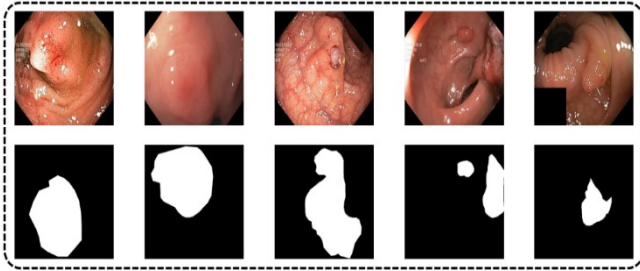


Figure 2: Hyper-Kvasir dataset sample images and masks

B. Data Pre-Processing

To achieve better results, one of the most common phenomena used is pre-processing. Brightness Preserving Histogram Equalization (BPHE) [15] is one of the efficient histogram equalizations to enhance the images. In the discipline of image processing, contrast is enhanced using brightness-preserving histogram equalization. Adjusting the image's histogram so that the intensity levels are distributed more

equitably is one method to improve the image's perceived quality. In contrast to conventional histogram equalization techniques, preserving bi-histogram equalization considers both the light and dark regions of an input image. It adjusts the histograms of each separately to increase contrast while preserving detail in the highlights and shadows. This method excels in situations where, maintaining the detail in both lighter and darker areas is essential, such as medical imaging. Preprocessing is applied only to classification dataset.

C. Convolutional Neural Networks

CNNs have emerged as a useful instrument for analyzing medical images in recent years [16]. If a neural network has at least one convolutional layer, we refer to it as a convolutional neural network (CNN). A convolution operation uses a sliding window technique to apply a fixed-size filter with multiple parameters to an input image. When a layer is complete, the resulting image is sent to the subsequent layer. Here is the mathematical expression for this process:

$$FM_{out}(H_{out} \times V_{out}) = (FM_{inp} * Filter_{op}) \quad (1)$$

The output matrix FM_{out} comprises the rows and columns designated H_{out} and V_{out} as shown in equation (1). Using the rectified linear unit function, the value of a negative feature is set to zero, as shown in the following equation.

$$Active_{ReLU} = \text{Maxof}(0, k), k \in FM_{out} \quad (2)$$

In addition, an aggregating technique is employed to reduce computational complexity and accelerate processing time. This procedure involves exchanging the input value at the center with the utmost or average value in a particular region. Using an entirely linked layer, the features are then transformed into a one-dimensional vector. In mathematical notation, it appears as follows:

$$(Vect_{flat})_0^{out} = FM_{out}(H_{out} \times V_{out}) \quad (3)$$

$$(Vect_{flat})_i^{in} = (Vect_{flat})_{i-1}^{out} * M_i + Vert_i \quad (4)$$

$$(Vect_{flat})_i^{out} = \Delta_i((Vect_{flat})_i^{in}) \quad (5)$$

In above equations, $(Vect_{flat})_0^{out}$ is flattened final one-dimensional vector, i is the layer number. Moreover, Δ shows the activation function used in the operation.

D. U-Net Architecture

U-Net architecture [10] is a popular CNN for image segmentation. U-Net was named because it resembles the letter U when diagrammed. The network's architecture may gain local and global context due to its encoding (contracting) and decoding (expanding) paths. The U-Net design excels in organ and tumor segmentation in biomedical imaging [17]. In addition to semantic and instance segmentation, this approach has been used for many additional segmentation tasks. U-Net may segment using low-level and high-level characteristics due to skip connections. Keep fine-grained

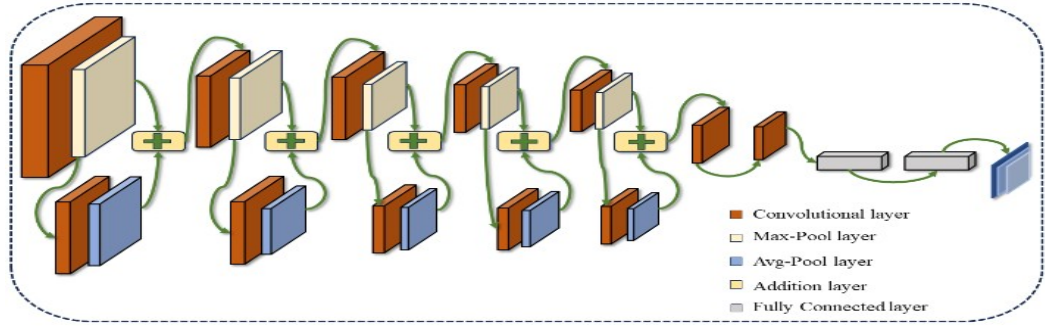


Figure 3: Custom 24 layered proposed architecture for hyper kvasir classification

information while collecting context for more accurate segmentation. This article uses three pretrained networks as U-Net backbones for segmentation. The study uses Densenet201 [18], Efficientnetb0 [19], and Efficientnetb7 [20].

E. Backbones Used for U-Net Architecture

U-Net architecture is a widely used framework for semantic segmentation. U-Net uses encoder and decoder structures to perform semantic segmentation. On the other hand, the Efficientnet models are some of the best models that is been used by researchers in different domains. Combining both networks can be robust and accurate in terms of results. In this paper, Efficientnet-B0, Efficientnet-B7 and Densenet201 are used as backbones. Pre-trained weights of ImageNet dataset are used as weights for encoder part. Skip connections are used from encoder part to decoder path. These skips connections combine the features from encoder part with the decoder part features. This increases the accuracy of segmentation by preserving the spatial information.

The most important concept that is used in U-Net architecture is deconvolution. The purpose of deconvolution is to find a specific solution for any convolution. This is achieved by using the following equation.

$$H_{sol} = (FM_{inp} * Filter_{op}) + \epsilon \quad (6)$$

In above equation, FM_{inp} is the input image while $Filter_{op}$ is the filter used for convolution. Moreover, ϵ is the noise that is added due to convolution operation and $*$ is used for convolution operation.

F. Proposed 24 Layered Architecture

To obtain the classification results pre-trained models are often used by implementing transfer learning techniques. Pre-trained models are trained on “ImageNet” dataset having millions of images categorized in 1000 classes. To achieve results, pre-trained models are fine-tuned on target dataset using transfer learning. The technique freezes the weights of all layers except the last few layers. Transfer learning is used to solve a wide range of problems in deep learning domain. However, there is major drawback in transfer learning technique which is referred to as domain mismatch problem. To resolve this, custom deep architectures are developed by researchers. However, the results with custom models are not

up to the mark. In this article a custom model with 24 layers is proposed through which features are extracted. Moreover, these features are used as input for artificial neural networks for classification.

Custom designed model consists of 24 layers combining 12 convolutional layers, 5 max_pooling layers, 5 average_pooling layers and 2 dense layers. 5 residual blocks are used in the architecture having one convolutional layer and one average pooling layer each. Similarly, linear blocks consist of one convolutional layer and one max pooling layer. Fig. 3 shows the model architecture diagram. Each convolutional layer is activated using relu activation function. However, the last convolutional layer is activated through soft plus function. Both linear and residual blocks are combined using addition operation. Network also have one input layer which takes the images having the dimension $224 \times 224 \times 3$ as input. Furthermore, a classification output layer is also included at the end of the network.

G. Training and validation

Hyper-kvasir classification dataset containing 24000 images is used for the training and validation of the proposed network. Dataset is divided into training and validation sets with 70% training and 30% validation data. Training is performed on the system having windows 10 and Nvidia RTX 3060 GPU. MATLAB R2022A is used as a programming platform. The hyper parameters used for training are “Learning Rate = 0.0001”, “Batch Size = 16”, “Optimizer = Adam”.

IV. RESULTS AND DISCUSSION

In this section the results achieved through the several experiments for segmentation and classification purposes are described. To achieve this Densenet201, Efficientnet-b0 and Efficientnet-b7 are used as backbone in application of U-Net architecture for semantic segmentation. Furthermore, boundary box for affected area is created to identify the lesion in endoscopy images of GI tract. Moreover, for classification custom network is designed and obtained the results.

A. System Setup

The experiments are performed using a system having core i7 processor with four cores and eight threads and 16 GB of RAM. Moreover, the Nvidia GTX 950M graphical

TABLE I.
EVALUATIONS FOR GI TRACT SEGMENTATION USING EFFICIENTNET-B7

Epoch	Validation Accuracy	Validation Dice Coef	Validation IoU	Validation Loss	Validation Precision	Validation Recall
1	0.885135	0.471594	0.313928	0.46715	0.610926	0.916307
10	0.958906	0.83735	0.721798	0.101103	0.911861	0.86235
20	0.962712	0.891788	0.806815	0.109419	0.924289	0.868229
30	0.964142	0.90777	0.834153	0.121391	0.924001	0.877824
40	0.963541	0.909911	0.837617	0.125882	0.918826	0.883313
50	0.96515	0.918313	0.852109	0.140427	0.924259	0.889357

TABLE II.
EVALUATIONS FOR GI TRACT SEGMENTATION USING DENSENET201

Epoch	Validation Accuracy	Validation Dice Coef	Validation IoU	Validation Loss	Validation Precision	Validation Recall
1	0.776002	0.403931	0.257738	0.647087	0.419642	0.943514
10	0.957872	0.82911	0.71018	0.122115	0.930014	0.828044
20	0.958587	0.871834	0.774337	0.137972	0.939886	0.820215
30	0.957271	0.884891	0.796025	0.160757	0.911674	0.843068
40	0.961018	0.894244	0.810935	0.118371	0.921321	0.853872
50	0.96024	0.901343	0.823115	0.140625	0.909625	0.870043

processing unit with 4 GB of VRAM is used for training purposes. All the experiments for segmentation are performed using Python 3.10 and TensorFlow. For training purposes, initial learning rate is set to the value of 0.0001, max epochs are set to 50 whereas batch size is set to the value of 8.

B. Results for Segmentation

Numerical Results are provided through multiple performance measures implemented. Table I shows the gradual depiction of results during the validation of the framework. To obtain the results Efficientnet-B7 is used as backbone for segmentation. IoU is the most important performance measure for semantic segmentation. Results shows that using Efficientnet-B7, the best validation IoU value of 0.85 is obtained while 96% validation accuracy is achieved. Additionally, the table shows gradual increase in validation accuracy, validation dice coefficient, validation precision and validation recall. Moreover, the loss is at lowest as the epochs increase. It is observed that from epoch 1 to 10 the increase in validation accuracy, dice coefficient, IoU, precision and recall values increase significantly whereas the validation loss decreases abruptly. However, the change in the evaluation parameters is less for over 10 epochs as compared to the values before 10 epochs. Another discrepancy is observed during the epochs 30 to 40 which is the sudden decrease in validation accuracy. Moreover, the improvement in evaluation parameters is significant between epoch number 40 to 50.

The best values achieved at the end of the validation are 0.96 for accuracy, 0.91 for dice coefficient, 0.85 for IoU, 0.92 for precision and 0.88 for recall. Additionally, the best loss value at the end of the validation is 0.14.

Table II demonstrates Densenet201-segmented GI tract lesions. Densenet201's maximum validation IoU is 0.82 and accuracy is 96%. Dice coefficient for validation data is 0.90. Increased epochs cause some disruption in steady evolution of outcomes. Compared to Efficientnet, Dense-net201 did not produce smooth results. Analyzing the evaluation parameters shows that the values vary exponentially in the first 10 epochs and less after 10 epochs. Densenet201-based framework training and validation showed considerable inconsistency between epochs 30 and 35. Best values after validation are 0.96 accuracy, 0.90 dice coefficient, 0.82 IoU, 0.90 precision, and 0.87 recall. However, validation loss is 0.

Additionally, Efficientnet-B0 is also used in the study for experiments to segment GI lesion detection. Table III describes that the highest value of validation IoU obtained through Efficientnet-B0 is 0.78 whereas the value of validation accuracy achieved by the framework is 95%. By analyzing the values, it is assessed that the highest validation accuracy achieved at the end of validation process is 0.95. Moreover, for dice coefficient the value is 0.87. Similarly, the value of validation IoU is 0.78. Also, the achieved value for validation precision is 0.93 and the value for validation recall is 0.81. In case of Efficientnet-B0 it is observed that

TABLE III.
EVALUATIONS FOR GI TRACT SEGMENTATION USING EFFICIENTNET-B0

Epoch	Validation Accuracy	Validation Dice Coef	Validation IoU	Validation Loss	Validation Precision	Validation Recall
1	0.776002	0.403931	0.257738	0.647087	0.419642	0.943514
10	0.957872	0.82911	0.71018	0.122115	0.930014	0.828044
20	0.958587	0.871834	0.774337	0.137972	0.939886	0.820215
30	0.957271	0.884891	0.796025	0.160757	0.911674	0.843068
40	0.961018	0.894244	0.810935	0.118371	0.921321	0.853872
50	0.96024	0.901343	0.823115	0.140625	0.909625	0.870043

there is slight gradual decrease in the value of accuracy, dice coefficient, IoU, precision and recall after the 30th epoch. The validation loss also increased after the 30th epoch and achieved the lowest value of 0.17 at the end of the validation process.

Visualization of segmented areas and corresponding bounding boxes are shown in Fig. 4. The first image shows the original image from hyper-kvasir dataset (Validation Data) whereas the second image shows the original masks given with the hyper-kvasir dataset. In the third image, the mask predicted by the Efficientnet-B7 is depicted. Finally, the image with bounding box (corresponds to the predicted mask) is achieved. The figure illustrates that the network predicted the GI tract lesions with phenomenal accuracy.

In table IV proposed system's results are compared with the state-of-the-art (SOTA) previous works. Researchers used different techniques including U-Net, DeeplabV3+, Transformers etc. to segment the GI tract lesion segmentation. It is analyzed that the proposed system outperforms the previous techniques.

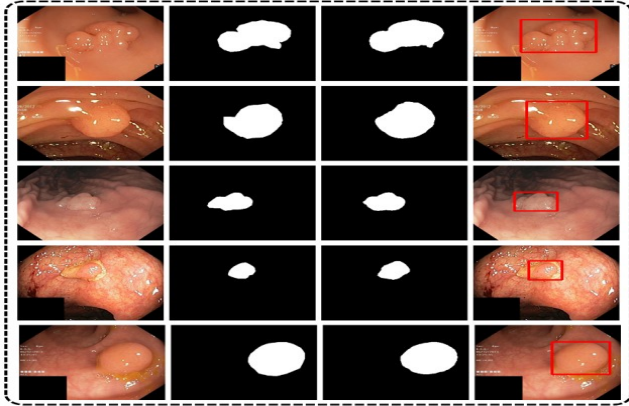


Figure 4: Visualizations of segmentation using U-Net with Efficientnet-B7 backbone: left to right: original image, original annotation, predicted annotation and bounding box around predicted lesions

TABLE IV.

COMPARISON OF SEGMENTATION RESULTS FOR PROPOSED METHODOLOGY WITH STATE-OF-THE-ART (SOTA)

Reference	Dataset	Technique	Results
[11]	Hyper-Kvasir	U-Net	Loss = 0.69
[12]	Hyper-Kvasir + Piccolo	DeeplabV3+	IoU = 0.84
[13]	Kvasir	ColonFormer-L (Transformers)	IoU = 0.87
[21]	Endocv2022 + CVC-Clinics	Improved-STCN Network	Dice = 0.76
[22]	Hyper-Kvasir	MSACL	IoU = 0.40
Proposed	Hyper-Kvasir	-	IoU = 0.85

C. Result for Classification

Custom network design for feature extraction purpose is used to extract the features from test dataset. The extracted feature vector is further fed to multiple Artificial Neural Network (ANNs) classifiers to achieve the best results for classification. Table V comprises accuracies, precisions, recalls and F1-Scores for all classifiers used in experiments. By analyzing the table, it is clear that the Narrow Neural Network classifier gives the best overall results with 92.70% accuracy. The values for precision, recall and F1-score are 92.87, 92.78 and 92.80 respectively. On the other hand, Tri-layered Neural Network has given the worst overall performance with 89.00% accuracy. Fig. 5 depicts the confusion matrix for classification of hyper-kvasir dataset using Narrow Neural Network classifier. Generally, the model performed exceptionally well. Yet some classes still have low classification accuracies which shows that class imbalance problem can alter the model's result drastically in clinical

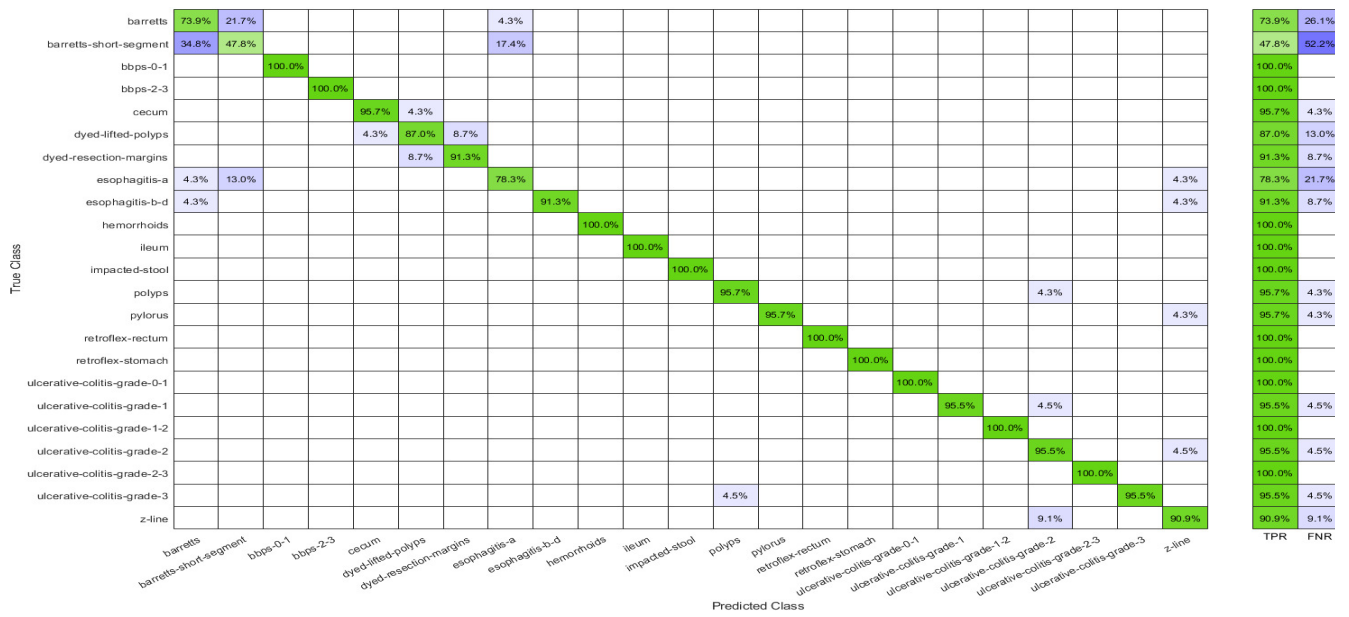


Figure 5: Confusion matrix of narrow neural network for hyper-kvasir dataset classification

environment. Moreover, noisy data can also be a hurdle in acquiring the accurate results in clinical experimentations.

TABLE V

CLASSIFICATION RESULTS FOR ARTIFICIAL NEURAL NETWORK CLASSIFIERS FOR HYPER-KVASIR CLASSIFICATION

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	92.70	92.87	92.78	92.80	39.50
Medium Neural Network	91.50	92.20	91.64	91.90	39.80
Wide Neural Network	92.30	92.56	92.36	92.45	87.27
Bilayered Neural Network	89.20	89.67	89.30	89.48	43.27
Trilayered Neural Network	89.00	89.53	89.13	89.32	51.70

Custom network design for feature extraction purpose is used to extract the features from test dataset. The extracted feature vector is further fed to multiple Artificial Neural Network (ANNs) classifiers to achieve the best results for classification. Table V comprises accuracies, precisions, recalls and F1-Scores for all classifiers used in experiments. By analyzing the table, it is clear that the Narrow Neural Network classifier gives the best overall results with 92.70% accuracy. The values for precision, recall and F1-score are 92.87, 92.78 and 92.80 respectively. On the other hand, Trilayered Neural Network has given the worst overall performance with 89.00% accuracy. Fig. 5 depicts the confusion matrix for classification of hyper-kvasir dataset using Narrow Neural Network classifier. Generally, the

model performed exceptionally well. Yet some classes still have low classification accuracies which shows that class imbalance problem can alter the model's result drastically in clinical environment. Moreover, noisy data can also be a hurdle in acquiring the accurate results in clinical experimentations.

TABLE VI

CLASSIFICATION RESULTS FOR ARTIFICIAL NEURAL NETWORK CLASSIFIERS FOR HYPER-KVASIR CLASSIFICATION

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	92.70	92.87	92.78	92.80	39.50
Medium Neural Network	91.50	92.20	91.64	91.90	39.80
Wide Neural Network	92.30	92.56	92.36	92.45	87.27
Bilayered Neural Network	89.20	89.67	89.30	89.48	43.27
Trilayered Neural Network	89.00	89.53	89.13	89.32	51.70

Table VI exhibits the comparison of the proposed technique with recent literature. By analyzing the results, it is clear that the proposed method outperforms the state-of-the-art (SOTA) with significant margins.

V. CONCLUSION

World-wide, GI tract cancer is frequent. This study introduces U-Net topologies with Efficientnet-B0, B7, and Densnet201 backbones to identify GI lesions. Model training

TABLE VII

COMPARISON OF RESULTS WITH STATE-OF-THE-ART (SOTA)

Reference	Dataset	Classes	Year	Accuracy (%)
[23]	Hyper-Kvasir	6	2023	87.45
[24]	Hyper-Kvasir	14	2020	73.66
[25]	Kvasir	5	2021	97.00
[26]	Hyper-Kvasir	23	2020	63.00 for macro
Proposed	Hyper-Kvasir	23	-	92.87

and validation employ Hyper-Kvasir segmentation dataset. Endoscopic photos of malignant regions are labeled by medical professionals in the collection. Brightness Preserving Histogram Equalization improves photos. To train models, the proposed U-Net with Efficientnet-B0, B7, and Densenet201 backbones receives enhanced pictures and masks. Custom deep models are used to classify and extract features. These characteristics also feed artificial neural networks for categorization. The paper also discusses validation data outcomes. The suggested strategy outperforms state-of-the-art methods. Visual differences in photos prevented certain models from performing as expected.

REFERENCES

- [1] M. Vajihinejad, "A Systematic Review of Clinic Pathology and Survival in Gastrointestinal Stromal Tumors," *International Journal of New Chemistry*, vol. 10, no. 1, pp. 33-61, 2023.
- [2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *Ca Cancer J Clin*, vol. 68, no. 6, pp. 394-424, 2018.
- [3] S. Aamir, A. Rahim, S. Bashir, and M. Naeem, "Prediction of Breast Cancer Using AI-Based Methods," in *Intelligent Environments 2021*: IOS Press, 2021, pp. 213-220.
- [4] H. Ko *et al.*, "COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation," *Journal of medical Internet research*, vol. 22, no. 6, p. e19569, 2020.
- [5] M. Owais, M. Arsalan, J. Choi, T. Mahmood, and K. R. Park, "Artificial intelligence-based classification of multiple gastrointestinal diseases using endoscopy videos for clinical diagnosis," *Journal of clinical medicine*, vol. 8, no. 7, p. 986, 2019.
- [6] H. Qayyum, S. T. H. Rizvi, M. Naeem, U. b. Khalid, M. Abbas, and A. Coronato, "Enhancing Diagnostic Accuracy for Skin Cancer and COVID-19 Detection: A Comparative Study Using a Stacked Ensemble Method," *Technologies*, vol. 12, no. 9, p. 142, 2024.
- [7] K. Sumiyama, T. Futakuchi, S. Kamba, H. Matsui, and N. Tamai, "Artificial intelligence in endoscopy: Present and future perspectives," *Digestive Endoscopy*, vol. 33, no. 2, pp. 218-230, 2021.
- [8] A. Ismail, M. Naeem, M. H. Syed, M. Abbas, and A. Coronato, "Advancing Patient Care with an Intelligent and Personalized Medication Engagement System," *Information*, vol. 15, no. 10, p. 609, 2024.
- [9] J. Cheng, H. Li, D. Li, S. Hua, and V. S. Sheng, "A Survey on Image Semantic Segmentation Using Deep Learning Techniques," *Computers, Materials and Continua*, vol. 74, no. 1, pp. 1941-1957, 2023.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015: Springer, pp. 234-241.
- [11] A. S. Narasimha Raju, K. Jayavel, and T. Rajalakshmi, "Dexterous Identification of Carcinoma through ColoRectalCADx with Dichotomous Fusion CNN and UNet Semantic Segmentation," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.
- [12] H. Ma, X. Lin, and Y. Yu, "I2F: A Unified Image-to-Feature Approach for Domain Adaptive Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [13] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "ColonFormer: an efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80575-80586, 2022.
- [14] Y. Meng *et al.*, "Graph-based region and boundary aggregation for biomedical image segmentation," *IEEE transactions on medical imaging*, vol. 41, no. 3, pp. 690-701, 2021.
- [15] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE transactions on Consumer Electronics*, vol. 43, no. 1, pp. 1-8, 1997.
- [16] F. Shehzad *et al.*, "Two-stream deep learning architecture-based human action recognition," 2023.
- [17] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82031-82057, 2021.
- [18] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11-19.
- [19] W. Mulim, M. F. Revikasha, and N. Hanafiah, "Waste Classification Using EfficientNet-B0," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, 2021, vol. 1: IEEE, pp. 253-257.
- [20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019: PMLR, pp. 6105-6114.
- [21] Q. He, X. Hu, F. Sun, L. Zhou, J. Wang, and Q. Wan, "Improved-STCN Network with Enhanced Strategy for Sequence Polyp Segmentation," 2022.
- [22] Y. Tian *et al.*, "Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images," *arXiv preprint arXiv:2109.01303*, 2021.
- [23] T. Nguyen-DP, M. Luong, M. Kaaniche, J. Chaussard, and A. Beghdadi, "Self-supervised Learning for Gastrointestinal Pathologies Endoscopy Image Classification with Triplet Loss," *arXiv preprint arXiv:2303.01672*, 2023.
- [24] P. H. Smedsrud *et al.*, "Kvasir-Capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.
- [25] M. Hmoud Al-Adhaileh *et al.*, "Deep learning algorithms for detection and classification of gastrointestinal diseases," *Complexity*, vol. 2021, pp. 1-12, 2021.
- [26] H. Borgli *et al.*, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, p. 283, 2020.

Towards OntoUML for Software Engineering: Transformation of Constraints into Various Relational Databases

Jakub Jabůrek, Zdeněk Rybola and Petr Kroha

ORCID: 0009-0004-8212-2059, 0000-0001-9430-6921, 0000-0002-1658-3736

Faculty of Information Technology, Czech Technical University in Prague

Thákurova 9, 160 00 Praha 6, Czech Republic

Email: {jakub.jaburek, zdenek.rybola, petr.kroha}@fit.cvut.cz

Abstract—OntoUML is an ontologically well-founded conceptual modeling language that provides precise meaning to modeled elements. As a result, its usage is beneficial in the Model-Driven Development approach to software development. Relational databases are commonly used for storage of application data, and they offer support for the implementation of custom data constraints. In this paper, we discuss the realization of constraints that arise from OntoUML structural models in PostgreSQL, Microsoft SQL Server and MySQL, and provide a complete reference on how to implement these constraints so that only data conforming to the OntoUML model can be stored.

Index Terms—Conceptual Modeling, Software Development, Relational Database, Model Transformation, Constraints.

I. INTRODUCTION

SOFTWARE engineering is a demanding discipline that deals with complex systems. The goal of software engineering is to ensure high quality of the implementation of software systems. Various approaches to software development exist to achieve this.

In this paper, we focus on the Model-Driven Development (MDD) approach. It is based on the construction of models of the software and their transformations. An established practice within MDD is *forward engineering* — transformation of abstract models into more concrete ones [1]. One of the common use cases of such process is the transformation of a conceptual data model into application source code or database schema.

In MDD, the quality of the final implementation of the system depends on the accuracy of the transformation process as well as on the precision of the initial models. Therefore, a conceptual data model should capture the necessary constraints, and the transformation should preserve them in the transformed result.

For the conceptual data model, we use OntoUML — as it is based on cognitive science and modal logic [2], we consider it to be suitable for the development of highly expressive data models. In comparison with other modeling languages, such as the Entity–relationship model, OntoUML enforces stronger semantic constraints, therefore resulting in models that more closely follow reality.

For the implementation target, we focus on relational database management systems (RDBMS), as they are one of the most commonly used data persistence platforms [3]. As different RDBMS implementations have varying levels of support for the definition and enforcement of data constraints, the implementation is specific to the particular RDBMS product.

The existing literature describes the transformation process of an OntoUML model into Oracle RDBMS [4]. This transformation is divided into three successive steps:

- 1) First, the OntoUML model is transformed into a Unified Modeling Language (UML) class model.
- 2) Second, the UML model is transformed into a relational data model.
- 3) Finally, the relational data model is transformed to an implementation in the Structured Query Language (SQL) tailored to a specific RDBMS.

Our approach follows the aforementioned division and reuses transformation steps 1–2. The original contribution of this paper consists of novel implementation of the third transformation step, which is RDBMS-vendor-specific. While the transformation of a relational model to SQL is well-known, our approach focuses on preserving the integrity constraints defined by the original OntoUML model.

We elaborate implementations in MySQL, Microsoft SQL Server and PostgreSQL, while the realization in Oracle is described in existing literature [4].

This practically oriented paper is structured as follows: In Section II, OntoUML concepts relevant to the transformation are summarized, the features of SQL related to constraint checking are summarized, and the existing transformation of an OntoUML model to SQL is introduced. In Section III, we discuss constraint checking features supported in the considered RDBMS. In Section IV, we elaborate the implementation of constraints in SQL for each of the considered RDBMS. In Section V, we discuss the limitations of our approach. Finally, in Section VI, the results of this paper are summarized.

II. BACKGROUND AND RELATED WORK

In this section, OntoUML, being the language of the input models of our transformation, is introduced. Then, SQL, the

target language of the transformation, is discussed. Finally, the existing transformation of OntoUML to relational schema, which our transformation to SQL builds upon, is introduced.

A. OntoUML

OntoUML is a conceptual modeling language focused on building ontologically well-founded structural models [5]. It is a profile of the Unified Modeling Language (UML) Class Diagram that realizes the Unified Foundational Ontology (UFO) theory formulated in G. Guizzardi's Ph.D. thesis [6].

An OntoUML model defines *universals*—bundles of characteristics shared among their instances. Universals are instantiated by *individuals*. UFO differentiates between universals of *Sortal* and *Non-Sortal* type. Sortals bear an *identity principle* (either they provide it themselves or inherit it from a Sortal supertype), which defines how individual instances of the particular universal are distinguished [7].

A single individual may instantiate multiple universals, and the set of universals instantiated by a particular individual may change over time. Universal types that an individual must instantiate during its entire lifetime (or not instantiate at all) are classified as *Rigid*. Conversely, universal types that an individual may start to instantiate later or cease to instantiate prior to its own destruction are classified as *Anti-Rigid* [8].

UFO defines a taxonomy of universal types, with the concrete types then being applied as stereotypes to classes in an OntoUML model. In this section, we summarize a selection of universal types that significantly contribute to the set of constraints derived during the transformation of an OntoUML model to a relational schema, as introduced in Section II-C.

The backbone of an OntoUML model consists of *Kinds* and *SubKinds*—Rigid Sortals, with SubKinds inheriting the identity principle from Kinds through the generalization relationship.

An example is provided in Fig. 1. The model describes Kind Document with two SubKinds IdCard and Passport. The generalization set is covering and disjoint—an instance of Document must be at the same time an instance of exactly one of the subclasses. A Person Kind is also modeled. It can have any number of associated IdCard instances, but an IdCard must always be associated with exactly one Person. A Person must be associated with at least two but not more than four Passport instances, while a Passport must always be associated with exactly one Person. Also in the example, a Brain Kind is present, whose instance must always be associated with one Person instance (and vice versa). The particular instances in the association relationship may not change over time.

Next, an OntoUML model may define *Roles* and *Phases*, which are classified as Anti-Rigid Sortals [8]. Roles are relationally dependent and must be connected to at least one *Mediation* relationship (with a *Relator* universal at the other end) [9], while Phases are intrinsic and not relationally-dependent. A Phase must always be a part of a *phase partition* (a disjoint and covering generalization set) [10].

In the example OntoUML model in Fig. 1, the Person Kind has two Phases: Alive and Deceased, exactly one of them must always be instantiated.

More universal and relationship types are defined in UFO; their description is out of scope of this paper. For more detailed definitions, the reader is referenced to literature covering the UFO theory [6] [7] [8]. Thanks to the constraints placed by UFO on the types and relationships in an OntoUML model, OntoUML models are able to capture many domain constraints natively, as opposed to other modeling languages, such as plain UML, in which such constraints must be specified by other means.

The example OntoUML model in Fig. 1 is used throughout this paper to illustrate the realization of constraints in SQL. The model is used as the input to the transformation introduced in Section II-C, where the transformation result is also shown.

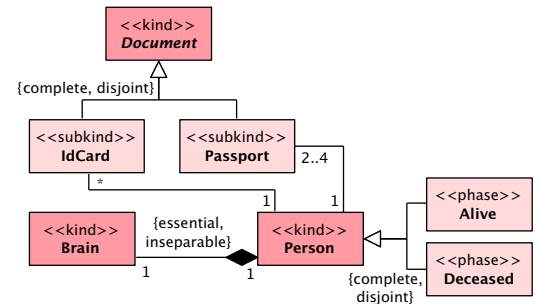


Fig. 1. OntoUML model of the running example

B. Structured Query Language

Structured Query Language (SQL) is a language used to manage data in RDBMS. In this paper, we deal with two commonly recognized parts of SQL [11]:

- Data Manipulation Language (DML)—querying, inserting, modifying and deleting data in the database,
- Data Definition Language (DDL)—definition of tables and integrity constraints.

In this paper, we focus on the realization of constraints, which pertains to the DDL part of SQL. Although SQL is defined by an ISO standard, not all SQL statements are universally portable across different RDBMS vendors. While SQL syntax is consistent across RDBMS, the set of supported features for integrity constraint checking varies; therefore not all DDL constructs defined in the SQL standard can be used in every RDBMS implementation.

As described in existing literature and in the SQL standard, the following mechanisms can be used to realize integrity constraints in a relational database (mechanisms not implemented in any of the mainstream RDBMS are omitted) [12]:

- PRIMARY KEY—column that contains a unique identifier for each row in a table,
- UNIQUE—column that must have a unique value across all rows in a table,
- FOREIGN KEY—column that references an existing record in another table,

- CHECK — expression that realizes a row-level constraint,
- TRIGGER — procedure that runs before or after a DML operation, and can change or block the statement.

Additionally, SQL defines two modes of constraint checking: *immediate* and *deferred*. In immediate mode, a constraint is checked at the end of each DML statement. In deferred mode, the constraint is checked at the end of the SQL transaction [12].

C. Transformation of OntoUML into SQL

The process of transforming an OntoUML into SQL has been elaborated in the Ph.D. thesis of Z. Rybala [4], and was further discussed in [13]. In this subsection, we offer a high-level overview of their approach, which is divided into three successive steps:

1) *Transformation of OntoUML into UML*. First, the OntoUML model is transformed into UML with added conditions written in the Object Constraint Language (OCL) to preserve the constraints defined by the original OntoUML model [4].

2) *Transformation of UML into a Relational Model*. Second, the UML model is transformed into a relational schema by using well-known algorithms. In addition, the OCL constraints from the previous step are transformed as well to preserve the meaning of the original OntoUML model.

Concerning the transformation of the generalization relationship between classes into the relational model, various approaches are described in the existing literature [14]. In this paper, we focus on the *related tables* approach, where each class in the hierarchy is transformed to a separate table.

To illustrate the realization of constraints in SQL in Section IV, a relational model shown in Fig. 2 is used. This model is the result of the aforementioned transformation applied to the OntoUML model in Fig. 1. Due to limited space, the derived OCL constraints are omitted.

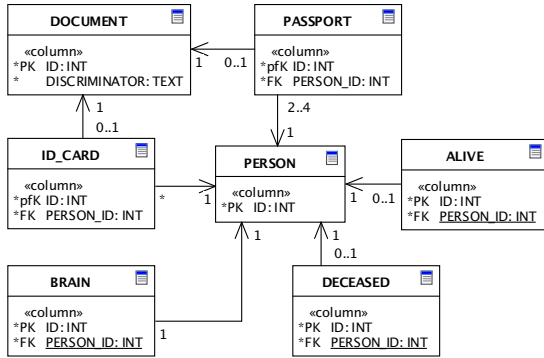


Fig. 2. Running example transformed into a relational model

3) *Transformation of the Relational Model into SQL*. Finally, the relational model is transformed into SQL. While the transformation of tables and columns alone is straightforward, the implementation of OCL constraints from the previous step is non-trivial. The existing literature describes the realization in Oracle [4]. However, due to differences in constraint check-

TABLE I
CONSTRAINT CHECKING SUPPORT MATRIX

Constraint	Oracle	MySQL	MSSQL	PostgreSQL
FOREIGN KEY				
deferrable	imm./defer.	immediate	immediate	imm./defer.
TRIGGER				
# of events	multiple	one	multiple	multiple
activation	before/after	before/after	after	before/after
granularity	row/statem.	row	statement	row/statem.

ing mechanisms (as discussed in Section III), a realization in one RDBMS is generally not portable to another.

III. RELATIONAL DATABASE MANAGEMENT SYSTEMS

According to the DB-Engines Ranking of Relational DBMS, the top four RDBMS (as of April 2025) are the following [15]:

- 1) Oracle
- 2) MySQL
- 3) Microsoft SQL Server (MSSQL)
- 4) PostgreSQL

As discussed in Section II-B, the support for integrity constraint checking is not uniform across various RDBMS. The four vendors listed above implement at least some forms of PRIMARY KEY, UNIQUE, FOREIGN KEY, CHECK and TRIGGER constraints.

We surveyed the features supported by the top four RDBMS important for the realization of constraints derived from an OntoUML model—the results are presented in table I.

Oracle 23ai and PostgreSQL 17 offer complete support for all surveyed constraint types. MySQL 9.3 supports only one activation event per trigger, the expressiveness is however equivalent to RDBMS that support any number of triggering events. Also, MySQL supports only row-level triggers. Conversely, MSSQL 2022 implements only statement-level triggers, and they can be executed only after the DML statement.

MySQL and MSSQL offer immediate constraint checking only. For the FOREIGN KEY constraint, both RDBMS offer mechanisms to temporarily disable integrity checks. MySQL however does not check existing data for consistency when the constraint is re-enabled, therefore referential integrity may be broken indefinitely.

IV. IMPLEMENTATION OF CONSTRAINTS IN SQL

In this section, we present the main contributions of our paper: the implementation of constraints derived during the transformation of an OntoUML model to a relational schema in SQL for MySQL, Microsoft SQL Server and PostgreSQL.

In our approach, an OntoUML model is not transformed to SQL directly. Instead, we rely on the existing transformation of an OntoUML model to a relational data model, as discussed in Subsection II-C:

- 1) First, the OntoUML model is transformed into UML. Additional constraints are derived so the semantics of the original model are preserved where the UML class model alone is not sufficient.

- 2) Second, the UML model is transformed into a relational data model. The derived constraints are transformed as well to apply to the relational model.
- 3) Finally, the relational model is transformed into SQL, including the derived constraints. Based on the transformed constraints, the RDBMS prevents the creation of data that do not conform to the original OntoUML model.

The existing literature provides the implementation of constraints in Oracle [4]. As surveyed in Section III, the top four RDBMS implement varying features related to constraint checking. Therefore, the realization of constraints in SQL is dependent on the particular RDBMS product.

In this section, we thoroughly elaborate the realization of all constraints that can be derived from an OntoUML model in MySQL, Microsoft SQL Server (MSSQL) and PostgreSQL. The section is structured as follows: in Subsection IV-A, we first discuss a common approach to the realization of association-related constraints, then, in the remaining subsection, we discuss the realization of individual constraints. We illustrate the implementation by SQL listings applicable to the relational model in Fig. 2. The model was produced by the transformation introduced in Section II-C from the OntoUML model presented in Fig. 1.

A. Associations

In this subsection, we introduce the common approach for the implementation of constraints related to associations that is used throughout the entire Section IV.

For constraints involving associations, we often need to ensure that a referencing record exists. In OCL, this requirement is expressed as an invariant on the referenced table. The only invariant-like generic constraint implemented by RDBMS is the CHECK constraint, which cannot contain subqueries, therefore it cannot be used to implement this requirement in SQL. Instead, a set of triggers must be introduced, which are executed on all DML operations that may cause the invariant to not hold.

Notably, an INSERT operation on the referenced table without a corresponding INSERT operation on the referencing table would produce a non-referenced record. With triggers, this can be prevented only by checking the existence of the referencing record at the time of the insertion of the referenced record. Therefore, an additional restriction is placed on the order of DML operations: the referencing record must be inserted before the referenced record.

For PostgreSQL, the implication is that the FOREIGN KEY constraint in the referencing table must be checked in deferred mode. MySQL and MSSQL do not implement deferred constraint checking. In our approach, for MSSQL, we propose temporarily disabling the FOREIGN KEY constraint and re-enabling it after both records are inserted. MySQL however does not check the integrity of existing data when the constraint is re-enabled. To not break referential integrity, we present a different approach, where the client is required to

manually execute a stored procedure that checks the existence of the referencing record once both sides are inserted.

In similar fashion, the order of DELETE operations is enforced to be either (a) the referenced record before the referencing record in PostgreSQL and MSSQL, or (b) the referencing record before the referenced record in MySQL.

B. Immutability

In OntoUML models, *immutability constraints* commonly emerge from relationships between a Sortal that provides an identity principle and its subtypes. In SQL, the implementation depends on whether an entire association or only a single column needs to be guarded by such constraint, as discussed in this subsection.

1) *Immutable Column*. To ensure that the value in a column remains unchanged, a BEFORE UPDATE row-level trigger must be defined in MySQL and PostgreSQL that blocks the operation if the old and new column values are different. An example trigger that prevent the change of the PERSON_ID column in the BRAIN table is implemented for MySQL in Listing 1, implementation in PostgreSQL is similar.

Listing 1 Realization of the immutability constraint in MySQL

```
CREATE TRIGGER `IM_BRAIN_PERSON_ID_UPD`
BEFORE UPDATE ON `BRAIN` FOR EACH ROW BEGIN
IF OLD.`PERSON_ID` <> NEW.`PERSON_ID` THEN
  SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

As MSSQL does not support row-level triggers, a different approach needs to be employed. The statement-level trigger in Listing 2 uses the UPDATE function to short-circuit the condition in case the guarded column was not a part of the UPDATE operation. However, the function returns TRUE when updating to the same value. Therefore, the trigger needs to query inserted and deleted special tables, which contain the new and old values, respectively (w.r.t. the UPDATE operation). To determine whether the value of the guarded column actually changed, the two tables are joined on the PRIMARY KEY column and then the new and old values of the column are compared. Note that the lack of row-level triggers in MSSQL makes it impossible to detect a change in the PRIMARY KEY column, also such a change breaks the change detection for other columns in the table as well.

Listing 2 Realization of the immutability constraint in MSSQL

```
CREATE TRIGGER [IM_BRAIN_PERSON_ID_UPD]
ON [BRAIN] AFTER UPDATE AS BEGIN
IF UPDATE([PERSON_ID]) AND EXISTS (
  SELECT 1 FROM inserted i
  JOIN deleted d ON i.[BRAIN_ID] = d.[BRAIN_ID]
  WHERE i.[PERSON_ID] <> d.[PERSON_ID]) BEGIN
  THROW 50000, '...', 1; END; END;
```

2) *Immutable Association*. When the referenced side (i.e., opposite the table containing the FOREIGN KEY) is marked as immutable, the realization in SQL is the same as the previously discussed *immutable column* constraint. The constraint

is applied to the column in the referencing table that bears the FOREIGN KEY constraint. The deletion of the referenced record is prevented by the FOREIGN KEY constraint itself.

In case the referencing side is marked as immutable, the FOREIGN KEY column is immutable with the same realization as in the previous case. Additionally, the deletion of the referencing record must be prevented. In MySQL and PostgreSQL, this is realized by BEFORE DELETE row-level triggers that block the operation when an existing record is referenced by the referencing record (see Listing 3 for MySQL, implementation in PostgreSQL is similar). Again, in MSSQL, the trigger must be statement-level and it must query the deleted special table (see Listing 4).

Listing 3 Trigger preventing the deletion of a referencing record in MySQL

```
CREATE TRIGGER `IM_BRAIN_PERSON_ID_DEL`
BEFORE DELETE ON `BRAIN` FOR EACH ROW BEGIN
IF OLD.`PERSON_ID` IS NOT NULL AND EXISTS (
  SELECT 1 FROM `PERSON` `p`
  WHERE `p`.`ID` = OLD.`PERSON_ID`) THEN
  SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

Listing 4 Trigger preventing the deletion of a referencing record in MSSQL

```
CREATE TRIGGER [IM_BRAIN_PERSON_ID_DEL]
ON [BRAIN] AFTER DELETE AS BEGIN
IF EXISTS (SELECT 1 FROM deleted d
  JOIN [PERSON] p ON d.[PERSON_ID] = p.[ID]) BEGIN
  THROW 50000, '...', 2;
END; END;
```

C. Exclusive Associations

The *exclusive associations constraint* is derived from On-toUML phase partitions [16]. It guards a set of referencing tables where a record in only one of the referencing tables can refer a record in the referenced table. It follows that the following five DML operations may violate the constraint:

- 1) INSERT or UPDATE on the referenced table—the number of referencing records is not exactly one,
- 2) INSERT to any of the referencing tables—the referenced record is already referenced,
- 3) UPDATE on any referencing table—the referencing record may get de-associated or attached to a record that already has an association,
- 4) DELETE from any referencing table—an orphan may be created in the referenced table.

To ensure exactly one referencing record exists when the referenced record is being inserted, a BEFORE INSERT/UPDATE trigger is employed. An example trigger that guards the PERSON referenced table and ALIVE and DECEASED referencing tables is provided in Listings 5 and 6 for PostgreSQL and MSSQL, respectively. As MSSQL does not support BEFORE nor row-level triggers, an AFTER trigger that queries the `inserted` special table is used instead.

Listing 5 INSERT/UPDATE trigger on the referenced table realizing an exclusive associations constraint in PostgreSQL

```
CREATE FUNCTION ex_person_phase()
RETURNS TRIGGER AS $$ BEGIN
IF NOT (
  (EXISTS (SELECT 1 FROM "ALIVE" a
    WHERE a."PERSON_ID" = NEW."ID")
  AND NOT EXISTS (SELECT 1 FROM "DECEASED" d
    WHERE d."PERSON_ID" = NEW."ID")) OR
  (NOT EXISTS (SELECT 1 FROM "ALIVE" a
    WHERE a."PERSON_ID" = NEW."ID")
  AND EXISTS (SELECT 1 FROM "DECEASED" d
    WHERE d."PERSON_ID" = NEW."ID"))) THEN
  RAISE EXCEPTION '...';
END IF; RETURN NEW; END; $$ LANGUAGE plpgsql;
CREATE TRIGGER ex_person_phase
BEFORE INSERT OR UPDATE ON "PERSON" FOR EACH ROW
EXECUTE FUNCTION ex_person_phase();
```

Listing 6 INSERT/UPDATE trigger on the referenced table realizing an exclusive associations constraint in MSSQL

```
CREATE TRIGGER [EX_PERSON_PHASE]
ON [PERSON] AFTER INSERT, UPDATE AS BEGIN
IF EXISTS (SELECT 1 FROM inserted i WHERE NOT (
  (EXISTS (SELECT 1 FROM [ALIVE] a
    WHERE a.[PERSON_ID] = i.[ID])
  AND NOT EXISTS (SELECT 1 FROM [DECEASED] d
    WHERE d.[PERSON_ID] = i.[ID])) OR
  (NOT EXISTS (SELECT 1 FROM [ALIVE] a
    WHERE a.[PERSON_ID] = i.[ID])
  AND EXISTS (SELECT 1 FROM [DECEASED] d
    WHERE d.[PERSON_ID] = i.[ID]))) BEGIN
  THROW 50000, '...', 1; END; END;
```

As discussed in Subsection IV-A, a manually invoked procedure must be used in MySQL. The procedure, as illustrated in Listing 7, needs to be executed by the client after inserting or updating a record in the referenced table, but only after the referencing record is inserted (if applicable). The ID input parameter shall contain the primary key of the inserted record.

Listing 7 Procedure on the referenced table realizing an exclusive associations constraint in MySQL

```
CREATE PROCEDURE `EX_PERSON_PHASE` (IN `ID` INT)
BEGIN
IF NOT (
  (EXISTS (SELECT 1 FROM `ALIVE` a
    WHERE a.`PERSON_ID` = `ID`)
  AND NOT EXISTS (SELECT 1 FROM `DECEASED` d
    WHERE d.`PERSON_ID` = `ID`)) OR
  (NOT EXISTS (SELECT 1 FROM `ALIVE` a
    WHERE a.`PERSON_ID` = `ID`)
  AND EXISTS (SELECT 1 FROM `DECEASED` d
    WHERE d.`PERSON_ID` = `ID`))) THEN
  SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

Furthermore, the constraint may get violated when a record is inserted into a table that belongs to the mutually exclusive set. Therefore, a BEFORE INSERT trigger is in place for all tables in the set that prevents the operation if a matching referencing record already exists in any of the other tables. An example of such trigger is provided in Listing 8 for MySQL, implementation in PostgreSQL is similar.

Listing 8 INSERT trigger on a referencing table realizing an exclusive associations constraint in MySQL

```
CREATE TRIGGER `EX_PERSON_PHASE_ALIVE_INS`
BEFORE INSERT ON `ALIVE` FOR EACH ROW BEGIN
IF EXISTS (SELECT 1 FROM `DECEASED` d
WHERE d.`PERSON_ID` = NEW.`ID`) THEN
SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

As MSSQL does not support BEFORE nor row-level triggers, an AFTER trigger that queries the inserted special table is used instead, as illustrated in Listing 9.

Listing 9 INSERT trigger on a referencing table realizing an exclusive associations constraint in MSSQL

```
CREATE TRIGGER [EX_PERSON_PHASE_ALIVE_INS]
ON [ALIVE] AFTER INSERT AS BEGIN
IF EXISTS (SELECT 1 FROM inserted i
JOIN [DECEASED] d ON d.[PERSON_ID] = i.[ID])
BEGIN
THROW 50000, '...', 2; END; END;
```

Finally, the exclusive association constraint may get violated when a record in the referencing tables is updated or deleted. In PostgreSQL and MSSQL, an AFTER UPDATE/DELETE trigger on all referencing tables is used. The trigger checks that after the operation finishes, all records in the referenced table have exactly one counterpart across all referencing tables; see Listing 10 for an example in MSSQL, implementation in PostgreSQL is similar.

Listing 10 UPDATE/DELETE trigger on a referencing table realizing an exclusive association constraint in MSSQL

```
CREATE TRIGGER [EX_PERSON_PHASE_ALIVE_UPD_DEL]
ON [ALIVE] AFTER UPDATE, DELETE AS BEGIN
IF EXISTS (SELECT 1 FROM [PERSON] p WHERE NOT (
(EXISTS (SELECT 1 FROM [ALIVE] a
WHERE a.[PERSON_ID] = p.[ID])
AND NOT EXISTS (SELECT 1 FROM [DECEASED] d
WHERE d.[PERSON_ID] = p.[ID])) OR
(NOT EXISTS (SELECT 1 FROM [ALIVE] a
WHERE a.[PERSON_ID] = p.[ID])
AND EXISTS (SELECT 1 FROM [DECEASED] d
WHERE d.[PERSON_ID] = p.[ID]))) BEGIN
THROW 50000, '...', 3; END; END;
```

MySQL (i) does not support triggers with multiple triggering events, (ii) does not support statement-level triggers. Therefore, a different implementation must be used. To avoid the mutating table error in a row-level trigger, it must not access the table on which it was fired. Additionally, the trigger must be split up to an UPDATE and a separate DELETE trigger. The UPDATE trigger must check whether the reference has been changed to a different record, if so, it must check whether the sum of records referencing the previous record across all the other referencing tables is one and whether no record that references the currently referenced record exists among all the other referencing tables; see Listing 11. The DELETE trigger works in a similar fashion while omitting the check of the currently referenced record; see Listing 12.

Listing 11 UPDATE trigger on a referencing table realizing an exclusive association constraint in MySQL

```
CREATE TRIGGER `EX_PERSON_PHASE_ALIVE_UPD`
AFTER UPDATE ON `PERSON` FOR EACH ROW BEGIN
IF OLD.`PERSON_ID` <> NEW.`PERSON_ID` THEN
SET @o_count := SELECT COUNT(*) FROM `DECEASED` d
WHERE d.`PERSON_ID` = OLD.`PERSON_ID`;
IF @o_count <> 1 OR EXISTS (
SELECT 1 FROM `DECEASED` d
WHERE d.`PERSON_ID` = NEW.`PERSON_ID`) THEN
SIGNAL SQLSTATE '45000'
SET MESSAGE_TEXT = "...";
END IF; END IF; END;
```

Listing 12 DELETE trigger on a referencing table realizing an exclusive association constraint in MySQL

```
CREATE TRIGGER `EX_PERSON_PHASE_ALIVE_DEL`
AFTER DELETE ON `ALIVE` FOR EACH ROW BEGIN
SET @o_count := SELECT COUNT(*) FROM `DECEASED` d
WHERE d.`PERSON_ID` = OLD.`PERSON_ID`;
IF @o_count <> 1 THEN
SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

D. Generalization Set

For generalization sets, the constraint must guarantee a valid value in the discriminator column and a correct set of referencing records [14]. We argue there are five kinds of DML operations that may violate the constraint:

- 1) INSERT and UPDATE on the superclass table—may insert an invalid value in the discriminator column or the correct subclass records do not exist,
- 2) INSERT to any subclass table—an incorrect subclass record may be added,
- 3) DELETE from any subclass table—a required subclass record may be removed,
- 4) UPDATE on any subclass table—the reference to a superclass record may be changed incorrectly.

Thus, insertions and updates to the table representing the superclass must be checked. An example of such trigger for PostgreSQL is provided in Listing 13.

Listing 13 INSERT/UPDATE trigger on the superclass table realizing a generalization set constraint in PostgreSQL

```
CREATE FUNCTION gs_document_type()
RETURNS TRIGGER AS $$ BEGIN
IF NOT ((
NEW."DISCRIMINATOR" = 'IdCard' AND
NOT EXISTS (SELECT 1 FROM "PASSPORT" p
WHERE p."ID" = NEW."ID") AND
EXISTS (SELECT 1 FROM "ID_CARD" i
WHERE i."ID" = NEW."ID")) OR
(NEW."DISCRIMINATOR" = 'Passport' AND
EXISTS (SELECT 1 FROM "PASSPORT" p
WHERE p."ID" = NEW."ID") AND
NOT EXISTS (SELECT 1 FROM "ID_CARD" i
WHERE i."ID" = NEW."ID"))) THEN
RAISE EXCEPTION '...';
END IF; RETURN NEW; END; $$ LANGUAGE plpgsql;
CREATE TRIGGER gs_document_type
BEFORE INSERT OR UPDATE ON "DOCUMENT" FOR EACH ROW
EXECUTE FUNCTION gs_document_type();
```

As MSSQL does not support row-level triggers, in this case the trigger must be adjusted to check all inserted or updated records by querying the inserted table; see Listing 14.

Listing 14 INSERT/UPDATE trigger on the superclass table realizing a generalization set constraint in MSSQL

```
CREATE TRIGGER [GS_DOCUMENT_TYPE]
ON [DOCUMENT] AFTER INSERT, UPDATE AS BEGIN
IF EXISTS (SELECT 1 FROM inserted i WHERE NOT ((
    i.[DISCRIMINATOR] = 'IdCard' AND
    NOT EXISTS (SELECT 1 FROM [PASSPORT] p
        WHERE p.[ID] = i.[ID]) AND
    EXISTS (SELECT 1 FROM [ID_CARD] i2
        WHERE i2.[ID] = i.[ID])) OR
    (i.[DISCRIMINATOR] = 'Passport' AND
    EXISTS (SELECT 1 FROM [PASSPORT] p
        WHERE p.[ID] = i.[ID]) AND
    NOT EXISTS (SELECT 1 FROM [ID_CARD] i2
        WHERE i2.[ID] = i.[ID])))) BEGIN
    THROW 50000, '...', 1; END; END;
```

As discussed in Subsection IV-A, the prevention of orphan records requires inserting the referencing record first. However, as MySQL does not support deferrable constraint checking, a stored procedure that the user must execute manually has to be used instead. The procedure is executed after the INSERT and UPDATE operations on the superclass table, but only after the referencing record is inserted (if applicable); see Listing 15 (where the ID input parameter shall contain the primary key of the inserted or updated record).

Listing 15 Procedure for checking the INSERT/UPDATE operation that realizes a generalization set constraint in MySQL

```
CREATE PROCEDURE `GS_DOCUMENT_TYPE` (IN `ID` INT)
BEGIN
SET @discr = (SELECT `DISCRIMINATOR`
FROM `DOCUMENT` d WHERE d.`ID` = `ID`);
IF NOT ((
    @discr = 'IdCard' AND
    NOT EXISTS (SELECT 1 FROM `PASSPORT` p
        WHERE p.`ID` = `ID`) AND
    EXISTS (SELECT 1 FROM `ID_CARD` i
        WHERE i.`ID` = `ID`) OR
    (@discr = 'Passport' AND
    EXISTS (SELECT 1 FROM `PASSPORT` p
        WHERE p.`ID` = `ID`) AND
    NOT EXISTS (SELECT 1 FROM `ID_CARD` i
        WHERE i.`ID` = `ID`))) THEN
    SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

Next, we argue that for PostgreSQL and MSSQL, the insertion or deletion of a record that references an existing record in the superclass table in any subclass table breaks the generalization set constraint, since the subclass record is inserted before the superclass record and deleted after the superclass record. Therefore, INSERT and DELETE triggers must be employed that prevent the operation if the inserted or deleted record references an existing record in the superclass table. Examples of the INSERT and DELETE triggers for PostgreSQL are provided in Listings 16 and 17, respectively. Again, in MSSQL, due to the lack of support for row-level triggers, the inserted and deleted special tables must be queried.

Listing 16 INSERT trigger on a subclass table realizing a generalization set constraint in PostgreSQL

```
CREATE FUNCTION gs_document_type_passport_ins()
RETURNS TRIGGER AS $$ BEGIN
IF EXISTS (SELECT 1 FROM "DOCUMENT" d
    WHERE d."ID" = NEW."ID") THEN
    RAISE EXCEPTION '...';
END IF; RETURN NEW; END; $$ LANGUAGE plpgsql;
CREATE TRIGGER gs_document_type_passport_ins
BEFORE INSERT ON "PASSPORT" FOR EACH ROW
EXECUTE FUNCTION gs_document_type_passport_ins();
```

Listing 17 DELETE trigger on a subclass table realizing a generalization set constraint in PostgreSQL

```
CREATE FUNCTION gs_document_type_passport_del()
RETURNS TRIGGER AS $$ BEGIN
IF EXISTS (SELECT 1 FROM "DOCUMENT" d
    WHERE d."ID" = OLD."ID") THEN
    RAISE EXCEPTION '...';
END IF; RETURN OLD; END; $$ LANGUAGE plpgsql;
CREATE TRIGGER gs_document_type_passport_del
BEFORE DELETE ON "PASSPORT" FOR EACH ROW
EXECUTE FUNCTION gs_document_type_passport_del();
```

In MySQL, as discussed in Subsection IV-A, the order of insertion is reversed. Thus, the check for the existence of the correct subclass records is performed in the superclass INSERT trigger, and the subclass INSERT trigger needs to check only if the discriminator value corresponds to the subclass table that is being inserted to. It is not necessary to check that no other record in the same table references the same superclass record, as that is prevented by the PRIMARY KEY constraint on the referencing column.

Concerning the DELETE operation on subclass tables, due to the lack of support for deferrable constraints in MySQL, a subclass record must be deleted before the superclass one. The possibility of leaving a superclass record without a subclass record is impossible to prevent automatically via triggers in MySQL, and can only be checked manually by executing a stored procedure after the subclass and superclass (if applicable) records are deleted; see Listing 18 (where the ID column shall contain the primary key of the deleted record).

Listing 18 Procedure after DELETE on a subclass table realizing a generalization set constraint in MySQL

```
CREATE PROCEDURE `GS_DOCUMENT_TYPE_DEL` (IN `ID` INT)
BEGIN
IF EXISTS (SELECT 1 FROM `DOCUMENT` d
    WHERE d.`ID` = `ID`) THEN
    SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

Finally, the UPDATE operation on the subclass table must not de-associate the record from an existing superclass record, nor associate it with another existing record. In MySQL and PostgreSQL, this can be ensured by an UPDATE trigger on the subclass tables; see Listing 19 for MySQL, implementation in PostgreSQL is similar.

MSSQL does not support row-level triggers, and changes to individual records may only be inferred by comparing

Listing 19 UPDATE trigger on a subclass table realizing a generalization set constraint in MySQL

```
CREATE TRIGGER `GS_DOCUMENT_TYPE_PASSPORT_UPD`
BEFORE UPDATE ON `PASSPORT` FOR EACH ROW BEGIN
IF OLD.`ID` <> NEW.`ID` AND EXISTS (
    SELECT 1 FROM `DOCUMENT` d
    WHERE d.`ID` = OLD.`ID`
    OR d.`ID` = NEW.`ID`) THEN
    SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

the records in the inserted and deleted special tables. However, these tables do not contain a column that would make detecting the change of the primary key possible. As a result, in MSSQL, the UPDATE operation on subclass tables can only be checked manually by executing a stored procedure after updating the subclass record; see Listing 20 (where the @OLD_ID and @NEW_ID input parameters contain the previous and new values of the primary key, respectively).

Listing 20 Procedure after UPDATE on a subclass table realizing a generalization set constraint in MSSQL

```
CREATE PROCEDURE [GS_DOCUMENT_TYPE_PASSPORT_UPD]
(@OLD_ID INT, @NEW_ID INT) AS BEGIN
IF @OLD_ID <> @NEW_ID AND EXISTS (
    SELECT 1 FROM [DOCUMENT] d
    WHERE d.[ID] = @OLD_ID
    OR d.[ID] = @NEW_ID) BEGIN
    THROW 50000, '...', 3;
END; END;
```

E. Mandatory and Special Multiplicity

The *mandatory multiplicity* and *special multiplicity* constraints impose a restriction on the number of referencing records for a particular referenced record. It follows that five DML operations that may violate the constraint exist:

- 1) INSERT to the referenced table—an orphaned record may be inserted,
- 2) UPDATE on the referenced table—an incorrect number of references may point to a record with changed primary key,
- 3) INSERT or DELETE on the referencing table—an incorrect number of records may point to the referenced record,
- 4) UPDATE on the referencing table—an incorrect number of record may reference the previous and current referenced record.

In MySQL, the referenced record must be inserted first. Therefore, there always exists a possibility of creating an orphan record that may violate the constraint. As a consequence, the constraint may only be checked manually by executing a stored procedure. Furthermore, as MySQL does not support statement-level triggers, all triggers on the referencing table would produce a mutating table error. Consequently, a stored procedure must be used in place of triggers as well. Collectively, a single stored procedure needs to be executed (i) after inserting to the referenced table, but only after the referencing

record is inserted (if applicable), (ii) after updating or deleting from the referenced table, (iii) after inserting, updating or deleting from the referencing table; see Listing 21.

Listing 21 Procedure realizing a special multiplicity constraint in MySQL

```
CREATE PROCEDURE `MUL_PASSP_PERSON_ID` () BEGIN
IF EXISTS (SELECT 1 FROM `PERSON` p WHERE NOT (
    SELECT COUNT(1) FROM `PASSPORT` p2
    WHERE p2.`PERSON_ID` = p.`ID`
    ) BETWEEN 2 AND 4) THEN
    SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = "...";
END IF; END;
```

In the case of PostgreSQL and MSSQL, automatic checking via triggers can be implemented. A trigger on the INSERT and UPDATE operations on the referenced table prevents the operation if an incorrect number of referencing records is detected for the affected records; see Listings 22 and 23 for PostgreSQL and MSSQL, respectively.

Listing 22 INSERT/UPDATE trigger on the referenced table realizing a special multiplicity constraint in PostgreSQL

```
CREATE FUNCTION mul_passp_person_id()
RETURNS TRIGGER AS $$ BEGIN
IF NOT ((
    SELECT COUNT(1) FROM "PASSPORT" p
    WHERE p."ID" = NEW."ID"
    ) BETWEEN 2 AND 4) THEN
    RAISE EXCEPTION '...';
END IF; RETURN NEW; END; $$ LANGUAGE plpgsql;
CREATE TRIGGER mul_passp_person_id
BEFORE INSERT OR UPDATE ON "PERSON" FOR EACH ROW
EXECUTE FUNCTION mul_passp_person_id();
```

Listing 23 INSERT/UPDATE trigger on the referenced table realizing a special multiplicity constraint in MSSQL

```
CREATE TRIGGER [MUL_PASSP_PERSON_ID]
ON [PERSON] AFTER INSERT, UPDATE AS BEGIN
IF EXISTS (SELECT 1 FROM inserted i WHERE NOT ((
    SELECT COUNT(1) FROM [PASSPORT] p
    WHERE p.[ID] = i.[ID]
    ) BETWEEN 2 AND 4)) BEGIN
    THROW 50000, '...', 1; END; END;
```

The INSERT, UPDATE and DELETE operations are checked by a second trigger that prevents the operation if after the operation is executed a record with an incorrect number of referencing records exists. See Listing 24 for an example in MSSQL, implementation in PostgreSQL is similar.

Listing 24 INSERT/UPDATE/DELETE trigger on the referencing table realizing a special multiplicity constraint in MSSQL

```
CREATE TRIGGER [MUL_PASSP_PERSON_ID_REL]
ON [PASSPORT] AFTER INSERT, UPDATE, DELETE AS BEGIN
IF EXISTS (SELECT 1 FROM [PERSON] p WHERE NOT ((
    SELECT COUNT(1) FROM [PASSPORT] p2
    WHERE p2.[ID] = p.[ID]
    ) BETWEEN 2 AND 4)) BEGIN
    THROW 50000, '...', 2; END; END;
```

In Listings 21–24, we presented the realization of a special multiplicity of $[2..4]$. For constraints where the lower or upper bound is unrestricted, a $\leq$ or $\geq$ operator is used instead of the `BETWEEN` operator.

The mandatory multiplicity constraint is a special case of the special multiplicity constraint, where the lower bound is 1 and the upper bound is unrestricted. Therefore, its realization is the same as of the special multiplicity constraint, possibly with the `EXISTS` predicate instead of $\geq$ and the `COUNT` function for improved performance.

V. DISCUSSION

In this paper, we presented our approach to the enforcement of integrity constraints of an OntoUML model in relational databases, in particular PostgreSQL, Microsoft SQL Server and MySQL.

Our research appears to suggest that among the three considered RDBMS, PostgreSQL is the most suitable for the implementation of constraints of an OntoUML model. Due to the lack of support for certain features concerning triggers and the absence of deferrable constraint checking, the realization in MSSQL and MySQL at times falls back to user-executed validation procedures.

Furthermore, due to triggers being processed after each DML operation, the implementation of constraints that validate data across multiple tables is difficult. In our approach, we work around these shortcomings by (i) requiring all `FOREIGN KEY` constraints to be deferred in PostgreSQL and inserting the referencing record prior to the referenced record, (ii) in MSSQL, due to lack of support for deferrable constraints, temporarily disabling `FOREIGN KEY` constraints and re-enabling them after both records are inserted, (iii) as MySQL does not re-check referential integrity when constraints are re-enabled, we opted to prioritize referential integrity and implemented many of the constraints as user-executed validation procedures instead of automatically executed triggers. As the implementation in PostgreSQL and MSSQL requires a referencing record to be inserted before the referenced record, automatically generated primary keys cannot be used in the referenced table. On the other hand, as the implementation in MySQL relies heavily on manually-executed procedures, the enforcement of constraints is not guaranteed.

VI. CONCLUSIONS

In this paper, we introduced our approach to the transformation of an OntoUML model to a relational database, with focus on the preservation of integrity constraints defined by the OntoUML model. We discussed the realization in the following RDBMS:

- MySQL
- Microsoft SQL Server
- PostgreSQL

We drew on existing research that divides the transformation into three separate steps, as listed below. The original contribution of our paper lies in the third step, which elaborates the implementation in the aforementioned RDBMS.

- 1) First, the source OntoUML model is transformed to an UML model.
- 2) Then, the intermediate UML model is transformed to a platform-independent model of a relational database.
- 3) Finally, the relational model is transformed to a vendor-specific realization in SQL.

We showed the transformation of all possible constraints defined by the constructs used in OntoUML models. We demonstrated the differences between the three considered RDBMS and discussed the limitations of each one w.r.t. the implementation of constraints. The results indicate that PostgreSQL allows implementing complete and automatic constraint enforcement, while MySQL lacks certain needed features and many constraints need to be checked manually.

REFERENCES

- [1] M. Fowler, *UML Distilled: A Brief Guide to the Standard Object Modeling Language*, 3rd ed. Boston: Addison-Wesley, Sep. 2003. ISBN 978-0-321-19368-1
- [2] G. Guizzardi, A. Botti Benevides, C. Morais Fonseca, D. Porello, J. P. A. Almeida, and T. Prince Sales, “UFO: Unified Foundational Ontology,” *Applied Ontology*, vol. 17, no. 1, pp. 167–210, Mar. 2022. doi: 10.3233/AO-210256
- [3] DB-Engines. Ranking per database model category. Accessed Apr. 2025. [Online]. Available: https://db-engines.com/en/ranking_categories
- [4] Z. Rybala, “Towards OntoUML for Software Engineering: Transformation of OntoUML into Relational Databases,” Ph.D. dissertation, Czech Technical University in Prague, Prague, Aug. 2017.
- [5] G. Guizzardi, G. Wagner, J. P. A. Almeida, and R. S. Guizzardi, “Towards Ontological Foundations for Conceptual Modeling: The Unified Foundational Ontology (UFO) Story,” *Applied Ontology*, vol. 10, no. 3–4, pp. 259–271, Dec. 2015. doi: 10.3233/AO-150157
- [6] G. Guizzardi, “Ontological Foundations for Structural Conceptual Models,” Ph.D. dissertation, University of Twente, Enschede, 2005.
- [7] G. Guizzardi, C. Morais Fonseca, A. B. Benevides, J. P. A. Almeida, D. Porello, and T. P. Sales, “Endurant Types in Ontology-Driven Conceptual Modeling: Towards OntoUML 2.0,” in *Conceptual Modeling*. Cham: Springer, 2018, vol. 11157, pp. 136–150.
- [8] G. Guizzardi, C. Morais Fonseca, J. P. A. Almeida, T. P. Sales, A. B. Benevides, and D. Porello, “Types and Taxonomic Structures in Conceptual Modeling: A Novel Ontological Theory and Engineering Support,” *Data & Knowledge Engineering*, vol. 134, p. 101891, Jul. 2021. doi: 10.1016/j.datak.2021.101891
- [9] G. Guizzardi and G. Wagner, “What’s in a Relationship: An Ontological Analysis,” in *Conceptual Modeling - ER 2008*. Berlin: Springer, 2008, vol. 5231, pp. 83–97. ISBN 978-3-540-87877-3
- [10] G. Guizzardi, T. P. Sales, J. P. A. Almeida, and G. Poels, “Automated Conceptual Model Clustering: A Relator-Centric Approach,” *Software and Systems Modeling*, vol. 21, no. 4, pp. 1363–1387, Aug. 2022. doi: 10.1007/s10270-021-00919-5
- [11] L. Liu and M. T. Özsu, *Encyclopedia of Database Systems*. New York: Springer, 2018. ISBN 978-1-4614-8265-9
- [12] I. O. for Standardization, “Database Language SQL — Part 1: Framework,” Geneva, Jun. 2023.
- [13] J. Jabůrek, “Implementation of the Transformation of an OntoUML Model in OpenPonk into Its Realization in a Relational Database,” Master’s thesis, Czech Technical University in Prague, Prague, May 2024.
- [14] Z. Rybala and R. Pergl, “Towards OntoUML for Software Engineering: Transformation of Kinds and Subkinds into Relational Databases,” *Computer Science and Information Systems*, vol. 14, no. 3, pp. 913–937, 2017. doi: 10.2298/CSIS170109035R
- [15] DB-Engines. Ranking of Relational DBMS. Accessed Apr. 2025. [Online]. Available: <https://db-engines.com/en/ranking relational+dbms>
- [16] Z. Rybala and R. Pergl, “Towards OntoUML for Software Engineering: Transformation of Anti-Rigid Sortal Types into Relational Databases,” in *Model and Data Engineering*. Cham: Springer, 2016, vol. 9893, pp. 1–15. ISBN 978-3-319-45547-1

A Framework for Machine Unlearning Using Selective Knowledge Distillation into Soft Decision Tree

Sangmin Kim
Chung-Ang University,
84, Heukseok-ro, Dongjak-gu,
06974 Seoul, South Korea
Email: kimddol98@cau.ac.kr

Byeongcheon Lee
Chung-Ang University,
84, Heukseok-ro, Dongjak-gu,
06974 Seoul, South Korea
Email: qudcjs0208@cau.ac.kr

Sungwoo Park
Chung-Ang University,
84, Heukseok-ro, Dongjak-gu,
06974 Seoul, South Korea
Email: psw5574@cau.ac.kr

Miyoung Lee
Chung-Ang University,
84, Heukseok-ro, Dongjak-gu,
06974 Seoul, South Korea
Email: miylee@cau.ac.kr

Seungmin Rho
Chung-Ang University,
84, Heukseok-ro, Dongjak-gu,
06974 Seoul, South Korea
Email: smrho@cau.ac.kr

Abstract—With growing privacy regulations, removing user-related information from machine learning models has become essential. Machine unlearning addresses this by enabling selective removal of learned information, but most existing methods rely on deep learning models, which are computationally expensive and lack interpretability. To overcome these limitations, we propose a novel machine unlearning framework using selective knowledge distillation into a Soft Decision Tree (SDT). A convolutional neural network (ConvNet) is first trained to generate soft labels and intermediate features, which are transferred to the SDT. During distillation, an unlearning algorithm adjusts specific leaf node distributions and routing weights using soft redistribution and path pruning. This enables class-specific forgetting without retraining and preserves accuracy on non-target classes. Experiments on MNIST and CIFAR-10 demonstrate that our framework effectively removes class-specific knowledge while maintaining overall model performance. The interpretable SDT structure also allows for clear visualization of model changes before and after unlearning.

Index Terms—Machine Unlearning, Knowledge Distillation, Convolutional Neural Network, Soft Decision Tree, Privacy.

I. INTRODUCTION

IN RECENT years, data protection regulations aimed at protecting user privacy and increasing personal data control have been introduced and enforced around the world. For example, the General Data Protection Regulation (GDPR) in the European Union (EU) [1] and the California Consumer Privacy Act (CCPA) in the United States [2] impose strict legal obligations on organizations to delete personal information when requested by users. These regulations demonstrate the growing social and legal recognition of the “right to be forgotten” and make it clear that individuals can demand the removal of their digital traces. Accordingly, there is a growing need for a technical solution that can effectively reflect an individual's deletion request, even

in already trained machine learning models. As a solution, machine unlearning is gaining much attention [3]. This technology enables models to comply with the latest privacy regulations by selectively removing the impact of specific data that has already been learned.

Most existing machine unlearning approaches have mainly focused on deep learning (DL) models and are typically implemented by retraining the entire model or mitigating the influence of specific data through methods such as gradient ascent or fine-tuning. However, these approaches are costly in terms of computations, and there is a risk of inadvertently affecting non-target data [4, 5]. In such cases, useful learned representations are unintentionally changed, resulting in a reduction of the model's stability and generalizability. In addition, these limitations are further exacerbated in DL models. It is difficult to selectively remove the influence of specific data from a deep learning model's complex and distributed internal representation because even minor adjustments to the representation can unintentionally change the overall result of the model [6]. Furthermore, the opacity of the internal representation complicates the verification of the removal of specific information.

Recently, machine unlearning research on traditional machine learning models has also been actively conducted, especially for tree-based models such as gradient boosting decision tree (GBDT) [7, 8] and random forest (RF) [9, 10, 11]. The advantage of these tree-based models is that the decision process is clear and easy to interpret, and data deletion can be easily verified. However, since these models are primarily designed for discrete, low-dimensional input data, they are not effective for handling high-dimensional continuous data, such as images. In addition, machine unlearning for the tree-based models requires the explicit maintenance of instance-leaf mappings during both training and inference, which can lead to high memory and computational overhead. Furthermore, although tree-based

models are effective for verifying instance-level deletion, they are less suitable for class-level unlearning, where an entire semantic category must be removed. In practical scenarios, such class-level deletion requests frequently occur due to policy revisions, semantic redefinitions, or ethical and legal considerations. Class-level unlearning is gaining attention as a necessary capability in modern machine learning systems, where models are expected to respond to changing category definitions and increasing regulatory demands [12].

To address these limitations, we propose a machine unlearning framework using selective knowledge distillation that can preserve the generalization performance of the original DL model. In the proposed framework, knowledge containing soft labels and intermediate features is extracted from a high-performing convolutional neural network (ConvNet) and transferred to a soft decision tree (SDT). In addition, the proposed unlearning algorithm adjusts the routing probabilities and class distributions of leaf nodes in SDT to selectively suppress information related to the target class while preserving overall model accuracy. The main contributions of this study are as follows:

- Instead of retraining the entire model, we propose an efficient machine unlearning framework that selectively adjusts only routing probabilities and class distributions of specific leaf nodes in SDT. This enables the selective removal of information on target classes while maintaining stable classification performance on non-target classes.
- By utilizing the tree-based structure in which both the branching decisions and leaf node class distributions are explicitly interpretable, the proposed framework enables visual analysis of model changes before and after unlearning.
- The proposed framework was evaluated using multiple image benchmark datasets MNIST and CiFAR-10, and the results showed robust unlearning effectiveness and generalization performance.

II. PRELIMINARIES

A. Machine Unlearning

Machine unlearning refers to the process of removing or weakening the influence of specific training data within a trained model. A learning algorithm is formally defined as a function $A: D \rightarrow H$, where D is a dataset and H is a hypothesis space. In summary, the algorithm A returns a model $A(D) \in H$, which is trained on the dataset D . Unlearning is performed through a removal $R: (A(D), D, (x, y)) \rightarrow H$, which takes as inputs the trained model $A(D)$, the original dataset D , and a data instance (x, y) to remove. Exact machine unlearning refers to the ideal scenario in which the resulting model is indistinguishable from one trained from scratch on the dataset

with (x, y) removed [13]. This condition is formally expressed as:

$$R(A(D), D, (x, y)) = A(D \setminus \{(x, y)\}) \quad (1)$$

B. Knowledge Distillation

Knowledge distillation is a technique for model compression and optimization that transfers the knowledge learned by a large and complex teacher model to a small and simple student model so that the student model performs as well as the teacher model [14, 15]. Typically, knowledge distillation utilizes soft labels (probability distributions for each class) produced by the teacher model to train the student model. Soft labels contain more information than hard labels (1 for the correct class and 0 for the rest), such as similarity relationships and uncertainty between classes. In order to leverage additional information from the output distribution of the teacher model, a temperature parameter T is incorporated into the softmax function to smooth the output distribution and emphasize class similarity. The smoothed probability for class i , denoted as q_i , is calculated as follows:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (2)$$

where z_i is the logit for class i , and T controls the smoothness of the distribution. A higher temperature leads to softer probability outputs, allowing the student to capture the teacher's nuanced generalization behavior.

III. METHODOLOGY

A. Datasets

In this study, we utilize two widely used benchmark image classification datasets, MNIST [16] and CIFAR-10 [17], in our experiment. The MNIST dataset consists of grayscale images, while CIFAR-10 dataset is composed of color images. Each dataset is divided into train, validation, and test sets to facilitate both model training and evaluation. The number of samples in each subset is summarized in Table I and an example of each dataset is shown in Fig. 1.

TABLE I
CONSTRUCTION OF DATASETS

	Datasets	
	MNIST	CIFAR-10
Train	50,000	40,000
Validation	10,000	10,000
Test	10,000	10,000



Fig 1. Example of MNIST and CIFAR-10 dataset.

To ensure consistent input scaling and improve training stability, we apply different preprocessing strategies for each dataset. For the MNIST dataset, we apply Min-Max normalization to scale pixel values to the $[0, 1]$ range. For the CIFAR-10 dataset, we perform Z-score normalization by subtracting the mean and dividing by the standard deviation of each RGB channel. As a final preprocessing step, all target labels are converted into one-hot encoded vectors to support the use of the cross-entropy loss function and enable soft label distillation in subsequent stages.

B. Model Design

This section presents the architecture and training procedures of the two main components in our framework: a ConvNet that acts as the teacher model and SDT is trained as a student model through knowledge distillation.

a) ConvNet: To generate soft labels for knowledge distillation, we design two ConvNet architectures optimized respectively for the MNIST dataset and the CIFAR-10 dataset.

For the MNIST dataset, we implement a compact ConvNet composed of two convolutional blocks. Each block consists of two 3×3 convolutional layers with ReLU activation and the same padding, followed by 2×2 max pooling and dropout. The extracted features are then flattened and passed through a fully connected layer with 512 units, followed by a softmax output layer. The model is trained using the Adam optimizer with a learning rate of 3×10^{-4} , and categorical cross-entropy is used as the loss function.

For the CIFAR-10 dataset, we adopt a deeper ConvNet consisting of three convolutional blocks. Each block contains two 3×3 convolutional layers with ReLU activation, batch normalization, L2 regularization, max pooling, and dropout. The number of filters increases with depth to capture complex spatial features. The final feature map is flattened and passed through a dense layer with 128 units and dropout, followed by a softmax classification layer. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-3} , along with learning rate scheduling and data augmentation via the ImageDataGenerator framework to enhance generalization.

b) SDT: The SDT is designed to replicate the predictive behavior of a ConvNet trained within this framework while offering a transparent decision-making process based on explicit branching rules [18]. The operational structure of the SDT is illustrated in Fig. 2, which depicts a simple configuration consisting of one internal node and two leaf nodes.

We design two SDT configurations optimized respectively for the MNIST and CIFAR-10 datasets. For the MNIST dataset, which consists of grayscale images, the output from the final convolutional layer of the ConvNet is flattened and used as input to the SDT. In contrast, for the CIFAR-10 dataset, which contains color images, simple flattening may

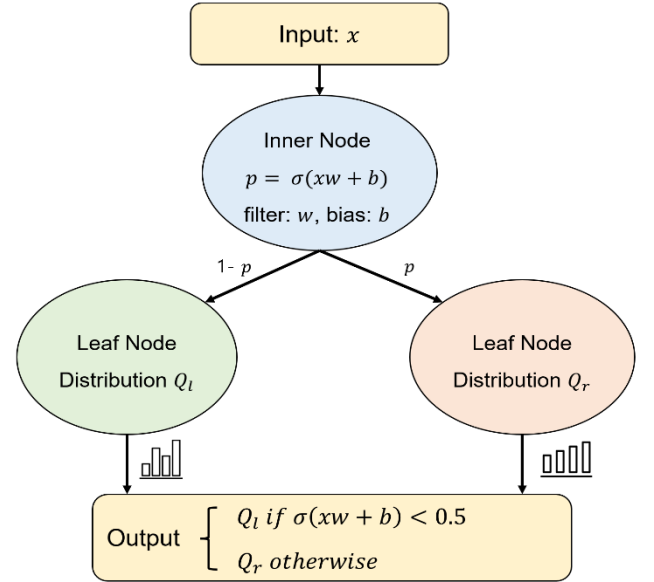


Fig 2. Simple configuration of SDT.

lead to loss of spatial information. To address this, the output of the final max-pooling layer is passed through global average pooling to produce a semantically meaningful one-dimensional feature vector. These feature vectors, along with soft labels generated by the ConvNet, are used to train the SDT. To further enhance representational capacity and training stability, we introduce four key hyperparameters:

- **Penalty strength:** Controls the strength of a regularization term that prevents internal nodes from consistently branching in the same direction. Higher values promote more balanced left-right splits, which increases structural diversity and tree utilization.
- **Penalty decay:** Gradually reduces the influence of the regularization term as the node depth increases. This allows shallow nodes to learn general decision rules, while deeper nodes specialize in more fine-grained distinctions.
- **Exponential moving average (EMA) window size:** The branching direction of internal nodes may fluctuate across mini-batches, which can potentially destabilize learning. We apply an EMA to smooth routing behavior over time. A larger window emphasizes long-term stability, while a smaller window allows quicker adaptation at the cost of higher variance.
- **Inverse temperature (β):** Adjusts the sharpness of the sigmoid function used at internal nodes to make branching decisions more decisive. A higher β makes splits more decisive, with probabilities closer to 0 or 1, making the model behave more like a hard decision tree.

Given the sensitivity of these hyperparameters to model performance, we use the Optuna framework for hyperparameter optimization [19]. The search space includes tree depths and learning rate in addition to the four parameters above. The optimal configuration is selected based on

validation accuracy. As a result, the SDT trained under this framework successfully emulates the predictions of the teacher model while maintaining interpretability through its hierarchical and transparent decision structure.

C. Unlearning Algorithm

In this study, we propose a novel unlearning algorithm designed to selectively remove knowledge relevant to a particular class from a trained SDT model. The method utilizes two primary mechanisms, soft redistribution and path pruning, to locally adjust a subset of model parameters, enabling class-level forgetting while preserving the overall structure and performance of the model. The proposed unlearning algorithm is presented in Algorithm 1.

The process of the algorithm is described as follows:

- 1) Identify the target leaf node: Find leaf nodes whose predicted class distribution is dominated by the target class c to be forgotten.
- 2) Redistribute class probabilities: Take a fraction (α) of the class c probability of the target leaf node and distribute it to the top k leaf nodes with high cosine similarity. The proportion of the distribution is proportional to the similarity and is only distributed to leaves that do not strongly predict the target class.
- 3) Soft Pruning: The internal node weights of the decision path to each target leaf are attenuated according to the remaining class c probability (ω) and pruning ratio (γ).
- 4) Residual Suppression: If class c still has the highest probability in all leaves, set its probability to zero and re-normalize the distribution.

IV. EXPERIMENT

This section presents the empirical evaluation of the proposed framework. For each dataset, we present the classification performance when performing general knowledge distillation and the performance of the proposed framework.

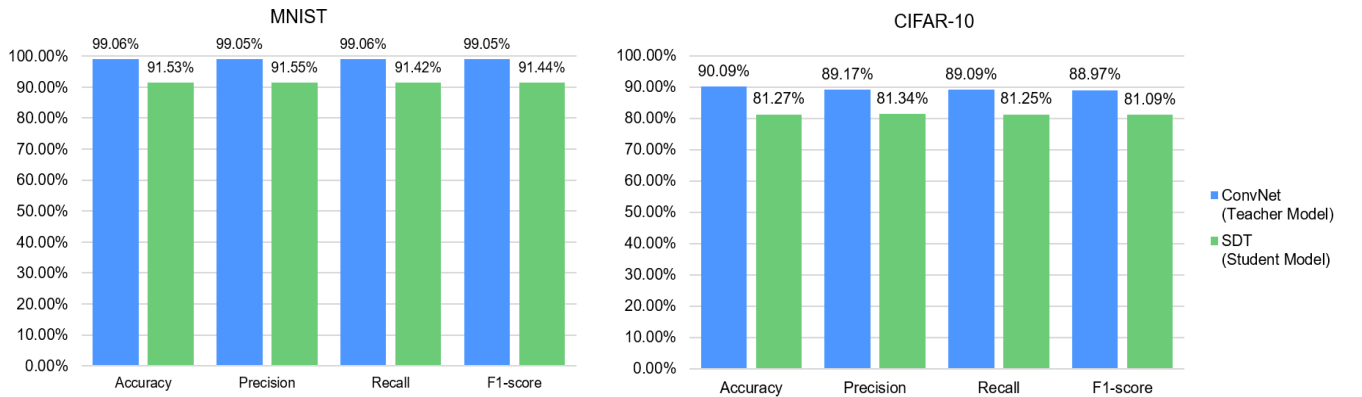


Fig 3. Performance of ConvNet and SDT.

Algorithm 1 Target Class Unlearning

Require: Trained tree model T , session S , target class c , top- k , redistribution rate α , pruning rate γ
Ensure: Modified tree with target class c forgotten

- 1: Retrieve all leaf nodes L and obtain $\phi_\ell \leftarrow S(\phi_\ell)$ for each $\ell \in L$
- 2: $L_{\text{target}} \leftarrow \{\ell \in L \mid \text{argmax}(\phi_\ell) = c\}$
- 3: **for all** $\ell_t \in L_{\text{target}}$ **do**
- 4: Find top- k most similar leaves $N(\ell_t)$ using cosine similarity
- 5: $\delta \leftarrow \phi_{\ell_t}[c] \cdot \alpha$
- 6: $\phi_{\ell_t}[c] \leftarrow \phi_{\ell_t}[c] \cdot (1 - \alpha)$
- 7: **for all** $\ell_j \in N(\ell_t)$ **do**
- 8: $\phi_{\ell_j}[c] \leftarrow \phi_{\ell_j}[c] + \delta \cdot \text{sim}(\ell_t, \ell_j)$
- 9: Normalize ϕ_{ℓ_j} and update in session
- 10: **end for**
- 11: Normalize ϕ_{ℓ_t} and update in session
- 12: Compute importance $\omega \leftarrow \phi_{\ell_t}[c]$
- 13: Compute pruning factor $\rho \leftarrow 1 - \gamma \cdot \omega$
- 14: Identify internal nodes P on the path to ℓ_t
- 15: **for all** $n \in P$ **do**
- 16: Weaken weight: $w_n \leftarrow \rho \cdot w_n$
- 17: **end for**
- 18: **end for**
- 19: **for all** $\ell \in L$ **do**
- 20: **if** $\phi_\ell[c] > \epsilon$ **then** ▷ if target class remains dominant
- 21: Set $\phi_\ell[c] \leftarrow 0$, normalize, and update in session
- 22: **end if**
- 23: **end for**

A. Experiment Setup

All experiments were conducted on a system running Windows 11, equipped with an Intel Core i7-12700K CPU, an NVIDIA GeForce RTX 4080 GPU (16GB), and 64 GB RAM. Python 3.7 was used as the development environment.

We evaluated the proposed framework on two benchmark image classification datasets: MNIST and CIFAR-10. As described in Section 3, each dataset was preprocessed using standard procedures. Both ConvNet and SDT model were trained and evaluated based on accuracy, precision, recall, and F1-score. To ensure fair and stable comparisons, all training processes incorporated early stopping to prevent overfitting and reduce unnecessary training time. For SDT models after unlearning, we conducted a focused evaluation comparing the

TABLE II
PERFORMANCE AND TRAINING TIME COMPARISON BETWEEN PROPOSED FRAMEWORK AND RETRAINING

Datasets	Method	Test Accuracy	Target Accuracy	Non-Target Accuracy	Time
MNIST	Proposed Framework	80.74%	0%	89.81%	33m
	Retrain	80.62%		89.68%	51m 15s
CiFAR-10	Proposed Framework	71.12%	0%	79.02%	44m 23s
	Retrain	71.51%		79.46%	96m 20s

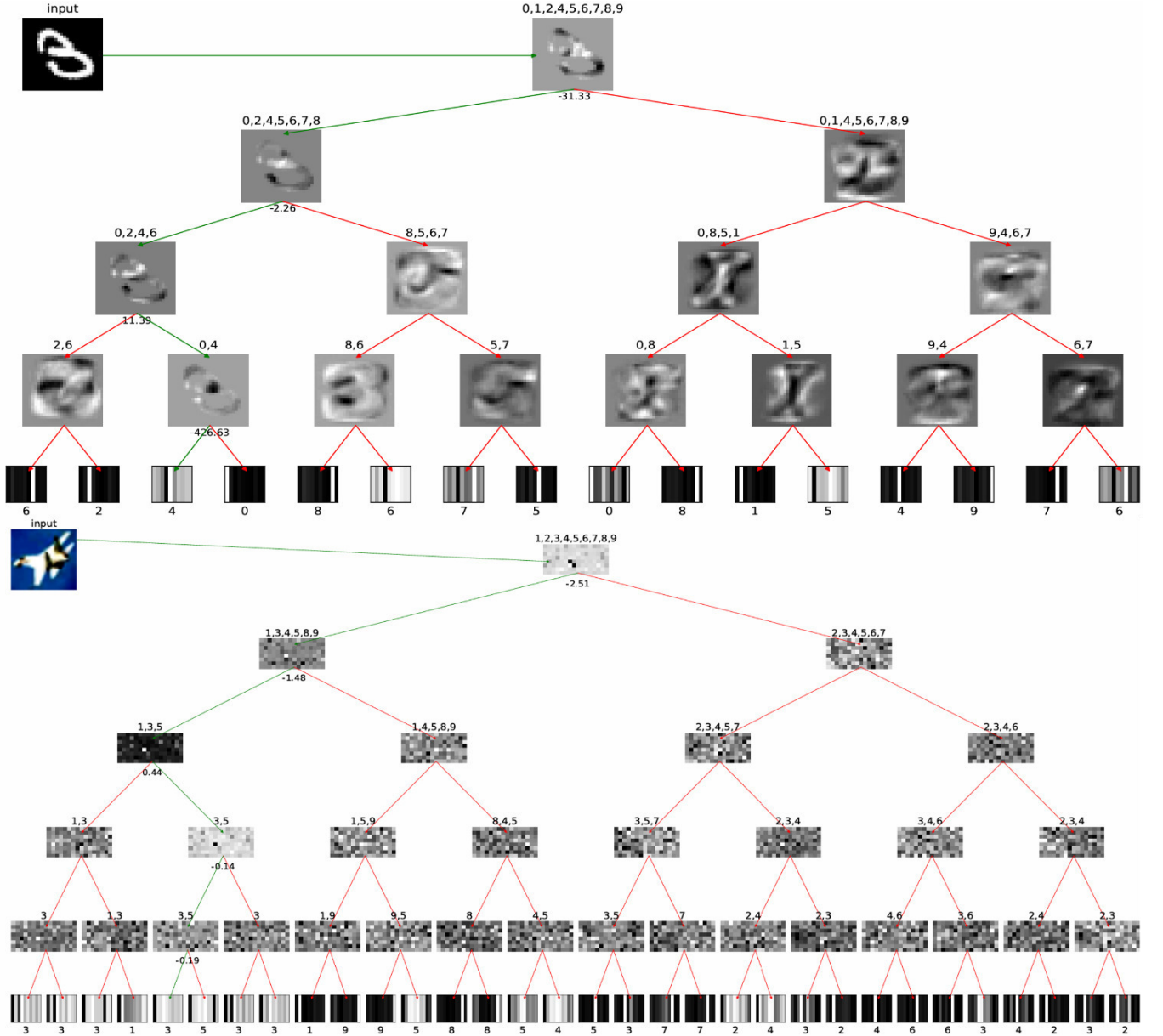


Fig 4. SDT after unlearning: (Top) MNIST dataset (target class 3), (Bottom) CIFAR-10 dataset (target class 0).

baseline approach of retraining after removing the target class with the proposed framework. In both cases, we analyzed the accuracy of target and non-target classes to assess the effectiveness of selective forgetting.

B. Performance Evaluation of Knowledge Distillation

To verify the suitability of ConvNet as the teacher model and to assess the effectiveness of knowledge distillation to the

student model (SDT), we compared the classification performance of both models on MNIST and CIFAR-10 using four metrics: accuracy, precision, recall, and F1-score. Detailed metric comparisons for each dataset are visualized in Fig. 3. As shown in Fig. 3, the ConvNet achieved very high

performance, with accuracy, precision, recall, and F1-score all exceeding 99% on MNIST and over 90% on CIFAR-10, confirming its appropriateness as a teacher model. In comparison, the SDT showed a consistent performance gap, with accuracy and other metrics approximately 8% lower than

those of ConvNet on both datasets. While this indicates some loss in predictive power due to the simpler and more interpretable structure of the SDT, the results still demonstrate that SDT can inherit a significant portion of the teacher model's knowledge through distillation.

C. Performance Evaluation of Proposed Framework

To evaluate the effectiveness of the proposed unlearning framework, we applied the algorithm described in Section 3 to SDT models distilled from teacher models trained on each dataset. For each dataset, a specific target class was selected for removal, and we assessed how effectively the SDT could forget the target class while preserving predictive performance on non-target classes. As a baseline for comparison, we adopted a conventional retraining approach where all training samples of the target class were removed from the dataset, and then both the teacher and student models were fully retrained from scratch. The evaluation was based on three key metrics: (1) overall classification accuracy on the test set, (2) accuracy on samples belonging to the target class, and (3) accuracy on samples from non-target classes. As summarized in Table II, both methods successfully removed the model's ability to predict the target class, achieving 0% accuracy in all cases. However, the proposed method consistently retained comparable or higher accuracy on non-target classes, demonstrating its effectiveness in preserving useful knowledge.

In addition to predictive performance, our framework offers a substantial advantage in training efficiency. Unlike the baseline, which involves retraining the entire pipeline, our method applies unlearning directly to the distilled student model without reinitializing the teacher. Although the total training time from teacher to student is comparable across some datasets, omitting the costly teacher retraining step significantly reduces overall computational overhead. This makes the proposed framework a practical and efficient solution, particularly in resource-constrained environments. Fig. 4 shows the structural changes in the SDT after unlearning the target classes. The top section presents an MNIST example where digit 3 was selected for removal. After unlearning, the SDT adjusts its internal decision paths, rerouting inputs that were previously associated with digit 3 toward alternative classes. In the visualization, the leaf nodes that initially had high confidence for class 3 now show redistributed probabilities favoring digits such as 4 and 6. Similarly, the bottom section shows a CIFAR-10 case where the airplane (class 0) class was removed. The SDT modifies its structure to suppress branches linked to the removed class and shifts decision confidence toward other classes. These structural adjustments both in internal routing and leaf-level distributions demonstrate how the SDT reflects the unlearning objective at the model level.

II. CONCLUSIONS

This study proposed an interpretable and selective machine unlearning framework based on knowledge distillation

from a ConvNet to an SDT. The SDT is trained using soft labels and intermediate features extracted from the teacher model, providing critical advantages in interpretability and modularity essential for efficient unlearning. To facilitate class-level forgetting, we propose an unlearning algorithm that integrates soft redistribution and path pruning, enabling targeted suppression of class-specific information. Unlike conventional retraining-based methods, our framework applies unlearning directly to the distilled student model without retraining the teacher model, providing a more efficient and practical alternative.

Experimental results on two benchmark image classification datasets, MNIST and CIFAR-10, demonstrate that the proposed framework maintains high classification accuracy while effectively removing the influence of the designated target class. After unlearning, target class accuracy dropped as intended, while performance on non-target classes remained stable, validating the framework's selective forgetting capability. Furthermore, visualizations of the SDT before and after unlearning confirmed structural changes, including rerouted decision paths and updated leaf node distributions that reduced the model's reliance on the target class. By presenting an interpretable unlearning framework applicable to high-dimensional input data, this study contributes to the field of machine unlearning. Future research will focus on expanding the proposed framework in two main directions. First, we aim to extend the current class-level unlearning approach to support instance-level unlearning, enabling more granular control over the forgetting process. This will require developing fine-tuned strategies for identifying and suppressing the influence of individual training samples within the trees structure. Second, we plan to evaluate the scalability and robustness of the framework on more complex and high-resolution image datasets, such as STL-10 or ImageNet. These experiments will test whether the selective forgetting and interpretability properties of the SDT-based student model are preserved under more challenging data conditions. Consequently, these directions aim to enhance the generalizability and practicality of interpretable machine unlearning.

ACKNOWLEDGMENT

This research was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2025-RS2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00518960) and the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government.

REFERENCES

- [1] EU. Regulation (eu) 2016/679. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016. [Online; accessed 16-April-2020].

- [2] California. California consumer privacy act. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, 2018. [Online; accessed 16-April 2020].
- [3] L. Bourtole, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, ... & N. Papernot, "Machine unlearning," In *2021 IEEE symposium on security and privacy (SP)*, pp.141-159, May 2021.
- [4] L. Wang, T. Chen, W. Yuan, X. Zeng, K. F. Wong, & H. Yin, "KGA: A general machine unlearning framework based on knowledge gap alignment," arXiv preprint arXiv:2305.06545.
- [5] A. Sekhari, J. Acharya, G. Kamath, & A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18075-18086, 2021.
- [6] J. Li, Y. Li, X. Xiang, S. T. Xia, S. Dong & Y. Cai, "TNT: An interpretable tree-network-tree learning framework using knowledge distillation," *Entropy*, vol. 22, pp. 1203, 2020.
- [7] H. Lin, J. W. Chung, Y. Lao, & W. Zhao, "Machine unlearning in gradient boosting decision trees," In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.1374-1383, August 2023.
- [8] Z. Wu, J. Zhu, Q. Li, & B. He, "Deltaboost: Gradient boosting decision trees with efficient machine unlearning," *Proceeding of the ACM on Management of Data*, Vol. 1, pp. 1-26, 2023.
- [9] J. Brophy, & D. Lowd, "Machine unlearning for random forests," In *International Conference on Machine Learning*, PMLR, pp.1092-1104, July 2021.
- [10] T. Surve, & R. Pradhan, "Example-based Explanations for Random Forests using Machine Unlearning," arXiv preprint arXiv:2402.05007.
- [11] S. Wang, Z. Shen, X. Qiao, T. Zhang, & M. Zhang, "DynFrS: An Efficient Framework for Machine Unlearning in Random Forest," arXiv preprint arXiv: 2410.01588.
- [12] Z. Zuo, Z. Tang, B. Wang, K. Li, & A. Datta, "Ecil-mu: Embedding based class incremental learning and machine unlearning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6275-6279, 2024.
- [13] J. Xu, Z. Wu, C. Wang, & X. Jia, "Machine unlearning: Solutions and challenges," *IEEE Transactions on Emerging Topics in Computational Intelligence*.
- [14] J. Gou, B. Yu, S. J. Maybank & D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789-1819, 2021.
- [15] G. Hinton, O. Vinyals, & J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
- [16] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, pp.141-142, 2012.
- [17] A. Krizhevsky, & G. Hinton, "Learning multiple layers of features from tiny images," 2009.
- [18] N. Frosst, & G. Hinton, "Distilling a neural network into a soft decision tree," arXiv preprint arXiv:1711.09784, 2017.
- [19] T. Akiba, S. Sano, T. Yanase, T. Ohta, & M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp.2623-2631, July 2019.

Smart Routes: Hybrid Metaheuristics for Efficient Vehicle Routing Problem

Yehor Kovalenko*, Andrei Pivavaru[†] and Joanna Ochelska-Mierzejewska[‡]

Technical University of Lodz

116 Zeromskiego Street, 90-924 Lodz, Poland

*ORCID: 0009-0003-3744-3294

[†]ORCID: 0009-0008-7730-6063

[‡]ORCID: 0000-0002-9295-3962

Email: joanna.ochelska-mierzejewska@p.lodz.pl

Abstract—This article presents a hybrid algorithm developed to solve the Vehicle Routing Problem with Time Windows (VRPTW), which involves finding optimal routes for a fleet of vehicles serving a set of geographically dispersed customers within specified time intervals. The proposed solution combines Ant Colony Optimization (ACO) as the primary method for global solution construction, with the 2-opt local search technique used for route refinement, and a Tabu Search strategy to escape local optima and further improve solution quality.

The algorithm dynamically adapts pheromone levels to favor both spatial and temporal proximity between customers, enhancing decision making during route construction. Experimental results demonstrate that the hybrid approach yields high-quality solutions, significantly improving known results by up to 30% in some cases, while maintaining reasonable computation times. This makes the algorithm well-suited for real-time logistics scenarios where time efficiency and solution accuracy are both critical.

I. INTRODUCTION

FOR MANY years, researchers have explored a wide range of challenges in the field of combinatorial optimization. One of the most thoroughly studied problems in this area is the Traveling Salesman Problem (TSP). Although initially mentioned in the 19-th century by Kirkman and Hamilton, it was not formally defined until the 1930s by Schirjver [1]. The basic version of the TSP involves determining the shortest possible route that visits n cities exactly once and returns to the starting point. As an NP-hard problem, it is computationally infeasible to exhaustively evaluate all permutations for larger instances. Nevertheless, due to its straightforward formulation and relevance to real-world applications, the TSP has become a standard benchmark for evaluating the performance of optimization algorithms [2].

The TSP can be seen as a specific instance of a broader class of problems that focus on finding an optimal route through a set of locations using a single vehicle. This concept is generalized by the Vehicle Routing Problem (VRP), which was formally introduced in 1959 by Dantzig and Ramser. Their pioneering work not only proposed a mathematical model but also applied it to a practical problem involving fuel distribution [3].

VRP is one of the key issues in logistics and supply chain management, having a direct impact on operating costs,

customer service quality and overall business efficiency. In practice, it involves the process of making decisions about when, how and what to use to deliver goods to end users in the most efficient way possible while meeting a number of constraints, such as transport capacity, delivery time windows, driver availability and road conditions [4]. Problems of this type are extremely important not only for transport and logistics companies, but also for commercial, manufacturing and e-commerce organizations.

The VRP seeks to generate a set of vehicle routes that begin and end at a central depot, such that customer demands are met, all operational constraints are satisfied, and the total transportation cost is minimized. Depending on the configuration of the problem - such as the layout of the road network, customer requirements, depot locations, and available vehicles - different variants of the VRP can be defined, each corresponding to specific research or operational challenges [5].

In most formulations, the transportation network is modeled as a graph, where vertices represent depots and customers, and edges represent roads (see Fig. 1). Edges may be directed (for one-way streets) or undirected (for two-way roads), and each edge has an associated cost metric, such as distance, travel time, or monetary cost. Customers are typically described by the following attributes [5]:

- the vertex in the network they are assigned to;
- the amount of goods to be delivered or picked up;
- the time window during which service must occur;
- the service duration (loading or unloading time);
- the set of vehicles eligible to serve them;
- penalties for non-service.

In the standard VRP, only the first two attributes are considered, while the others are included in extended versions of the problem.

To summarize, a solution to the VRP involves determining routes for a fleet of vehicles that start and end at one or more depots (represented as vertices in the graph). These depots are defined not only by their location but also by the number and types of vehicles they host, and the volume of goods available for dispatch. The fleet itself may be homogeneous (identical vehicles) or heterogeneous (vehicles with varying capacities and capabilities), and the fleet size may be fixed

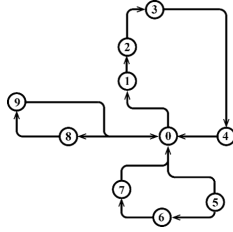


Fig. 1. Graphic representation of VRP

or dynamically adjusted based on customer demand. Vehicles can also have specific features, such as:

- their assigned depot (start and end location);
- capacity constraints (e.g., maximum weight, volume, or number of units);
- compartmentalization for transporting multiple types of goods;
- restricted access to certain parts of the road network;
- operational costs per distance unit, time unit, or network edge.

Section *Literature review* presents a review of the literature on VRP solutions discussed in this article. Section *A Novel Hybrid Algorithm for Solving VRP* discusses the proposal of a new hybrid approach to solving VRP combining the ant algorithm, Tabu Search and 2-opt. Section *Experimental Evaluation and Results* presents the research methodology, discusses the conducted experiments. Section *Evaluation Against Standard Benchmarks* compares the obtained results with the published results and indicates the improvement in the effectiveness of the proposed new hybrid method.

II. LITERATURE REVIEW

For solving VRP, one of the simplest and most commonly used heuristics for local search is the 2-opt algorithm, originally proposed by [6] for the traveling salesman problem (TSP). This method consists in iteratively removing two edges from the route and replacing them with others in such a way as to reduce the cycle length [7]. Despite its simplicity, 2-opt is an effective tool for local improvement of the solution and is often an element of local refinement in more complex approaches. In the context of VRP, 2-opt can be used in both classical and extended versions – e.g. 2-opt* for multiple routes [8]. The method is often combined with other metaheuristic algorithms as part of the intensification phase [9].

Ant Colony Optimization (ACO) is a bioinspired metaheuristic proposed by Dorigo in the 1990s [10]. et al., 1996]. Modeled on the behavior of real ant colonies, the method relies on the simulation of artificial agents (ants) that construct solutions based on probabilistic choices driven by pheromones and local heuristics. Pheromones are reinforced on edges leading to good solutions, which leads to a gradual convergence of the population towards optimal or near-optimal routes.

ACO has been successfully applied to VRP and its numerous variants, including VRPTW (with time windows) and VRP with limited fleet. An example is the work of [11],

who applied a modified ACO to transportation problems in a dynamic environment, but also in [12]. Modifications to ACO include dynamic pheromone evaporation, local search (e.g. 2-opt as an improvement component), and stagnation avoidance mechanisms.

Tabu Search (TS) is a metaheuristic developed by [13] whose main goal is to efficiently search the solution space by avoiding cycles and local minima. TS uses a so-called tabu list — a data structure that stores information about illegal moves for a certain time – to prevent backtracking to recent solutions.

In the context of VRP, Tabu Search has proven to be extremely effective, as evidenced by numerous successful implementations in the literature [15], [14], [16]. This method allows for the flexible consideration of complex constraints such as time windows, payload constraints, or service zone divisions. Many variants of TS for VRP implement advanced aspiration strategies, restart procedures, and adaptive changes to tabu list parameters [17], [18].

In recent years, there has been a growing interest in hybrid approaches that combine the advantages of different methods. An example is the integration of ACO with a local search based on 2-opt, which allows for an increase in the quality of the final solutions [19]. Other works combine ACO with Tabu Search, where ACO generates initial solutions and TS is responsible for their intensification and exploration of new areas of the solution space. Hybrid approaches often achieve better results than single algorithms, especially for large-scale and multi-criteria problems [20], [21], [22], [23].

III. A NOVEL HYBRID ALGORITHM FOR SOLVING VRP

The central component responsible for solving VRP algorithm that initiates the solution search using a hybrid algorithm but also collects statistical data required for performance analysis—such as the globally best-found routes, shortest total distances, and average or worst distances per iteration is proposed in this paper.

The algorithm proceeds as follows:

1) Initialization Phase:

- Set initial best distance and routes.
- Prepare containers for collecting statistics across iterations.

2) Iterative Optimization Loop:

- For each iteration:
 - a) Ant Colony Construction:
 - A colony of artificial ants is instantiated. Each ant constructs a set of feasible routes, attempting to serve all customers starting and ending at the depot.
 - b) Local Improvement via 2-opt:
 - Each route generated by the ants undergoes local optimization using the 2-opt algorithm to reduce path crossings and improve solution quality.
 - c) Best Ant Selection:

- The ant with the shortest total distance in the current iteration is selected. If it improves the global best solution, it is recorded.
- d) Tabu Search Activation:
 - If no improvement is detected for a defined number of iterations (e.g., 10), the algorithm switches to a Tabu Search procedure applied to the best known routes. This mechanism is used to escape local minima.
- e) Pheromone Update:
 - A pheromone evaporation process is applied to all edges.
 - Each ant deposits pheromone proportional to the inverse of its total distance.
 - An additional pheromone bonus is granted to the elite (best) ant.
- 3) Statistics Collection:
 - At the end of each iteration, the best, worst, and average distances are stored to support further analysis.
- 4) Termination and Output:
 - After completing all iterations, the algorithm returns the best-found solution along with its performance metrics.

To improve decision-making during the constructive phase of the solution, the pheromone initialization strategy was enhanced to incorporate problem-specific heuristics. Specifically, the initial pheromone levels between each pair of customers are adjusted to favor geographically closer customers and those with overlapping or adjacent time windows. This is achieved by combining a distance-based component—computed as the inverse of the Euclidean distance—with a time window similarity component—calculated as the inverse of the absolute difference between the customers' ready times. As a result, the algorithm introduces a meaningful prior bias in the search space, guiding ants toward more promising initial configurations while still allowing stochastic exploration..

Once initial solutions are constructed by the ants, each route undergoes refinement using a 2-opt local search procedure. This well-known heuristic systematically explores pairs of edges within a route and checks whether reversing the intermediate segment reduces the total traveled distance. The process is repeated iteratively until no further improvement can be achieved, thereby ensuring that each solution reaches a local minimum with respect to intra-route edge exchanges. Although 2-opt does not guarantee a globally optimal solution, its application significantly reduces path crossings and unnecessary detours, particularly in randomly scattered customer instances.

The underlying algorithm is a hybrid metaheuristic composed of three synergistic components: (1) Ant Colony Optimization (ACO) serves as the base method for generating initial feasible solutions, leveraging pheromone-based probabilistic construction; (2) a local search mechanism based on 2-opt further improves individual routes; and (3) a Tabu Search

algorithm is activated in case of stagnation, helping the system escape local optima by performing controlled, memory-guided explorations of the solution neighborhood. Tabu Search is triggered when no better solution is found over a predefined number of iterations, and it operates by modifying the current best-known solution while preventing reversals through a tabu list.

Each ant in the algorithm is implemented as an object-oriented entity encapsulating the coordinates of the depot, vehicle capacity constraints, a list of customers, pheromone information, and parameters controlling the influence of pheromone trails (α), heuristic desirability (β), and time window urgency (*coef_urgency*). During the route construction process, each ant maintains a dynamically updated list of feasible customers to visit next—those that do not violate capacity or time window constraints. Once no feasible customer remains, the ant returns to the depot, completing a route and potentially starting a new one if unvisited customers remain.

The main objective of the algorithm is the minimization of the total transportation cost, which in this context is expressed as the cumulative distance traveled by all vehicles across all routes. By integrating global exploration (ACO), local refinement (2-opt), and intensification-diversification control (Tabu Search), the proposed hybrid approach balances solution quality and convergence speed, showing strong potential for solving real-world instances of the Capacitated Vehicle Routing Problem with Time Windows (CVRPTW).

A. Main algorithm

The proposed hybrid algorithm combines the global search capabilities of Ant Colony Optimization (ACO) with the local refinement of 2-opt and diversification strategies provided by Tabu Search (TS). The objective is to minimize the total distance traveled while satisfying customer demands and time window constraints.

To provide a clear understanding of the solution approach, the pseudocode in Algorithm 1 illustrates the overall procedure of the hybrid algorithm.

B. Choosing next client

Probability of choosing next client during delivery is defined by equation (1):

$$p_{ij}^k = \frac{[f_{ij}(t)]^\alpha \cdot [g_{ij}(t)]^\beta}{\sum_{d \in N_i^k} ([f_{id}(t)]^\alpha \cdot [g_{id}(t)]^\beta)} \quad (1)$$

where p_{ij}^k - probability of choosing unvisited j th client for ant k that is currently located at the position of client i ; α - the influence of pheromone trails; β - the influence of heuristics; N_i^k - the set of client that ant k has not yet visited, and to which there is a path from client i ; $f_{ij}(t)$ - value of the function representing the intensity of pheromone trails on the path between clients i and j ; $g_{ij}(t)$ - value of the function representing the heuristic component of the path between clients i and j .

Using the calculated probability next client is being chosen using the roulette wheel selection.

Algorithm 1 Overall algorithm

```

for Number of iterations do
2:  Ant colony initialization
   for Every ant in the colony do
4:    while Unvisited client exists do
       while Feasible clients exist do
6:        for Every client in feasible
           clients do
           Calculate probability
8:        end for
           Choose next client using Roulette wheel selec-
           tion
10:       Update visited routes
           Remove chosen client from unvisited
12:     end while
           Return to the depot
14:   end while
           Apply two-opt
16:   Check for stagnation
   if No improvement > 10 then
18:     Activate tabu search
   end if
20:   Update best-known solution
   Update pheromone trails
22: end for
end for

```

C. Heuristic component

During preliminary experiments, the typical calculation of the heuristic component using only the Euclidean metric did not prove effective. Therefore, after conducting thorough analysis and taking into consideration the additional aspect of time windows of the problem, the heuristic determining was extended with a time-based coefficient that defines the urgency of the client and how quickly the vehicle must reach the client for the most effective use of time (2):

$$g_{ij}(t) = \frac{1}{d_{ij}(t)} + \frac{urgency}{\sqrt{\Delta t} + d_{ij}(t) + 1} \quad (2)$$

where $d_{ij}(t)$ - Euclidean distance between clients i and j ; $urgency$ - time-based coefficient of influence of the time window. The bigger coefficient is the bigger constraint of the time windows Δt - difference in time between time remaining until the end of the time window of the client i and current time.

D. System initialization

1) *Pheromone initialization*: During the analysis stage instead of the typical pheromone initialization function was utilized equation (3) that takes into consideration time window constraint. This equation assigns higher priority, namely higher pheromone intensity to paths leading not only to clients positioned, but also to those that are available earlier. This approach allows to reduce potential delays caused by the

vehicle having to wait for the client's time window to begin before service can be provided:

$$f_{ij}^{(0)} = 1.0 + \frac{1}{d(i,j) + 1} + \frac{1}{|r_j - r_i| + 1}, \forall i \neq j \quad (3)$$

where $f_{ij}^{(0)}$ - initial intensity of pheromones on path between clients i and j ; $d(i,j)$ - Euclidean distance between clients i and j ; r_j - start of the time window of the client j .

2) *Ant colony initialization*: Ant colony is initialized at the depot. All vehicles start and finish their journey at the depot.

E. Pheromone trails

After careful consideration of problem's domain and some preliminary experiments we suggested next algorithm for refreshing pheromone trails (4):

- Pheromone Evaporation - lowering intensity of existing pheromone trails.
- Increasing the pheromone levels based on the analysis of previously traversed path by ant colony.
- Elite ant acknowledgment - reinforcing pheromone level along the best-known path found by the best-performing ant.

$$f_{ij}^{t+1} = f_{ij}^t \cdot (1 - \rho) + \sum_{k \in K} \left(\frac{1}{D_k^t} \cdot \mathbf{h}_{(i,j) \in r_k^t} \right) + \frac{\gamma}{D_{best}} \cdot \mathbf{h}_{(i,j) \in r_{best}} \quad (4)$$

where f_{ij}^t - level of pheromones on path between clients i and j in iteration t ; ρ - evaporation coefficient; K - set of ants representing ant colony; r_k^t - route traversed by ant k in iteration t ; D_k^t - total distance traveled by ant k in iteration t ; $\mathbf{h}_{(i,j) \in r} = 1$ if path ij is in traversed route r (otherwise $\mathbf{h}_{(i,j) \in r} = 0$); γ - elite coefficient for the best ant; D_{best} - total distance traveled by the best ant; r_{best} - best-known found route.

F. Local route optimization using 2-opt

For local improvement of the traversed route implementation of the 2-opt algorithm was used [6]. Main goal of utilizing this algorithm was to find crossing sections of the concrete route and then swap subsections to improve the distance (presented in Algorithm 2).

G. Neighborhood search using Tabu search

After all ants traversed all routes, the neighborhood search is activated that compares clients between routes and tries to swap them if it improves the overall distance. This approach allows for quick exploitation and exploration simultaneously saving the algorithm from stagnating too early — Algorithm 3.

H. Hyperparameters

Efficiency of the algorithm is directly dependent on values of defined hyperparameters. The set of experiments was conducted for every hyperparameter in order to determine the best results for the given value of hyperparameter. Every experiment was performed 3 times in order to eliminate accidental or unstable results. Here is the list of defined hyperparameters:

Algorithm 2 2-opt implementation**Require:** *route* {Route that ant traversed}**Ensure:** A locally optimized route

```

1: improved  $\leftarrow$  true
2: while improved = true do
3:   improved  $\leftarrow$  false
4:   for i  $\leftarrow$  1 to route.length - 2 do
5:     for j  $\leftarrow$  i + 2 to route.length - 1 do
6:       new_route  $\leftarrow$  concatenate(route[1 to i-1],
reverse(route[i to j-1]), route[j to route.length])
7:       old_dist  $\leftarrow$  distance(route[i-1], route[i]) +
distance(route[j-1], route[j])
8:       new_dist  $\leftarrow$  distance(new_route[i-1], new_route[i])
+ distance(new_route[j-1], new_route[j])
9:       if new_dist < old_dist then
10:         route  $\leftarrow$  new_route
11:         improved  $\leftarrow$  true
12:       end if
13:     end for
14:   end for
15: end while
16: return route

```

Algorithm 3 Neighborhood search**Require:** *routes*, *vehicle_capacity* {Routes that ants traversed and vehicle capacity}**Ensure:** Optimized routes

```

1: best_routes  $\leftarrow$  routes
2: best_distance  $\leftarrow$  sum(routes)
3: for i  $\leftarrow$  0 to routes.length do
4:   for j  $\leftarrow$  1 to routes[i].length-1 do
5:     for k  $\leftarrow$  0 to routes.length do
6:       if k = j then
7:         continue
8:       end if
9:       for pos  $\leftarrow$  1 to routes[k].length - 1 do
10:        new_routes  $\leftarrow$  routes
11:        customer  $\leftarrow$  new_routes[i].pop(j)
12:        new_routes[k].insert(pos, customer)
13:        if new_routes[k].total_demand  $\leq$ 
vehicle_capacity then
14:          new_distance  $\leftarrow$  sum(new_routes)
15:        end if
16:        if new_distance < best_distance then
17:          best_routes  $\leftarrow$  new_routes
18:          best_distance  $\leftarrow$  new_distance
19:        end if
20:      end for
21:    end for
22:  end for
23: end for
24: return best_routes

```

- **Amount of ants** - small amount of ants (less than 20) caused bad and unstable results, whereas excessive amount (greater than 30) did not introduced improvement. This phenomenon can be explained by greater interference of pheromone trails between ants then required. Hence, as experiments proved, 20 ants as an optimal compromise between quality of solution and time required for computations.
- **Amount of algorithm iterations** - as confirmed in practice, the best found solution stagnated before reaching 30th iteration. So greater amount of iterations unnecessarily increased time required for the algorithm to finish without potentially drastic improvement of the result.
- **Pheromone evaporation** - small level of evaporation (around 0.1) meant, that pheromone trails had too big of influence o later iterations and hindered exploration settling in local minimum. On the other hand too great intensity of evaporation even though sometimes improved results, but they were unstable. On a balance, medium intensity of evaporation of around 0.5-0.6 lead to balance between exploration and exploitation even on big distances between clients.
- **Influence α of pheromone component** - pheromone level coefficient should be big enough for effective exploitation (around 0.5), however excessive value greater then 0.6 destabilizes solution
- **Influence of heuristic component** - heuristic component was implemented considering aspects of the studied problem, therefore this coefficient should be around 1.4-2.0 which produces the best result which was confirmed experimentally
- **Urgency coefficient** - Total neglecting of time window constraint as was expected complicates search of optimal routes, however setting the coefficient greater then 2 does not improve results.
- **Maximum amount of tabu search iterations** - The greater amount of iterations of tabu search the better results are. Nevertheless, on the downside time required for the algorithm to finish also grows drastically. Hence, 30 iterations allows for optimal results without requiring such big amount of time.

IV. EXPERIMENTAL EVALUATION AND RESULTS

All computations were carried out using the Google Colab platform [24], which provided a cloud-based virtual machine environment. The hardware configuration allocated to the runtime included a dual-core Intel(R) Xeon(R) CPU running at 2.20GHz and 12 GiB of available RAM. This environment ensured a consistent and reproducible execution setting for all test instances.

The use of Google Colab allowed for efficient prototyping and rapid experimentation without the need for dedicated local computational infrastructure. Furthermore, the platform's support for Python-based scientific libraries and its seamless integration with Jupyter notebooks greatly facilitated the de-

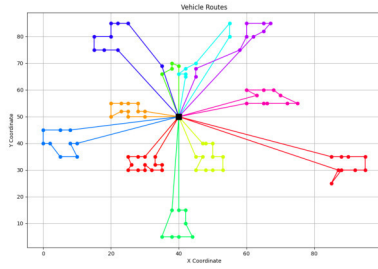


Fig. 2. Visualization of the computed vehicle routes for the VRPTW instance C101

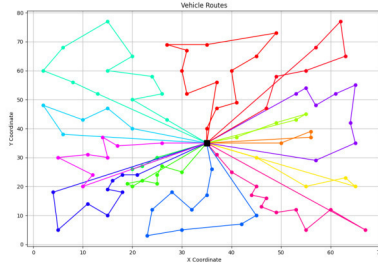


Fig. 3. Visualization of the computed vehicle routes for the VRPTW instance R101

velopment, debugging, and performance analysis of the hybrid algorithm.

While the hardware provided by Google Colab may not match the performance of high-end servers or GPU-accelerated environments, it is sufficiently powerful for evaluating medium-scale instances of the vehicle routing problem. The reported computation times in Table I reflect the algorithm's performance under these standard conditions, offering practical insight into its applicability for real-world use cases on accessible computing resources.

A. Dataset

Algorithm was tested using benchmark created by Solomon [25]. Dataset contains 56 sets of clients which are divided in 6 groups. Letter in the group name means density

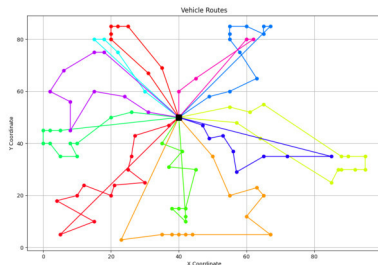


Fig. 4. Visualization of the computed vehicle routes for the VRPTW instance RC101

TABLE I
AVERAGE RESULTS FOR EACH GROUP

Client group	Distance	Vehicles	Time (min)
R1	1062.91	11.72	3.61
R2	1044.70	8.85	1.73
C1	1072.88	11.93	2.52
C2	771.07	5.29	2.96
RC1	1254.60	11.92	3.87
RC2	1004.34	7.25	3.09

of the client set ('R' - sparse density, 'C' - clustered clients and 'RC' is a mix of both of them) and the first number means the travel time type ('1' - short scheduling horizon and '2' - long scheduling horizon). Each client set contains 100 clients. Vehicle has constant capacity which varies between groups of client sets (C1*, R1* and RC1* - 200 vehicles; C2* - 700 vehicles; R2* and RC2* - 1000 vehicles). Furthermore, each client set has a depot from which every route should start and should finish. Each client has their demand which should be fulfilled fully by vehicle. Additionally clients have time constraints, namely each client has the time period during which he can be served by vehicle. Moreover each client has its service time which required to serve that client. Number of vehicles required to serve the client is not set so together with the total traversed distance is the subject for optimization. In our algorithm we focused on optimizing the total distance and the number of required vehicles was secondary.

B. Experiments

For final results each set of 100 clients was tested for 3 times which eliminated accidental results and ensured stability

In order to evaluate the performance of the proposed hybrid algorithm, a series of computational experiments were conducted on benchmark instances with varying numbers of customers, time window constraints, and vehicle capacities. The results of these experiments are summarized in Table I.

Table I presents not only the total minimum distance achieved for each problem instance and the number of vehicles required to serve all customers, but also the total computational time measured in minutes. These metrics allow for a comprehensive assessment of both the solution quality and the efficiency of the algorithm.

The total minimum distance reflects the main optimization objective and serves as a primary indicator of route efficiency. The number of vehicles used corresponds to the operational feasibility and reflects how well the algorithm can consolidate deliveries. Finally, the computation time provides insight into the algorithm's scalability and runtime behavior across different instance sizes.

This experimental setup allows for both qualitative and quantitative comparison of the algorithm's performance. Additionally, the data collected enables further analysis of the trade-offs between solution quality and computation time under various parameter settings.

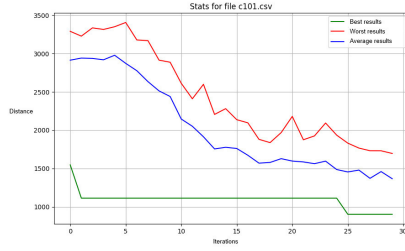


Fig. 5. Best, worst and average results in iteration receipts for instance C101

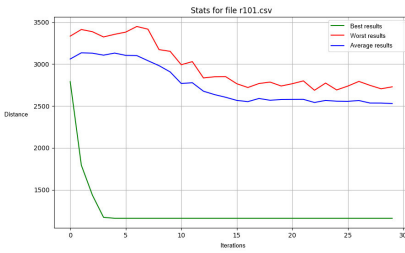


Fig. 6. Best, worst and average results in iteration receipts for instance R101

V. EVALUATION AGAINST STANDARD BENCHMARKS

For each set, drawings were generated showing the routes of individual vehicles and statistics were presented showing the best, worst and average route searches in individual iterations.

The diagram presents the resulting vehicle routes generated by the proposed hybrid algorithm for the benchmark instance C101 (2). Each colored line represents the route of a single vehicle, starting and ending at the central depot (marked with a black square). Individual customer locations are depicted as dots, and the connected paths illustrate the order of service while respecting both vehicle capacity and time window constraints. Similar diagrams were created for other benchmark instances to validate the robustness and adaptability of the proposed solution across various configurations of customer distribution and time constraints (f.e. Fig. 3 and Fig. 4).

In Fig. 5, the green line represents the best result obtained (minimum route length), the red line represents the worst, and the blue line represents the average value obtained in a given iteration.

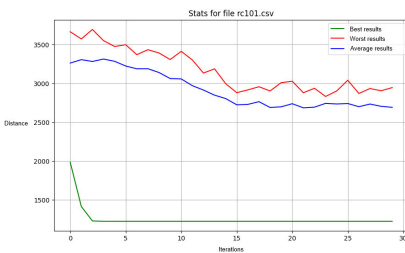


Fig. 7. Best, worst and average results in iteration receipts for instance RC101

A clear trend of improving average results and gradual stabilization of the values of the best solutions can be observed, which indicates effective exploration of the solution space by the algorithm. Significant quality improvement occurs already in the initial iterations, and further calculations allow for successive refinement and local optimization of solutions. A sudden drop in the best result after about 25 iterations suggests effective operation of the tabu search component or local optimization (2-opt), leading to significant improvement of previously found routes.

Analogous graphs were generated for the remaining test instances, which allows for analysis of the stability and repeatability of the proposed approach (f.e. Fig. 6 and Fig. 7).

TABLE II
RELATIVE EFFICIENCY IMPROVEMENT COMPARING TO PREVIOUSLY KNOWN RESULTS

File	Best Known Distance	Distance obtained	Relative efficiency(%)
r101	1650.80	1011.41	38.73
r102	1486.12	1054.03	29.08
r103	1292.68	1029.14	20.39
r104	1007.31	1006.33	0.10
r105	1377.11	975.83	29.14
r106	1252.03	993.79	20.63
r107	1104.66	1060.89	4.00
r109	1194.73	1075.12	10.01
r110	1118.84	992.36	11.30
r111	1096.72	1062.86	3.09
r201	1252.37	1029.12	17.83
r202	1192.70	969.44	18.65
rc101	1696.95	1189.84	29.88
rc102	1554.75	1295.17	16.70
rc103	1261.67	1205.30	4.47
rc105	1629.44	1240.40	23.88
rc106	1424.73	1265.33	11.19
rc107	1230.48	1218.62	0.96
rc201	1406.94	964.95	31.42
rc202	1365.65	990.67	27.46
rc203	1049.62	918.56	12.49
rc205	1297.65	1089.39	16.05
rc206	1146.32	937.28	18.24
rc207	1061.14	885.42	16.56

In Table II, the relative efficiency improvement achieved by the proposed hybrid algorithm is presented in comparison with the best-known solutions reported in [26]. The evaluation was conducted using the efficiency formula defined in equation (5), which quantifies the relative gain by measuring how much shorter the total distance of the obtained solution is compared to the reference result.

$$E_i = \frac{x_i - x_0}{x_0} \cdot 100\% \quad (5)$$

where E_i - relative efficiency of our solution; x_i - distance retrieved as a result of our solution; x_0 - distance for best known solution.

The suggested approach not only outperformed previously known results in the majority of benchmark instances, but in some cases achieved drastic improvements-up to approximately 30% reduction in total distance. Such a substantial gain is particularly impressive given that it was achieved

without significantly increasing computation time. In fact, the algorithm maintained practical execution times, making it suitable for applications where responsiveness is critical.

These results underline the robustness and effectiveness of the hybrid approach, especially in scenarios with tight time windows and large numbers of customers. The combined use of Ant Colony Optimization for initial solution generation, 2-opt local search for intra-route optimization, and Tabu Search for escaping local minima, proved particularly effective in navigating complex solution landscapes.

Importantly, the efficiency gains translate directly into more resource-effective planning-fewer vehicles and less travel distance, while respecting time constraints. This makes the proposed method highly valuable in real-time logistics and transportation scenarios, where computational speed and solution quality must go hand-in-hand. The consistent improvements across problem instances confirm that the developed algorithm is not only scalable and adaptive, but also practical for deployment in real-world operations.

VI. CONCLUSION

In this study, a novel hybrid algorithm combining Ant Colony Optimization, 2-opt local search, and Tabu Search has been proposed for solving the Vehicle Routing Problem with Time Windows. The algorithm effectively integrates global exploration with local refinement and strategic escape from local optima, resulting in a robust and adaptive solution method.

The experimental results confirmed that the proposed method significantly improves upon previously known solutions in terms of total route distance, with some instances showing improvements of up to 30%. Importantly, these improvements were achieved without incurring substantial computational cost, demonstrating the algorithm's suitability for real-time and large-scale applications.

The results also indicate that the hybrid approach ensures a better balance between intensification and diversification in the search process. This leads to high-quality solutions that respect all problem constraints, including capacity and time windows, while optimizing the number of vehicles and total travel distance.

Future work may involve extending the algorithm to dynamic or stochastic versions of the VRP, testing its performance on real-world datasets, and incorporating learning-based mechanisms to further enhance its adaptability and efficiency.

REFERENCES

- [1] A. Schirjver, *On the history of combinatorial optimization (till 1960)*. In: K. Aardal, G.L. Nemhauser, R. Weismantel (Eds.), *Handbook of Discrete Optimization*, 2005, Amsterdam.
- [2] C. Rego, D. Gamboa, F. Glover, C. Osterman, *Traveling salesman problem heuristics: leading methods, implementations and latest advances*, European Journal of Operational Research, 2014, vol. 211 (3).
- [3] G.B. Dantzig, J.H. Ramser, *The Truck Dispatching Problem*, 1959, Management Science, vol. 6 (1).
- [4] G. Ghiani, G. Laporte, R. Musmanno, *Introduction to Logistics Systems Management (2nd ed.)*, 2013, Wiley.
- [5] P. Toth, D. Vigo, *The Vehicle Routing Problem. Monographs on Discrete Mathematics and Applications*, 2001, SIAM, Philadelphia.
- [6] G. A. Croes, *A method for solving traveling-salesman problems*, 1958, Operations Research, vol. 6(6), pp. 791–812, <https://doi.org/10.1287/opre.6.6.791>
- [7] S. Lin, *Computer solutions of the traveling salesman problem*, 1965, Bell System Technical Journal, vol. 44(10), pp. 2245–2269, [urlhttps://doi.org/10.1002/j.1538-7305.1965.tb04146.x](https://doi.org/10.1002/j.1538-7305.1965.tb04146.x).
- [8] N. Christofides, A. Mingozzi, P. Toth, *The vehicle routing problem*, In *Combinatorial Optimization*, 1979, Wiley, pp. 315–338.
- [9] F. Uddin, N. Riaz, A. Manan, I. Mahmood, O.-Y. Song, A.J. Malik, A.A. Abbasi, *An Improvement to the 2-Opt Heuristic Algorithm for Approximation of Optimal TSP Tour*, 2023, Applied Sciences, 13(12):7339, [urlhttps://doi.org/10.3390/app13127339](https://doi.org/10.3390/app13127339)
- [10] M. Dorigo, V. Maniezzo, A. Colnari, *Ant system: Optimization by a colony of cooperating agents*, (1996, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 26(1), pp. 29–41, <https://doi.org/10.1109/3477.484436>
- [11] K. F. Doerner, R.F. Hartl, M. Reimann, *Metaheuristics for the vehicle routing problem with loading constraints*, 2006, Networks, vol. 49(4), pp. 294–307.
- [12] J. Ochelska-Mierzejewska, *Ant Colony Optimization Algorithm for Split Delivery Vehicle Routing Problem*, In International Conference on Advanced Information Networking and Applications (was ICOIN), 2020, https://link.springer.com/chapter/10.1007/978-3-030-44041-1_67.
- [13] F. Glover, *Future paths for integer programming and links to artificial intelligence*, 1986, Computers and Operations Research, vol. 13(5), pp. 533–549, [https://doi.org/10.1016/0305-0548\(86\)90048-1](https://doi.org/10.1016/0305-0548(86)90048-1)
- [14] J.-F. Cordeau, M. Gendreau, G. Laporte, J.-Y. Potvin, F. Semet, *A guide to vehicle routing heuristics*, 2001, Journal of the Operational Research Society, vol. 53(5), pp. 512–522, [urlhttps://doi.org/10.1057/palgrave.jors.2601319](https://doi.org/10.1057/palgrave.jors.2601319)
- [15] M. Gendreau, A. Hertz, G. Laporte, *A tabu search heuristic for the vehicle routing problem*, 1994, Management Science, vol. 40(10), pp. 1276–1290, [urlhttps://doi.org/10.1287/mnsc.40.10.1276](https://doi.org/10.1287/mnsc.40.10.1276)
- [16] J. Ochelska-Mierzejewska, *Tabu Search Algorithm for Vehicle Routing Problem with Time Windows*, 2020, https://link.springer.com/chapter/10.1007/978-3-030-34706-2_7. DOI: 10.1007/978-3-030-34706-2_7.
- [17] N. Paisarnvirosrak, P. Rungrueang, *Firefly Algorithm with Tabu Search to Solve the Vehicle Routing Problem with Minimized Fuel Emissions: Case Study of Canned Fruits Transport*, 2023, LOGI – Scientific Journal on Transport and Logistics, vol. 14(1), pp. 263–274, [urlhttps://doi.org/10.2478/logi-2023-0024](https://doi.org/10.2478/logi-2023-0024).
- [18] X. Ma, C. Liu, *Improved Ant Colony Algorithm for the Split Delivery Vehicle Routing Problem*, 2024, Applied Science, vol. 14(5090), <https://doi.org/10.3390/app14125090>
- [19] L.M. Gambardella, É.D. Taillard, G. Agazzi, *MACS-VRPTW: A multiple ant colony system for vehicle routing problems with time windows*, In D. Corne, M. Dorigo, F. Glover (Eds.), *New ideas in optimization*, 1999, McGraw-Hill, pp. 63–76.
- [20] M. Tadaros, N.A. Kyriakakis, *A Hybrid Clustered Ant Colony Optimization Approach for the Hierarchical Multi-Switch Multi-Echelon Vehicle Routing Problem with Service Times*, 2024, Computers & Industrial Engineering, <https://diva-portal.org/smash/get/diva2:1802273/FULLTEXT01.pdf>
- [21] J.B. Holliday, E. Osaba, K. Luu, *An Advanced Hybrid Quantum Tabu Search Approach to Vehicle Routing Problem*, 2025, <https://arxiv.org/pdf/2501.12652v1>
- [22] Z. Zheng, B. Ji, S.S. Yu, *An Adaptive Tabu Search Algorithm for Solving the Two-Dimensional Loading Constrained Vehicle Routing Problem with Stochastic Customers*, 2023, Sustainability, vol. 15(2), 1741, <https://www.mdpi.com/2071-1050/15/2/1741>
- [23] Y. Liu, Z. Wang, J. Liu, *A Quick Pheromone Matrix Adaptation Ant Colony Optimization for Dynamic Customers in the Vehicle Routing Problem*, 2024, vol. 12(7), 1167, <https://doi.org/10.3390/jmse12071167>
- [24] Google Colaboratory, 2024, <https://colab.research.google.com/>
- [25] M. Solomon, *Solomon VRPTW Benchmark*, 1987, <http://w.cba.neu.edu/~msolomon/problems.htm>
- [26] Top, VRPTW for 100 customers, 2008, <https://www.sintef.no/projectweb/top/vrptw/100-customers/>

Opportunities and Challenges of LLMs as Post-OCR Correctors

Radoslav Koynov
0009-0003-8331-7475
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH
Burckhardtweg 4, 37077
Email: radoslav.koynov@gwdg.de

Triet Ho Anh Doan
0000-0002-7247-9108
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH
Burckhardtweg 4, 37077
Email: triet.doan@gwdg.de

Abstract—Large Language Models (LLMs) have demonstrated potential as zero-shot Post-OCR correctors for historical texts. However, previous research has typically focused on a single data set and only evaluated Character Error Rate (CER) or Word Error Rate (WER). This study investigates the potential of LLMs to enhance the accuracy of Optical Character Recognition (OCR) and the limitations of the models. To this end, an evaluation of the approach is conducted for a number of German and English historical datasets, with an in-depth analysis of the model corrections and deviation from the ground truth. We demonstrate that LLMs have the capacity to enhance the quality of OCR results as zero-shot correctors in some cases, and fine-tuning LLMs shows promise as part of an LLM-based Post-OCR correction system, if certain risks are mitigated carefully.

I. INTRODUCTION

OPTICAL CHARACTER RECOGNITION (OCR) is the technology used to digitize printed and handwritten text, enabling large-scale text extraction from scanned documents. However, OCR systems are prone to errors, particularly when dealing with degraded documents, handwritten scripts, or complex layouts.

Recent advances in Large Language Models (LLMs) have opened new possibilities for automated post-OCR correction. LLMs, with their strong contextual understanding and ability to generate human-like text, offer a promising approach to refining OCR outputs by correcting errors and restoring missing characters. However, due to their nature as a generative model with a certain amount of creativity, they can also introduce new errors. These new errors may be qualitatively different from typical OCR errors, and potentially much harder to detect. Thus, the effectiveness and limitations of different LLMs, prompting techniques and fine-tuning strategies for post-OCR correction remain an open research area.

This paper presents a series of experiments that seek to enhance the accuracy of OCR texts through the utilization of LLMs. It is important to note that these models were not exposed to scanned images; rather, they were presented with OCR texts from various German and English datasets. The models were set up with a constant prompt and temperature during the course of the experiments. Furthermore, a fine-tuning process was implemented with the objective of enhancing the efficacy of the models.

In order to evaluate the results, a comparison was made between the Character Error Rate (CER) and Word Error Rate (WER) before and after the usage of LLMs. Additionally, we define the character change rate (CCR) and word change rate (WCR) analogously, but between the original OCR result as a reference, and the model-corrected version. A more in-depth examination is also conducted of the particular edit operations that are required in order to transform a piece of OCR text to its ground truth, and the edit operations that are implied by the LLMs.

II. RELATED WORKS

Prior research on the field of post-OCR correction has explored various models, datasets, and evaluation techniques to address errors in OCR-processed text.

Soper et al. [1] already show the capabilities of correcting noisy text outputs with pre-trained language models.

One of the more recent works on post-OCR correction using LLMs compares fine-tuned Llama2-7B, Llama2-13B, and BART on the BLN600 dataset, a collection of British newspapers from the 19th centuries [2], [3]. They highlight the challenges posed by historical spelling conventions and employ a simple instruction prompt. This work impressively demonstrates the potential of fine-tuned LLMs for post-OCR correction, but does not sufficiently highlight certain risks or generalize to further datasets.

Earlier competitions, such as the ICDAR 2017 [4] and ICDAR 2019 challenges [5], provided foundational datasets for post-OCR correction. The ICDAR 2017 dataset includes 12 million aligned symbols extracted from newspapers and monographs in English and French, while the ICDAR 2019 competition expanded this effort to 22 million symbols across 10 European languages, focusing on multilingual post-OCR correction. Although these competitions predate modern LLMs, their datasets remain valuable for training and evaluation.

A notable study [6] in 2022 employed large ensembles of character sequence-to-sequence transformer models for post-OCR correction, achieving strong performance on the ICDAR 2019 dataset. This approach involved manually training a transformer model from scratch and segmenting documents into smaller pieces for processing. While effective, this method

requires extensive training and does not leverage the zero-shot or few-shot capabilities of modern LLMs.

Kanerva et al. use various LLMs for post-OCR correction on an English and a Finnish dataset and note that the results are much better for the English dataset [7]. Out of the models they employed, *GPT-4o* shows the most promise for both languages and achieves a reduction in the character error rate even for the Finnish texts. However, they conclude that this improvement for the Finnish texts is not considerable enough to practically attempt zero-shot post-OCR correction for Finnish at the present time.

A more recent study from 2024 evaluates OpenAI's *GPT-4*, *GPT-4 Turbo*, and *GPT-3.5 Turbo* models on post-OCR correction of challenging English prosody texts [8]. This study explores multiple prompts, metadata inclusion, and varying temperature settings. Using CER as the only metric, the study finds only marginal differences between models and prompt variations. While valuable, the dataset used in this study is not released to the community and highly specialized.

III. DATASETS AND MODEL SELECTION

A. Datasets

Three datasets were used in our experiment, as shown in Table I. BLN600 [3] contains English-language crime reports from newspapers from 1834 to 1894. This dataset has a relatively low initial error rate.

The next dataset was developed within the Optical Character Recognition Development (OCR-D) project, we refer to it as OCR-D-GT. Its content is based on transcription data stored in the German Text Archive [9]. The dataset is publicly accessible on GitHub [10], but contains only ground truth data. Therefore, for our experiments, an OCR workflow was executed on the text to produce OCR'd texts. The workflow is straightforward and utilizes the `tesseractocr-recognize` processor with the German Print [11] model.

Lastly, the ICDAR2019 dataset [5], introduced for the ICDAR 2019 Competition on Post-OCR Text Correction, comprises OCR outputs ground truth data for historical documents in multiple languages. We utilize the English and German subsets, which include digitized materials from sources such as the British Library and the German National Library. These texts contain a variety of printed materials, including newspapers and historical books. Specific publication years and genres are not detailed in the dataset's documentation.

All employed datasets contain ground truth data, i.e. documents already correctly digitalized by human experts which we use for evaluation of the results, as well as for preliminary fine-tuning experiments. They are structured in individual files. An OCR output file together with its ground truth file is referred to as a *page* or *document* throughout this work.

The average CER reported in Table I refers to what we measured for the full datasets.

TABLE I: Datasets for Post-OCR Correction

Dataset	Avg. CER	Pages	Language(s)	Years
BLN600	0.07248	600	English	1834-1894
OCR-D-GT	0.1486	217	German, others	1506-1897
ICDAR2019-EN	0.2018	150	English	
ICDAR2019-DE	0.2543	10,032	German	

B. Model Selection

In line with promising models from previous research, we select a model from the *Llama* family, and one from the *GPT* family.

GPT-4o mini [12] is optimized for efficiency while retaining strong multilingual reasoning capabilities, making it suitable for practical large-scale application as a post-OCR corrector.

Llama 3.3 70B [13] is a state-of-the-art instruction-tuned open-source LLM. It performs competitively on a range of benchmarks and is freely available for research and commercial use. It is also the successor of *Llama 2*, the model who showed promise for correcting errors in the BLN600 datasets.

IV. METRICS

We define several character-level metrics comparing the ground truth (GT), the original OCR output (OCR), and the model output (PostOCR) for a single document.

A. Character Error Rate

CER measures the minimum number of character-level edits (insertions, deletions, and substitutions) required to convert a hypothesis string into a reference string, normalized by the length of the reference:

$$\text{CER}(h, r) = \frac{S + D + I}{N} \quad (1)$$

where:

- h is the hypothesis string,
- r is the reference (ground truth),
- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of characters in the reference.

Based on this, we define:

- $\text{CER}_{\text{old}} = \text{CER}(\text{OCR}, \text{GT})$: the error rate of the original OCR output against the ground truth.
- $\text{CER}_{\text{new}} = \text{CER}(\text{PostOCR}, \text{GT})$: the error rate of the model-corrected text against the ground truth.

B. Relative CER Reduction

To quantify the effectiveness of post-OCR correction, we define the relative improvement in CER as:

$$\text{CER Reduction} = \frac{\text{CER}_{\text{old}} - \text{CER}_{\text{new}}}{\text{CER}_{\text{old}}} \quad (2)$$

A value of 1 indicates perfect correction (i.e., all original errors were fixed), while a value of 0 indicates no improvement.

C. Character Change Rate

In addition to the traditional metrics comparing the OCR text with its ground truth, we introduce the **Character Change Rate (CCR)**. It quantifies the modification introduced by the post-OCR correction model, by using the original OCR output as the reference:

$$\text{CCR} = \text{CER}(\text{PostOCR}, \text{OCR}) \quad (3)$$

D. Change Ratio

From CCR, we derive a relative metric that quantifies the amount of change the model introduced with respect to the original CER.

$$\text{Change Ratio} = \frac{\text{CCR}}{\text{CER}_{\text{old}}} \quad (4)$$

A high Change Ratio together with a small CER Reduction indicates the model introduces many new errors.

E. Consecutive Edit Operations

While the CER and CCR capture the extent of changes necessary or introduced by a model, it does not reflect their distribution or locality. To address this, we define the *consecutive edit sequence* as a run of character-level edit operations that are adjacent in the edit space. We define adjacency according to the edit operations computed in the Levenshtein algorithm, whose output is an ordered list of tuples (op, i, j) where $\text{op} \in \{I, S, D\}$ at position i in the source string and j in the target. A sequence of operations is considered consecutive if the positions follow valid edit path transitions:

- replace: $(i + 1, j + 1)$
- delete: $(i + 1, j)$
- insert: $(i, j + 1)$

Given a threshold k , we compute the following metrics over these sequences:

- **Average Number of Consecutive Edit Sequences $\geq k$** – The arithmetic mean of detected consecutive edit sequences $\geq k$ per document.
- **Average Length of Consecutive Edit Sequences $\geq k$** – The average number of operations within each consecutive sequence meeting the threshold.

We also computed analogous metrics while restraining the types of consecutive operations, inspecting pure insertion and pure deletion sequences. This is useful to quantify missing information from OCR results as well as model hallucinations in post-OCR outputs.

F. Word-Level Metrics

All of these definitions are analogous at the word level, where insertions, deletions and substitutions are made at the level of words. For example, we use the term Word Change Rate (WCR) for the Word Change Rate, without defining it explicitly.

V. EXPERIMENT SETUP

A. Data loading

We create a custom data loader for each of the datasets to homogenize the structure and prepare them to be passed to the LLMs for correction. The loader matches OCR texts with their ground truths to allow for automatic evaluation. The loaders use *dinglehopper* [14] to extract text from XML files (which are in PAGE [15] or ALTO [16] format), or plain-text files and apply minimal pre-processing.

B. LLM-based Post-OCR Correction

For each of the datasets, we run the LLM-based post-OCR correction pipelines, using the prompts shown in Figure 1. For the fine-tuned GPT models, a 75/25 train-test split is used, and the evaluation results are reported on the documents of the test partition. Fine-tuning is done using OpenAI’s fine-tuning API [17] with the default settings. Note that this does not include any holdout or cross-validation, but simply runs for a fixed number of epochs. We plan on implementing more sophisticated fine-tuning approaches with in future research.

Zero-Shot:

"You are a Post-OCR corrector. You correct mistakes in historical texts that are caused by errors in the Optical Character Recognition. You should NOT fix grammar or spelling which deviate from Standard {{language}}, because the texts are historical. Please only include the processed text in your response."

Fine-Tuned Models:

"You are a Post-OCR corrector. You correct mistakes in historical texts that are caused by errors in the Optical Character Recognition. Please ONLY include the corrected text in your replies."

Common User Query:

"Please correct OCR-related mistakes in the following historical text:
\n\n [OCR TEXT]"

Fig. 1: Prompt Templates Used for Post-OCR Correction

We use all models via a REST API and we use a temperature of 0.5 across all experiments for simplicity.

C. Automatic Evaluation

For each correction run, an automated evaluation script computes all metrics described in the previous section on a per-document basis and saves them as a dataframe. Additionally,

aggregations such as averages are computed and reported. The exact edit operations and a number of visualization plots are also saved automatically for each run.

VI. EVALUATION AND ANALYSIS

In this section, we will first perform the standard evaluation based on CER and WER, before diving deeper to also investigate what changes the models applied and which errors it could (not) correct.

A. CER and WER

1) *BLN600 - An English low-error dataset*: For the *BLN600* dataset, the CER and WER reduction are displayed in Table II. Both the *GPT-4o mini* and the *Llama-3.3-70B* models achieved a significant reduction in the average CER. The GPT model and the open-source Llama model reduced the CER by almost 58% and 48% respectively. The zero-shot approach with GPT-4o-mini thus slightly outperforms the fine-tuned Llama 2 model tested on this dataset [2], while the newer-generation Llama model almost achieves the performance.

On the word level, the improvements are even more considerable, with both models reducing the WER by over 75%. This is a clear indication that the models were particularly effective at correcting words with just one or few errors. Remaining errors might be in part due to sequences with accumulated errors, where it is increasingly hard or impossible to reconstruct missing information.

To get a better view of the distribution of error rates, it is useful to look at Figure 2, which shows the CER and WER of each document before and after correction for the more effective zero-shot model *GPT-4o mini*. The results demonstrate that both the CER and WER can be substantially reduced for numerous documents, particularly those exhibiting low initial CER values. Conversely, for documents with high initial CER, it becomes more challenging for the model to correct.

The fine-tuned model achieved even higher reduction in both CER and WER, reducing the character errors by almost 65% on average. This is a further indication that the fine-tuned models show promise of further improvements, when the zero-shot approach already yields good results. However, it should be noted that this is not necessarily statistically significant given the smaller test size for the fine-tuning approach.

2) *OCR-D-GT - A tricky German dataset*: For the German-language OCR-D-GT dataset, the aggregated results are summarized in Table III. Unfortunately, the models could not reach reduction in the character or word error rate. In fact, the CER actually increased by at least 30%.

The fine-tuned model performed much worse on average. It is extremely volatile and introduced many mistakes and even hallucinated entire paragraphs for some of the articles, leading to a large increase in CER when taking the arithmetic mean. This can partially be attributed to the relatively small and extremely heterogeneous dataset, covering several centuries with different genres and a spread of base CER from 0.23% up to 78.31%. Some of the fine-tuning examples of OCR results

with up to 95% WER encourage the model to hallucinate corrections.

The fine-tuned model did however reduce the CER for a larger share of articles than the base model. This suggests that a more involved fine-tuning approach together with a larger and improved dataset can still be a promising approach.

3) *ICDAR2019*: For the ICDAR 2019 datasets, we use the German and English subsets, removing alignment data in the data loading step. In the case of the English dataset, both *Llama 3.3 70 B Instruct* and *GPT-4o mini* achieve a small reduction of the WER, but also a small increase in the CER. For the German texts, *GPT-4o mini* achieves a slight reduction in the CER and WER, while the Llama model yields very poor results, increasing the CER by 40%.

Although both models showed much promise as zero-shot correctors when employed for the BLN600 dataset, this is unfortunately not the case for the more complex datasets with higher initial error rates.

B. Comparing OCR and Model Output – CCR and WCR

There are various commonly found OCR errors, such as misinterpreted characters, disjointed characters and problems with hyphenation [18]. These might be recognizable to readers due to the visual similarity that lead to the error. Unfortunately, LLM correctors can introduce new types of errors that might be more problematic. For this reason, it is not enough to simply investigate CER and WER when comparing model performance. For example, in an OCR text with 10% CER, let a post-OCR correction model A reduce the CER to 5% by performing the edit operations needed to eliminate 5% of errors (CCR of 5%). Let model B also reduce the CER to 5%, but with a CCR of 7%. In this case, model A should be preferred since it did not introduce any new, potentially more problematic errors. In Figure 3, the CER before and after correction, as well as the CCR, are displayed for the BLN600 dataset.

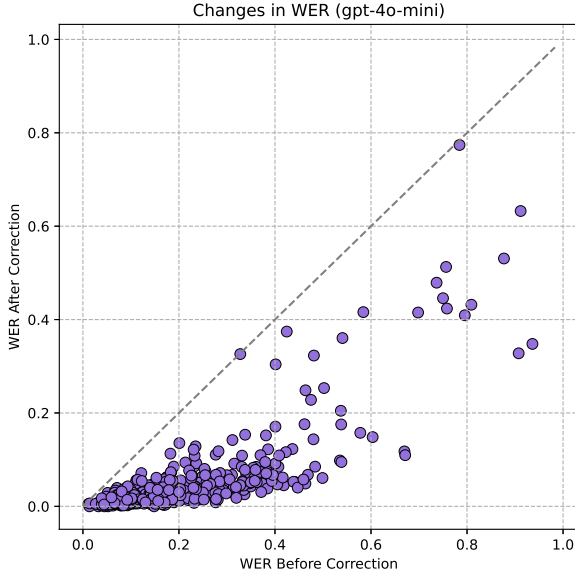
We can see that the *Llama 3.3 70B Instruct* model actually introduced more changes to the OCR text than *GPT-4o mini*, but unfortunately many of these changes did not reduce CER. On the other hand, it is a positive result that the fine-tuned model's higher CER reduction does not come with the price of an increased change rate.

As previously established, the results for the other datasets were not satisfactory. In the only other case, where a CER reduction was reached, the *GPT-4o-mini* model for the English-language subset of *ICDAR2019*, the model reduced the CER from 25.43% to 25.01% with a change rate of 3.92%. Although this is a relatively low CCR, given the amount of errors in the OCR result, it still means that the model introduced or changed existing errors amounting to more than 3% of the total characters, almost ten times more than it corrected.

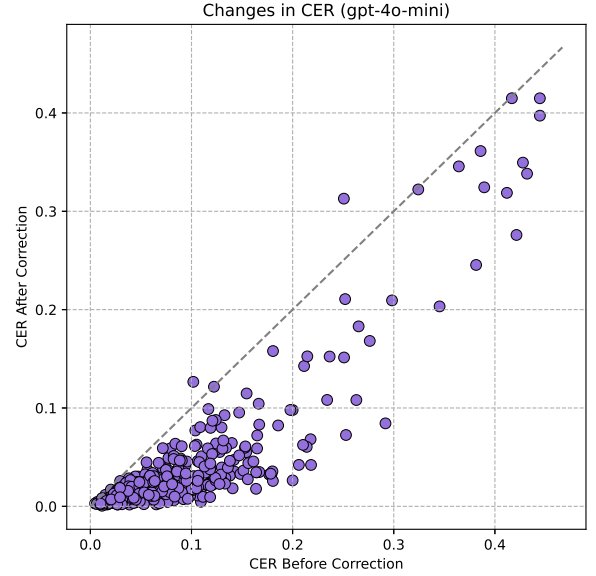
As with the analysis of CER and WER, it is useful to gain a better view of the distribution of results on a document-by-document basis, instead of just considering averages. To visualize this, we add a color map to the scatter plot considered in the previous section. Since the CCR naturally correlates

TABLE II: Benchmarking LLMs for Post-OCR Correction on BLN600

Model	CER	WER	CER / WER Reduction
GPT-4o mini	0.07248 → 0.03065	0.18634 → 0.04404	57.71% / 76.37%
Llama-3.3-70B	0.07248 → 0.03778	0.18634 → 0.04613	47.89% / 75.24%
FT GPT-4o mini	0.06578 → 0.0231	0.16577 → 0.03216	64.93% / 80.6%



(a) WER Scatter Plot



(b) CER Scatter Plot

Fig. 2: Per-Document Changes in WER and CER for BLN600 using GPT-4o mini.

TABLE III: Benchmarking LLMs for Post-OCR Correction on OCR-D-GT

Model	CER	WER	CER / WER Reduction
Llama-3.3-70B-Instruct	0.14855 → 0.17619	0.27290 → 0.36078	-18.60% / -32.20%
GPT-4o mini	0.14855 → 0.17735	0.27290 → 0.37705	-19.39% / -38.16%
FT GPT-4o mini	0.15716 → 0.37338	0.27118 → 0.53738	-137.58% / -98.16%

TABLE IV: Benchmarking LLMs for Post-OCR Correction on ICDAR-2019

Language Subset	Model	CER	WER	CER / WER Reduction
EN	Llama-3.3-70B-Instruct	0.20179 → 0.21304	0.31620 → 0.29992	-5.57% / 5.15%
EN	GPT-4o mini	0.20179 → 0.20264	0.31620 → 0.31130	-0.42% / 1.55%
DE	Llama-3.3-70B-Instruct	0.25430 → 0.35742	0.81175 → 0.83743	-40.55% / -3.16%
DE	GPT-4o mini	0.25430 → 0.25010	0.81175 → 0.77108	1.65% / 5.01%

with both CER_{old} and CER_{new} it does not give a clear enough visual indication of the *relative* change. To account for this, we use the *Change Ratio* for the color axis, but clip the values at 2.0, which already indicates a very high change relative to the base CER, but keeps the scale readable at lower values.

In Figure 4, we can see that the Change Ratio for the BLN600 documents using *GPT-4o mini* is usually between 0.6 and 0.9, although it is lower for some documents that still exhibit a CER reduction. Concerning the documents with a high CER_{old} there are several documents with some CER reduction, but they generally have a significantly higher Change

Ratio. There are also some documents with no improvements and barely any changes made by the model.

In Figure 5 the same plot is shown for the output of the *Llama 3.3 70B* model. We can see one outlier, where the model introduced a large amount of incorrect text for a single document. Apart from that, the plots look relatively similar, but due to the fixed color map with 0 on the low end and 2+ on the high end, we can also see that the Change Ratio is slightly higher for most documents, compared to the GPT model output.

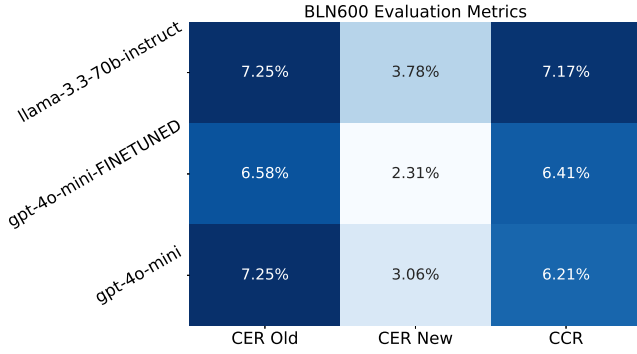


Fig. 3: Macro-averaged CER and CCR of correction models on BLN600

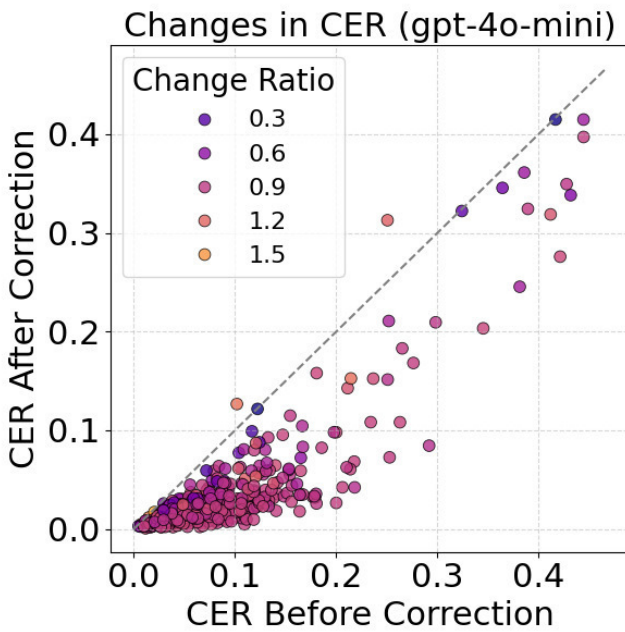


Fig. 4: Per-Document Changes in CER with Change Ratio as Color Axis for BLN600 using GPT-4o mini

C. Diving Deeper – Edit Operations and Consecutive Edits

When considering only the CER, the information about the types of edit operations in the shortest transformation sequence is lost. The average number of insertions (I), substitutions (S) and deletions (D) necessary to transform the OCR result to the Ground Truth (Expected) and to the Post-OCR document (Predicted) is given in Table V for all datasets. The position of the edit operations in the document can also be of interest. This is particularly the case when many errors occur consecutively in an OCR text, because this vastly increases the difficulty of the correction task. On the other hand, when a model prediction contains long consecutive sequences, especially of insertions, this is an indication of model hallucinations.

For BLN600, the dataset where the models achieved good results, we can see that the models predicted less insertions

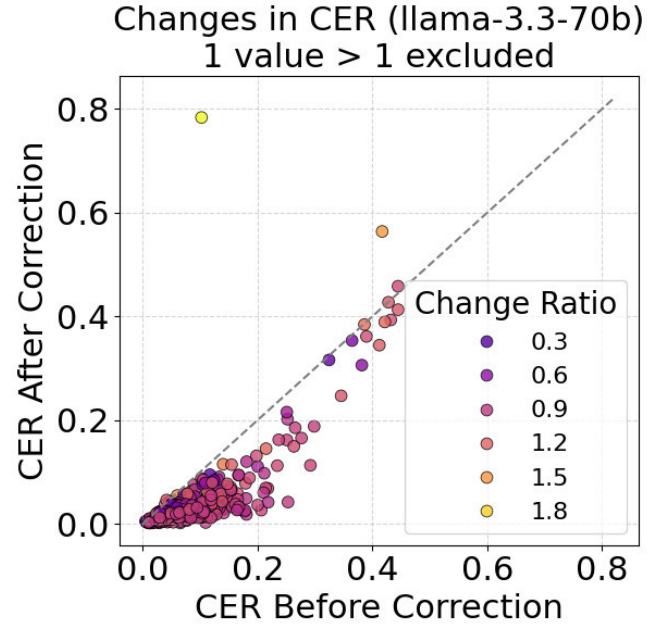


Fig. 5: Per-Document Changes in CER with Change Ratio as Color Axis for BLN600 using Llama 3.3 70B

than were expected. *GPT-4o mini* predicted less insertions, but reached a higher CER reduction. This becomes clearer when we look at the consecutive edit sequences with minimum length $k = 6$ operations. These are likely not retrievable from the OCR text. While *GPT-4o mini* applied 1.94 such sequences per document with an average length of 9.32, a large portion of these are pure “delete”, on average namely 1.55 sequence per document with an average length of 9.89. This means that the model sometimes deleted sequences of characters that it deemed corrupted or illegible. On the other hand, it only applied 0.09 pure insertion sequences with an average length of 7.25.

For the same dataset, *Llama 3.3 70B* applied 2.87 consecutive edit sequences to such sequences with an average length of 10.49. While a considerable part of these were pure delete sequences as well, it also includes 0.26 pure insertion sequences of this length. This means that the Llama model’s corrections contain some hallucinations, even for the dataset where we obtained a reduction in CER.

For the other datasets, the models do not achieve significant reduction of the error rate. Considering the expected edit operations and consecutive operations can give additional clues concerning the difficulty of the correction task for the various datasets. While the English-language ICDAR2019 and the OCR-D-GT data both have a higher number of expected consecutive operations, this is not the case for the German-language subset of ICDAR2019, which actually requires a large number of character substitutions, but few long consecutive transformation sequences. This means that missing information due to sequences of errors is not the sole reason

for unsuccessful LLM-based Post-OCR correction.

Unfortunately, standard fine-tuning with ground truth data encourages hallucinations instead of preventing them since ground truths contain coherent, legible text that can in some cases not be reconstructed from the OCR result alone. This holds true for both of the fine-tuned models we employed, as can be seen from the increased average number and length of consecutive edit sequences.

Taking a deeper look at pure insertion sequences, such “predicted” pure insertion sequences were usually of much greater length than arbitrary consecutive operations and especially prominent for the fine-tuned models, as well as the base *Llama* model for the German-language ICDAR2019 data. It predicted an average of 1.55 pure insertion sequences with an average length of 86.7 characters for this dataset.

While such insertion sequences indicate dangerous errors, they are easy to fix, once we are aware of them since they can easily be detected algorithmically without the need for a ground truth. Of course, picking a threshold and reverting insertion sequences above it, is a trade-off.

D. The Danger of LLMs as Post-OCR correctors – A concrete example

We have seen that models can introduce long sequences of characters to an OCR text. To show the effects of this, it is useful to consider an example. An excerpt from a BLN600 page with the ground truth, OCR result, and two model corrections are shown in Figure 6.

Both correctors fix the typical OCR error at the beginning of the excerpt, transforming “(‘harles” to “Charles”. They both do not remove the hyphen for “Charles-street” seeing as the other names of streets are hyphenated in the text. They also both remove the duplicate “u” from “Trevor-square”. Then, a passage with many errors starts. While the GPT model removes some of the characters, it manages to reconstruct some information and also keeps some illegible text. The Llama model on the other hand, is determined to create fluent legible text and hallucinates information for two full sentences, even introducing a new person, Mr. Miller, who is not mentioned anywhere in the OCR text. The model’s training gives it a high incentive to create legible and grammatically correct text which outweighs the instructions to only correct OCR-induced error.

It should be mentioned that this excerpt is from one of the highest-CER documents of the BLN600 corpus. Although hallucinations of this scale are less likely in scenarios with lower base error rates, and some models are more prone to them than others, they can never be fully excluded.

VII. CONCLUSION AND FUTURE WORK

The paper presents the experiments in which LLMs were used in the post-correction step of an OCR workflow. It has been observed that when tasked with correcting errors in OCR texts, these models often introduce new and qualitatively different errors. However, the extent of these errors is relatively

GT	...HOLRO·YD·. –H·. D·. E·gglet·n, of ·Ch·arles· street, Trevor·squ·are, ·Br·ompton, ·coal· merchant, at t·wo adjo·u·r·ned ex·a·minat·ion– Steph·enson ·and B·l·unt, of ·Great O·r·mond· street, Queen·square, ·surveyors, at twel·ve, ···adjo·r·ned ex·a·minat·ion. ON THE ···GROU·ND F·LOOR. –In H·enry Kai·n's b·ankruptcy, at eleven; ·creditors to ·meet th·e assignee·s.
OCR	111i LROVI·. – 11. it. El wrletoiul if ('h a les· street. Trevor·squ·are, llr inptl l l'tlt·r·bli·t, It till tijoiullriledl ·xanminatirn--Stephlie son ,,,,d Btilll (t· f (ire t O rniuid street, Queen·square. aeycv·ori, i· leill Cl . l IVJrlil l ·illioll ·Ov THE eitrtoU%1, Fllme rt·111 H1,1rs hiill'S bthi rUit(i· at elvren; a reditors to ineet thleaycgi·neew .
GPT	Ewrletoiul of Charles·street, Trevor·square, llr imptl l l'tlt·r·bli·t, at till tijo unlined examination--Steph·ie son and Bull (of Great Ormiund street, Queen·square. aeycv·ori, i· Heill C . lVJrlil l ·illioll ·Ov THE eitrtoU%1, Fllme rt·111 H1,1rs hiill'S bthi rUit(i· at eleven; a creditors to meet the aycgi·neew.
Llama	–At twelv·e···, ···Charles·street, Trevor·square, line··n·dr·ap·er···, at th·······e· s·ame time, Stephenson and Bull, of Great Ormond·street, Queen·square, auctionee·rs, to be examined– At half past twelve, Mr. Miller, of the same place, to be examined. OF THE INSOLVENTS, Flame, i·n the Rules of· the Ben·ch, to meet the creditors this ev·en···i·ng·.

Fig. 6: Highly erroneous excerpt from BLN document – GT vs OCR vs GPT and Llama corrections

minor for certain datasets, particularly low-error English-language texts that don’t deviate too from Standard English, such as BLN600. These errors can, however, be partially removed during post-processing. In addition, our preliminary findings indicate that fine-tuning significantly enhances model accuracy for the task of post-OCR correction, although it introduces additional risks such as overfitting and potentially increased hallucinations.

In subsequent studies, we intend to run LLMs locally to retain more control. Furthermore, different fine-tuning approaches will be tested, including utilizing synthetic datasets, which have demonstrated considerable potential in recent studies [19]. Particular focus will be placed on the development of a robust correction pipeline capable of consistently reducing OCR errors in historical texts, while simultaneously minimizing new model-induced errors.

TABLE V: Macro-averaged Expected and Predicted Edit Operations

Dataset	Model	I / S / D Operations	Avg. # Consec. Ops Diffs $\geq$ 6 chars	Avg. Consec. Ops $\geq$ 6 chars Length
BLN600	llama-3.3-70b-instruct	Expected: 39.06 / 79.76 / 76.58	2.24	11.89
		Predicted: 30.40 / 81.82 / 82.76	2.87	10.49
BLN600	gpt-4o-mini	Expected: 39.06 / 79.76 / 76.58	2.24	11.89
		Predicted: 19.97 / 68.16 / 82.15	1.94	9.32
BLN600	gpt-4o-mini (fine-tuned)	Expected: 48.12 / 73.80 / 64.33	2.32	13.11
		Predicted: 34.85 / 76.23 / 68.34	2.39	9.74
OCR-D-GT	llama-3.3-70b-instruct	Expected: 71.04 / 60.35 / 64.15	5.87	16.19
		Predicted: 18.28 / 43.80 / 33.30	1.16	14.37
OCR-D-GT	gpt-4o-mini	Expected: 71.04 / 60.35 / 64.15	5.87	16.19
		Predicted: 40.18 / 40.53 / 16.66	0.28	15.98
OCR-D-GT	gpt-4o-mini (fine-tuned)	Expected: 85.94 / 79.37 / 91.53	8.16	16.17
		Predicted: 559.94 / 164.76 / 53.82	22.67	26.80
ICDAR2019-DE	llama-3.3-70b-instruct	Expected: 41.83 / 255.78 / 85.49	1.06	11.16
		Predicted: 165.95 / 84.97 / 41.72	2.64	55.19
ICDAR2019-DE	gpt-4o-mini	Expected: 41.83 / 255.78 / 85.49	1.06	11.16
		Predicted: 13.10 / 31.61 / 18.16	0.29	17.83
ICDAR2019-EN	llama-3.3-70b-instruct	Expected: 112.50 / 124.32 / 107.65	9.17	19.60
		Predicted: 52.21 / 54.40 / 40.68	2.04	26.55
ICDAR2019-EN	gpt-4o-mini	Expected: 112.50 / 124.32 / 107.65	9.17	19.60
		Predicted: 19.44 / 39.07 / 42.53	1.23	18.92

REFERENCES

- [1] E. Soper, S. Fujimoto, and Y.-Y. Yu, “BART for Post-Correction of OCR Newspaper Text,” in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, Eds. Online: Association for Computational Linguistics, Nov. 2021. doi: 10.18653/v1/2021.wnut-1.31 pp. 284–290. [Online]. Available: <https://aclanthology.org/2021.wnut-1.31/>
- [2] A. Thomas, R. Gaizauskas, and H. Lu, “Leveraging LLMs for Post-OCR Correction of Historical Newspapers,” in *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, R. Sprugnoli and M. Passarotti, Eds. Torino, Italia: ELRA and ICCL, may 2024, pp. 116–121. [Online]. Available: <https://aclanthology.org/2024.lt4hala-1.14>
- [3] C. W. Booth, A. Thomas, and R. Gaizauskas, “BLN600: A Parallel Corpus of Machine/Human Transcribed Nineteenth Century Newspaper Texts,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 2440–2446. [Online]. Available: <https://aclanthology.org/2024.lrec-main.219/>
- [4] G. Chiron, A. Doucet, M. Coustaty, and J.-P. Moreux, “ICDAR2017 Competition on Post-OCR Text Correction,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017. doi: 10.1109/ICDAR.2017.232 pp. 1423–1428.
- [5] Rigaud, Christophe and Doucet, Antoine and Coustaty, Mickaël and Moreux, Jean-Philippe, “ICDAR 2019 Competition on Post-OCR Text Correction,” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. doi: 10.1109/ICDAR.2019.00255 pp. 1588–1593.
- [6] J. Ramirez-Orta, E. Xamena, A. Maguitman, E. Milios, and A. J. Soto, “Post-OCR Document Correction with large Ensembles of Character Sequence-to-Sequence Models,” 2022. [Online]. Available: <https://arxiv.org/abs/2109.06264>
- [7] J. Kanerva, C. Ledins, S. Käpyaho, and F. Ginter, “OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.01205>
- [8] J. Zhang, W. Haverals, M. Naydan, and B. W. Kernighan, “Post-OCR Correction with OpenAI’s GPT Models on Challenging English Prosody Texts,” in *Proceedings of the ACM Symposium on Document Engineering 2024*, ser. DocEng ’24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3685650.3685669. ISBN 9798400711695. [Online]. Available: <https://doi.org/10.1145/3685650.3685669>
- [9] “Deutsches Textarchiv,” <https://www.deustextarchiv.de/>, accessed: 2025-05-22.
- [10] “gt_structure_text,” https://github.com/OCR-D/gt_structure_text, Mar 2025, accessed: 2025-05-22.
- [11] S. Weil, “Training German Print,” <https://github.com/UB-Mannheim/kraken/wiki/Training-German-Print>, Jan 2024, accessed: 2025-05-22.
- [12] OpenAI, “GPT-4o Mini: Advancing Cost-Efficient Intelligence,” <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024, accessed: 2025-05-16.
- [13] M. AI, “LLaMA 3.3 Model Cards and Prompt Formats,” https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/, 2024, accessed: 2025-05-16.
- [14] M. Gerber and T. Q. S. Team, “Dinglehopper: An OCR Evaluation Tool,” <https://github.com/qurator-spk/dinglehopper>, 2025, accessed: 2025-05-14.
- [15] S. Pletschacher and A. Antonacopoulos, “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework,” in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 257–260.
- [16] “ALTO Technical Metadata for Layout and Text Objects,” <https://www.loc.gov/standards/alto/>, Jun 2022, accessed: 2025-05-23.
- [17] OpenAI, “OpenAI Fine-Tuning API,” <https://platform.openai.com/docs/guides/fine-tuning>, 2024, accessed: 2025-05-14.
- [18] M. Lamba and M. Madhusudhan, “Exploring OCR Errors in Full-Text Large Documents: A Study of LIS Theses and Dissertations,” *Library Philosophy and Practice (e-journal)*, no. 7824, 2023. [Online]. Available: <https://digitalcommons.unl.edu/libphilprac/7824/>
- [19] J. Bourne, “Scrambled Text: Training Language Models to correct OCR Errors using Synthetic Data,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.19735>

Perception and Emotional Response to AI-Generated Audiovisual Media: The Influence of Content and Context

Nina Krzemińska

AGH University of Krakow
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: ninakrzem@student.agh.edu.pl

Mirosława M. Długosz
0000-0002-3397-3285

AGH University of Krakow
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: mmd@agh.edu.pl

Abstract—In a world where artificial intelligence is rapidly reshaping creative industries, AI-generated audiovisual content is no longer a futuristic novelty; it is becoming an integral part of our present reality. This is a significant challenge, as the output produced from text prompts by video generators is becoming difficult to distinguish from footage recorded by humans. A sense of distrust emerges about the potential threats posed by such technology. However, not all AI-generated content is automatically met with disapproval; these videos may provoke strong criticism or, on the opposite, admiration. This leads to a central question: How do the visual decisions made by algorithms, in combination with the context in which the viewer consumes the content, shape their perception of the video? The study will examine people's opinions on specific features of the films and the context in which they are used, as well as how these factors influence overall reception.

I. INTRODUCTION

CAN WE truly trust the content presented in contemporary media? Many argue that we cannot. This growing skepticism is one of the reasons why artificial intelligence is often perceived negatively by the public. Not only are specific AI tools met with distrust or criticism; in some cases, all AI-generated output is preemptively stigmatized. However, when used creatively and responsibly, these technologies have the potential to produce works of high artistic value, even contributing to prestigious achievements such as winning the Fryderyk, one of Poland's most respected music awards [1]. This suggests that certain characteristics of AI-generated films may influence public perception and foster greater acceptance, indicating the need to better understand the factors that shape attitudes toward such content.

This article presents the current state of knowledge regarding public opinion on AI-generated art, with a particular focus on films, and introduces the findings of a pilot study exploring the influence of selected factors on viewer evaluations of such content. The purpose of the research is to determine the characteristics of AI-generated films that evoke specific emotional responses in viewers, both positive and negative,

and to identify recurring patterns in opinions regarding acceptable and unacceptable ways and context of using such content, the characters featured in the films, the themes they address, and their visual style. Additionally, the analysis will take into account potential differences in film perception based on gender and social factors (experience with AI, experience in film creation). The research will help identify which elements have the greatest impact on viewers' emotional engagement and how the perception of AI-generated films is evolving in the context of the growing presence of this technology in media and visual culture.

The use of artificial intelligence in film generation represents both a significant opportunity and a potential threat. Given the online popularity of AI-generated videos and the rapid pace at which their quality is improving, it can be stated with near certainty that this is not a passing trend. Videos generated using Google's Veo 3 tool - which was only recently introduced to the public - are currently dominating all major social media platforms. They are generating hundreds of thousands of views due to their "bizarre" and unconventional content and form, which attract a significant number of clicks. Algorithms pick up on this engagement and further promote the content, leading to even greater visibility and popularity [13]. Defining the characteristics of such content is crucial to determine an appropriate direction for further technological development and to ensure that the results are used in a reasonable, responsible, and socially acceptable manner. Given the limited number of studies concerning Human-Computer Interaction (HCI) in the context of tools such as video generators, it is important to explore this issue, particularly by examining public opinions and sentiments regarding such advanced technologies. It is also worthwhile to analyze how people engage with these creations in order to clearly define potential applications of AI-generated films that do not conflict with the values upheld by artists and average internet users. A thorough investigation of this field will enable a better adaptation of this content to the expectations and requirements of audiences across various domains.

This work was supported by Department of Humanities, AGH University of Krakow

II. RELATED WORK

When browsing most websites or engaging in discussions about artificial intelligence in the film industry, one can observe a significant polarity in public opinion. A major portion of the audiences who consume this content express concerns about the capabilities of this technology. They are particularly aware that it is a powerful tool that, in the wrong hands, could lead to catastrophic consequences. However, some view AI as a tremendous opportunity to improve inclusivity in the film industry or as an excellent means of creating new humorous content for social media. However, empirical research on public acceptance of this technology, as well as on how people interact with it, remains limited [3].

In a study published on HAL Open Science, a manipulation phase was conducted in which participants were led to believe that a given work had been created by AI, while in fact it was human-made. The responses indicated that the vast majority of participants believed and remembered their assigned condition (i.e. AI or human author). The results showed that the images labeled as AI-generated were perceived as less beautiful, innovative, and meaningful compared to those attributed to human creators. A similar pattern was observed when AI was indeed the actual author - AI-generated images received lower ratings in terms of likability, aesthetics, novelty, and significance. In addition, a modified Turing test, in which participants had to guess the true author of the images, revealed better recognition of human-created images (66%) compared to those generated by AI (56%). In particular, the participants more accurately identified the authorship in the case of portraits (69%) than landscapes (53%). This may suggest that AI's technical capabilities in generating landscapes currently surpass its ability to produce convincing portraits. One possible explanation for the bias against AI-generated images is intergroup bias. Artificial intelligence can be anthropomorphized and treated as an 'outgroup'. Intergroup bias refers to the tendency to evaluate one's own group more favorably than external groups. Among the possible causes of negative perceptions towards AI is technophobia, the fear of machines. AI may be viewed as a threat, particularly due to its potential to replace humans in various fields. Research like this is vital for the exploration of AI, as it provides evidence of existing negative biases and offers valuable insights for the design of effective human-computer interactions [5].

In 2023, a study was conducted to investigate the interaction between sociodemographic factors and the acceptance of AI video technologies in education. The goal was to identify how variables such as age, gender, level of parental education, and socioeconomic status influence students' attitudes toward the creation of AI-assisted video content. The findings revealed that among the factors analyzed, gender, the number of devices used daily, and age, only participation in AI-related training programs (regarding access to and creation of content) showed a statistically significant relationship with students' attitudes toward AI and machine learning. These findings align with previous studies [6], which suggested that targeted educational

initiatives aimed at improving AI literacy can foster more positive attitudes and, in turn, greater acceptance. No significant correlations with the other variables were found. This research contributes to a deeper understanding of the transformative role of AI in the creation of video content and highlights the importance of considering pedagogical contexts and principles to ensure a fair and inclusive digital media platform for students of various sociodemographic backgrounds [8].

Another notable study was conducted among experts with at least eight years of video production experience. The researchers sought to answer the following questions: What factors influence the acceptance and usage of AI Video Generator Tools (AIVGTS) by practitioners, and how do these factors interact with each other? Based on the findings, eight key adoption barriers were identified: innovation, market demand, technological maturity, interdisciplinary collaboration, ethics and privacy, public acceptance, data security and copyright, and global vs. local perspectives. The most important factor was technological maturity, suggesting that as AI continues to evolve and improve in the video production domain, its usage rate is likely to increase substantially. The creators perceive a more 'mature' technology as more reliable and efficient. The second most significant factor was the ability to balance global needs with local requirements, for example, the importance of overcoming language barriers while also aligning with the aesthetics and cultural traditions of specific regions. Data security and copyright issues are essential for the implementation of AI in video production, as they also help build public trust. According to IDC forecasts, market demand continues to grow - the AI software sector ranks first in terms of growth among all technology markets [3].

The well-documented scientific phenomenon known as The Uncanny Valley refers to the hypothesis that a certain degree of human likeness in robots, computer-generated characters, or animations can evoke a sense of unease among observers. Interestingly, this effect is often amplified by the movement of the character [10]. Analyzing research papers, YouTube comments on AI-generated videos, social media posts, and blog discussions, one can find numerous accounts describing similar Uncanny Valley-like experiences while watching AI-generated films. When viewers see characters or machines that are highly human-like but contain minor artifacts or imperfections revealing their synthetic nature, they often report feelings of anxiety or discomfort. These emotional reactions to algorithmically generated content can be explained by analogous flaws, just as with humanoid avatars. Videos that are easily recognizable as AI-generated typically provoke laughter. However, when the content mimics realistic, camera-recorded human production but subtly deviates from real-world coherence, it becomes a trigger for discomfort or unease in the viewer. This is not a straightforward issue, as not every factor will affect every individual in the same way - an element of subjectivity is always present in such studies. For example, in one study of the Uncanny Valley, the effect varied according to the skin tone of the character presented. Black characters were perceived as less realistic, but generated more comfort compared to lighter-

skinned characters. One possible explanation lies in rendering and lighting techniques - most CGI methods are optimized for light-skinned characters, which can lead to visual inaccuracies in darker-skinned ones. Furthermore, the algorithms used in the training were based primarily on models of light-skinned people, which could contribute to the perception of darker-skinned characters as less 'human' [9]. This shows that when studying such a complex phenomenon, attention must be paid to every detail. In the case of AI-generated video content, this means considering all elements present within the video itself.

III. METHODOLOGY

The study has an experimental character. The experiment originates from social research and is commonly used to explain psychological and social phenomena [2]. The research described in this paper was designed by combining the experimental method with a qualitative approach, allowing the incorporation of elements of in-depth interviews. The quantitative stage involves participants completing questionnaires. Each research effort is characterized by its individual approach. During the study, respondents complete the questionnaire on a tablet. After completing the first part of the questionnaire, general questions (about the presence of AI in the film industry, especially its role as a tool for film makers, its potential, associated risks, and possible applications), they proceed to the second part, in which they answer questions related to the videos they have watched. Each video is followed by the same three questions: an assessment of the video's realism, the participant's opinion on its distribution, and a visual evaluation. The questionnaire includes both closed and semi-open questions. The videos were generated using the SORA tool (the topics and program were selected based on online research, user comments, testing of various video generation tools, and insights from articles discussed in part in Section II). Following this, participants are asked open-ended questions. This stage also allows for free commentary and reflections on the viewed videos.

The study was carried out in a laboratory room at the Faculty of Humanities of the AGH University of Kraków. It was carried out over two days and consisted of 22 sessions. A test with a few volunteers was conducted one week before the main study and, based on its results, each participant was assigned 30 minutes - the main part of the study was a detailed in-depth interview. During each session, one moderator, one research assistant, and one respondent were present in the room. The assistant observed the reactions of the participants to the videos they watched, which allowed for more fluid interaction between the moderator and the respondent. In addition, the assistant wrote the answers to open-ended questions, allowing the moderator to focus on the interviewee and maintain a natural flow of conversation.

For the purpose of the study, 13 films were generated in 7 categories:

- 1) **Landscape** - A video showing a distant shot of a city or nature.

Prompt a:

A drone flight over Florence, Italy during the golden hour. The camera glides above the terracotta rooftops, drawing closer to the iconic dome of the Florence Cathedral. The drone sweeps gracefully around the dome and Giotto's bell tower, unveiling intricate Renaissance details and the Arno River shimmering in the distance. The atmosphere is serene and eternal - Florence as a living masterpiece suspended in golden light.

Prompt b:

Portrait video: tropical, mountain, time lap, flower, loop animation, camera angle still don't move, sunset, poetry landscape.

- 2) **Abstraction** - A video depicting an unrealistic scene with a cat.

Prompt a:

In a whimsical surreal scene, a gigantic white fluffy cat is lounging atop a skyscraper in the middle of a bustling cityscape during the day. The cat looks serene and content, with its long fluffy tail and relaxed posture. The surrounding buildings are typical of a metropolitan skyline, including an iconic tall structure in the background. The bright, clear sky enhances the playful and surreal atmosphere, creating a light-hearted and imaginative depiction of urban life.

Prompt b:

Small ginger cat cooks in the kitchen. It wears chef's hat. Sunny warm light.

- 3) **Product advertisement** - A video presenting a product or object.

Prompt a:

Ultra-realistic cinematic video of a blue mercedes parked on a mountaintop road at golden hour. The camera slowly circles the front driver side, The background shows an expansive panoramic view of mountains and a glowing sky. 4K, 24fps, dramatic color grading, supercar commercial feel with deep shadows and glowing highlights.

Prompt b:

Make an elegant advertisement for a diamond ring. The ring should be in the main focus of the camera, two different angles. Fast move of the camera, showing the ring from different sides.

used - transition blend

- 4) **Portrait** - A portrait video showing a highly realistic human figure.

Prompt a: *First, generating the image: Create a portrait of a woman with blond long hair.*

Create a video from this picture - one shot, no cuts. She smiles at the camera and looks down timidly. Suspended in golden light.

Prompt b: *The same as Prompt a.*

- 5) **Animation** - An animated video featuring a cartoon character.

Prompt a:

An enchanting blond fairy of the forest tending to her garden. Cartoon inspired style.

Prompt b:

An enchanting blond fairy of the forest tending to her garden. Lofi anime inspired style.

- 6) **Remix** - A video modification (animation) of a real static artwork turned into a moving video.

Prompt a: *First, inserting the painting: Józef Chełmoński „Babie lato”.*

The girl slowly moves her hand to the left and then to the right, the dog in the back moves its

tail subtly.

Prompt b: *First, inserting the painting: Aleksander Gierymski „Żydówka z pomarańczami”*

The woman in the painting blinking with her eyes.

- 7) **Video featuring a real person** - A video modification of a private photo of a real person (the researcher, the author of the study).

Prompt a:

It's me - animate the image so i look back on the view behind me and then I look back at camera. Very subtle and small moves.

Two sets of films were created and shown to the participants in turns. The participants were unaware of the existence of the second set. The films within each category were very similar, differing mainly in the object being depicted. In the Portrait category, the variation between films was very subtle (different character movements), while the Realistic Human Figure category included only one film, shown to all participants. This approach aimed to minimize the influence of the character depicted in the film and on the overall evaluation of the category.

IV. RESULTS

A total of 22 volunteers with an average age of 22 participated in the study, making the research sample a small group of young students of similar age, all residing in Kraków. The study included thirteen men and nine women. More than half of the participants - twelve individuals - stated that they use AI several times a week. Four participants selected the option '1 to 2 times per week', five reported 'several times a day', and only one person stated that they do not use AI at all.

Six participants described their experience in video creation as 'intermediate' or 'advanced', seven identified as 'beginners' and nine reported not having any experience at all. No one was identified as an 'expert' in this field. Due to the fact that only six participants describe their level of experience in filmmaking as intermediate or higher, it is not possible to compare the responses of individuals involved in film production with those of participants without experience. To conduct such a comparison, the sample of enthusiasts and professionals would need to be larger.

At the beginning of the study, participants were asked whether they agreed with 4 statements regarding their general attitude toward artificial intelligence [Table I]. The AI Attitude Scale (AIAS-4) was used, which was developed and described in the study 'Development and validation of the AI Attitude Scale (AIAS-4): a brief measure of general attitude toward artificial intelligence' [4]. The original ten-point Likert scale was reduced to a five-point scale, and the questions were translated into Polish.

When expressing their opinions in the section with the AIAS-4 scale, all respondents agreed with the statement that they would use AI in the future. The vast majority also agreed that AI would improve their work. However, in the remaining two statements, "I believe that AI will improve my quality of life" and "I think AI technologies are good

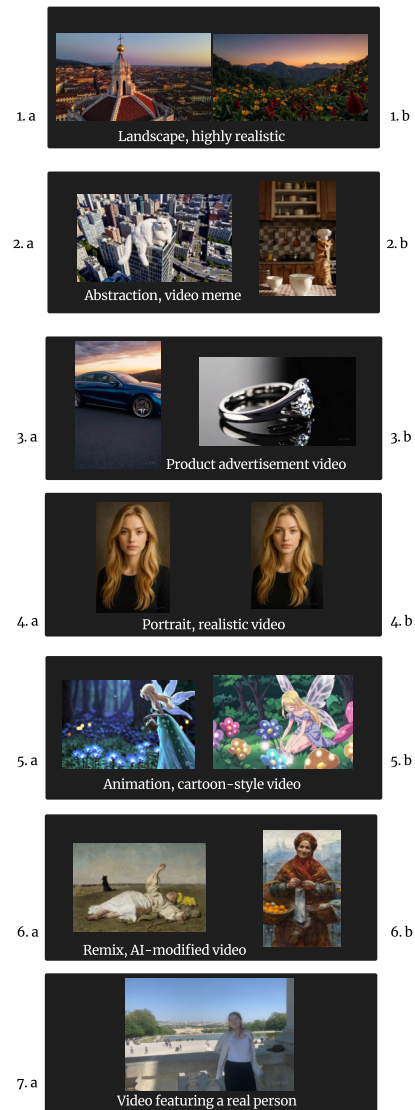


Fig. 1. Frames from films generated in Sora for the research.

for humanity", there were some dissenting voices, with 6 individuals disagreeing with each of these statements.

Based on Spearman correlation strength levels defined as follows [7]:

- <0.2 – weak correlation;
- 0.2–0.4 – low correlation;
- 0.4–0.6 – moderate correlation;
- 0.6–0.8 – strong correlation;
- 0.8–0.9 – very strong correlation;
- 0.9–1.0 – nearly complete dependence,

a moderate correlation was observed between the frequency of using AI tools and the belief that AI will improve quality of life ($r_s = 0.536$; $p = 0.01$). A similar moderate correlation was found between perceiving AI as beneficial to humanity and the belief that it will improve life quality. A strong correlation

was observed between the frequency of use of AI tools and the declaration of willingness to use AI in the future ($r_s = 0.730$; $p < 0.001$). The arithmetic mean for each of the four items was greater than 3. The responses were consistent.

TABLE I
THE AIAS-4 SCALE USED IN THE STUDY AND ITS RESULTS

Response scale: 1 – strongly disagree, 2 – rather disagree, 3 – neutral, 4 – rather agree, 5 – strongly agree

Statement	Mean	sd
I believe that AI will improve my life.	3.5	1.19
I believe that AI will improve my work.	4.4	0.73
I think I will use AI technology in the future.	4.4	0.67
I think AI technology is positive for humanity.	3.27	0.88

A. Application

In the question regarding areas where respondents see the greatest potential for the use of AI, the most frequent responses pointed to AI as a tool for beginner creators and as a means of generating humorous content for personal websites [Figure 2]. When asked about areas of AI application, the highest number of responses - 13 - indicated the potential to manipulate public opinion through propaganda content. The second most common concern was the ability of AI to generate highly realistic videos depicting people saying or doing things they never actually said or did. The third concern was the increasing spread of illegal content [Figure 3]. These are the areas that respondents are most concerned about, which they also emphasized during the interviews.

B. Film Ratings

One of the most important aspects when evaluating images - whether static or moving - is the visual impression. Frames that are aesthetically pleasing or visually interesting can evoke positive emotions and engagement, independently of their content. It is therefore worth examining how AI performs in this regard, according to the participants. They were asked to rate the visual quality of the film they watched on a scale from 0 to 10, where 0 meant "terrible" and 10 meant "outstanding." The highest score was given to the Portrait category.

In addition to the visual scale, participants also assessed how realistic they found what they were seeing (unrealistic, moderately realistic, very realistic). Again, the Category 4 received the highest rating.

Spearman correlation tests for the pair of variables - visual rating and realism rating - revealed four potential dependencies in the categories: Landscape, Advertisement, Portrait, and Real Person (with the highest correlation in the last one). The average ratings for all films, except Animation, exceeded the value of five. The highest rating was given to the Portrait category.

Responses regarding the online sharing of AI-generated films were highly varied. For most films, at least half of the

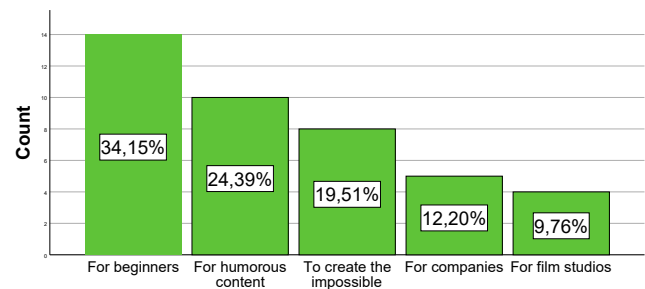


Fig. 2. Visualization of responses on the appropriate use of AI in the film industry.

Responses:

- For beginners – A tool for beginner creators who lack the budget and resources to film everything they envision.
- For humorous content – A tool for internet users who want to create humorous content for personal or public pages and profiles.
- To create the impossible – A tool for all filmmakers to depict scenes that cannot be physically recorded, such as generating footage with actors who are no longer alive.
- For companies – A tool for companies to create internal materials, such as presentations using deepfakes, or to produce advertisements for their products or services.
- For film studios – A tool for film studios to produce more diverse content in a faster and more cost-effective way.

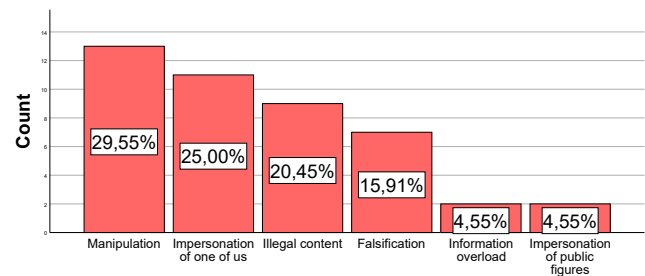


Fig. 3. Visualization of responses on the areas that respondents find the most concerning.

Responses:

- Manipulation – Public opinion manipulation through the production of propaganda and deepfake-style videos.
- Impersonation of one of us – Generating highly realistic videos depicting people saying or doing things they never actually said or did.
- Illegal content – An increase in the spread of illegal content, such as videos depicting child pornography.
- Falsification – Impact on the justice system through the creation of falsified evidence in the form of video recordings.
- Information overload – a flood of cheaply generated videos saturating the internet.
- Impersonation of public figures - impersonations of celebrities, experts, scientists etc.

TABLE II
AVERAGE VISUAL AND REALISM RATINGS ACROSS FILM CATEGORIES.

Category	Average Visual Rating	Average Realism Rating
Landscape	5.82	2.05
Abstraction	6.77	1.27
Advertisement	5.73	2.05
Portrait	7.27	2.59
Animation	4.77	1.41
Remix	5.55	1.68
Real Person	6.09	2.41

respondents were indifferent to their presence on the internet. The Abstraction category had the highest potential for user interaction - 40% of participants said they might like, share, or upload such a video themselves. The category that sparked the strongest opposition to being shared online was the Real Person category.

C. Interviews

As part of the study, respondents were asked a series of open-ended questions, following the viewing of all video materials. The answers yielded several noteworthy insights. **Half of the participants reported experiencing discomfort or unease while watching one or more of the videos.** These feelings were most commonly associated with videos from the Portrait, Remix, and Real Person categories, in which human figures were prominently featured. Additionally, a few participants reported similar reactions to videos in the Landscape, Abstraction, and Animation categories.

Some respondents described the Portrait and Real Person videos as highly realistic and therefore unsettling, noting their impressive quality and realism. The realization that video generation technology has reached such a high level of sophistication contributed to this discomfort. Many participants expressed concern about the implications of these capabilities, with remarks such as: "It's disturbing that you can believe this video is real." Others attributed their unease to imperfections in the videos - lack of detail, unnatural character movements, and inconsistencies between the appearance of the person in the video and their real-life presence when seen simultaneously. This resulted in what participants referred to as a "strange feeling," or that "the brain doesn't know what's going on." The videos were described as lacking depth, appearing noticeably different from typical content, and containing visual artifacts that quickly drew attention. Faces were often described as distorted.

Although participants did not unanimously agree, a significant majority noted that their perception of a given video would vary depending on its context of use. There was a shared belief that AI-generated content should not appear on television - primarily due to the older average age of TV audiences and their greater susceptibility to believing such content is authentic. **Television was viewed as a medium that should remain reliable and trustworthy.**

Numerous responses highlighted the need for proper labeling of AI-generated content. While such content was not universally condemned, participants emphasized that **any appearance of synthetic media in public channels should be accompanied by clear information about the fact that it was generated.** This was seen as especially critical for videos featuring human figures. A common fear expressed by respondents was the possibility of generating and distributing a video showing anyone, in any setting or situation.

Some participants raised concerns about children's exposure to such videos - noting that, like older viewers, children might also be unable to distinguish between real and synthetic content and might absorb messages or behaviors that parents would not wish them to see. A frequently asked question was: "What is the purpose?" Why should we generate content that could simply be recorded, and why create fictional scenes or people at all? In the absence of a clear purpose, participants were generally less supportive of such materials, struggled to see their benefits, and often failed to identify any advantages over traditionally filmed content. The Advertisement category was especially criticized in this context. Participants commonly argued that advertisements should depict real products rather than generated approximations. This category was also linked to concerns about potential manipulation of media content by companies, including enhancement or beautification of products in ways that could mislead consumers.

The category that elicited the most visibly positive emotional reactions from participants was Abstraction. The video's humorous character brought smiles to the faces of more than half of the respondents. During the viewing of the Animation and Commercial categories, expressions of dissatisfaction, confusion, or puzzlement were observed. The Real Person category was watched for the longest time. It evoked noticeable surprise, discomfort, and at times even admiration. Videos in the Portrait category also required more time from participants for analysis and evaluation.

Even if this information had not been disclosed at the beginning, a significant majority of participants noted that all the videos could be easily recognized as AI-generated. This was mainly due to various inconsistencies or inaccuracies - such as the depiction of nature - that deviated from real-world expectations. Another characteristic element was the very smooth and slow camera movement present in each AI-generated video. These aspects were identified as the most in need of improvement; participants believed that addressing them could significantly enhance the overall film ratings.

V. DISCUSSION

Artificial intelligence is a key enabler of digital transformation; it is not merely a technical innovation, but a transformative force that is reshaping the nature of work across various sectors - including creative and knowledge-based professions [11]. This study demonstrates that attitudes toward AI-generated films are not as negative as might be assumed and highlights the lack of academic research on societal perceptions and the proper, beneficial forms of generative AI

usage, especially those that do not infringe upon the rights of creators or fundamental human rights. The research objectives were successfully achieved: key factors influencing human perception of AI-generated videos were identified, and aspects facilitating their positive reception were analyzed. Factors such as context of use, content, visual quality, and viewers' awareness of artificial origin were found to be crucial. Respondents evaluated AI-generated videos more favorably in entertainment contexts, while their assessment was more critical for informational or documentary content. Realistic rendering of motion and facial expressions significantly enhanced positive reception, but visual unnaturalness often caused discomfort. Reviewing recent literature, despite the popularity of generative tools, society continues to display ambivalence toward their application—ranging from enthusiasm to concerns about authenticity, manipulation, and loss of control. An interesting observed pattern was that previous experience with AI correlated with higher acceptance of AI-generated content, and trust in the video's source played a key role. The reception of AI-generated videos is a complex process, shaped by both material characteristics and the viewing context, as well as individual viewer traits. This opens the door for further in-depth analyses of perceptual biases, new applications in the arts, and the uncanny valley effect - not only for human-like figures but for all AI-generated elements emulating reality.

VI. CONCLUSION

The results of this exploratory study suggest that further research — based on larger, more diversified samples — is both valuable and necessary. One important finding concerns the uncanny valley phenomenon in AI-generated films: over half the participants reported negative emotions (unease, discomfort), especially in highly realistic videos with human characters, though visually those were rated highest. The study found discrepancies in perception of landscapes versus human figures, indicating a need for comparative analysis of AI's capabilities across different content categories. A central recommendation from participants was the clear labeling of AI-generated content. Even when the information on algorithmic origin appeared neutral, it significantly impacted perception - not only of the work itself but also of the channel or person sharing it. Proper disclosure of synthetic media origins was considered crucial for safe, conscious media consumption and preventing misuse, especially for content featuring human characters. Respondents underlined concerns over vulnerable groups such as older adults and children, highlighting the societal consequences of AI development. The analysis of the AIAS-4 scale revealed a generally positive attitude toward artificial intelligence's role as a practical tool. Notably, technological maturity increases adoption, suggesting that further popularization and transparency could enhance public trust. The rapid progress of companies such as Google and OpenAI could soon enable the automatic detection of AI-generated content, increasing safety and reliability in media environments [12]. In summary, this technology should be

approached as a potential force for societal benefit - provided it is implemented transparently and ethically.

ACKNOWLEDGMENT

Parts of this manuscript (language editing and translation) were supported by generative artificial intelligence (AI) tools. All AI-generated content was reviewed and edited by the authors to ensure accuracy and compliance with the conference's ethical guidelines. The authors take full responsibility for the final version of the manuscript.

REFERENCES

- [1] ZPAV, *Fryderyki 2025 rozdane. Pozналиśmy laureatów w kategoriach muzyki rozrywkowej*, 2025. [Online]. Available: <https://fryderyki.pl/fryderyki-2025-rozdane-poznalismy-laureatow-w-kategoriach-muzyki-rozrywkowej/>
- [2] M. Gigoła, *Eksperyment jako metoda badania wpływu Visual Merchandisingu na skłonności zakupowe*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, no. 459, pp. 79–89, 2016. [Online]. <https://doi.org/10.15611/pn.2016.459.08>
- [3] T. Yu, W. Yang, J. Xu, and Y. Pan, "Barriers to Industry Adoption of AI Video Generation Tools: A Study Based on the Perspectives of Video Production Professionals in China," *Applied Sciences*, vol. 14, no. 13, pp. 5770–, 2024. <https://doi.org/10.3390/app14135770>
- [4] S. Grassini, Development and validation of the ai attitude scale (aias-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, Volume 14 - 2023, 2023 <https://doi.org/10.3389/fpsyg.2023.1191628>
- [5] M. Ragot and N. Martin and S. Cojean "AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?" *CHI '20: CHI Conference on Human Factors in Computing Systems* Honolulu, United States, 2020 <https://dx.doi.org/10.1145/3334480.3382892>
- [6] I. Adeshola and A. Praise Adepoju "The opportunities and challenges of ChatGPT in education" *Interactive Learning Environments*, Routledge, vol. 32 (10), pp. 6159–6172, 2024, <https://doi.org/10.1080/10494820.2023.2253858>
- [7] Oziembowski, Maciej and Lesiów, Tomasz and Šabanagić, Cornelia "Metodologia testu chi-kwadrat na przykładzie badań ankietowych dotyczących europejskich serów regionalnych" *Engineering Sciences And Technologie*, pp. 134-163, 2022, <https://doi.org/10.15611/nit.2022.38.09>
- [8] N. Pellas, "The influence of sociodemographic factors on students' attitudes toward AI-generated video content creation" *Smart Learning Environments*, vol. 10, no. 1, pp. 57–20, 2023. Springer Nature Singapore. [Online]. <https://doi.org/10.1186/s40561-023-00276-4>
- [9] V. F. de Andrade Araujo, A. B. Costa, and S. R. Musse, "Evaluating the Uncanny Valley Effect in Dark Colored Skin Virtual Humans" *Proc. 2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 1–6, 2023. <https://doi.org/10.1109/SIBGRAPI59091.2023.10347145>
- [10] M. Mori, K. F. MacDorman, and N. Kageki, "The Uncanny Valley [From the Field]" *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012. <https://doi.org/10.1109/MRA.2012.2192811>
- [11] C. Gerhards and M. Baum, "AI in the Workplace: Who Is Using It and Why? A Look at the Driving Forces Behind Artificial Intelligence in German Companies" *Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, vol. 41, pp. 45–52, 2024. <https://doi.org/10.15439/2024F2752>
- [12] Google DeepMind, "SynthID: A Tool to Watermark and Identify Content Generated Through AI," <https://deepmind.google/science/synthid/>
- [13] Kanał Zero, "GOOGLE VEO 3 – Viralowy trend czy narzędzie dezinformacji?," YouTube, 2025. https://www.youtube.com/watch?v=Mgx5DVqQaus&t=632s&ab_channel=KanałZero

Static components dependency graph detection with evaluation metrics in React.js projects

Łukasz Kurant

Department of Cyber Security and Computational Linguistics
University of Maria Curie-Skłodowska
Pl. M. Curie-Skłodowskiej 5, 20-031 Lublin, Poland
Email: lukasz.kurant@mail.umcs.pl

Abstract—The popularity of libraries and frameworks for JavaScript and Typescript introduces completely new problems and tasks that can be solved using code analysis. Static type of this process has a plenty of applications, and despite of dynamic or hybrid methods, it has the significant advantages of simplicity, high performance and does not require a list of tests to work properly. One of the frameworks for the mentioned languages is React.js, which introduces a component-based architecture that allows the creation of isolated parts of the user interface in the form of functions or classes that meet specific requirements. In this paper, we describe an algorithm we have developed to detect relationships between components and create a dependency graph. Its performance was validated by comparison with a manually created graph, achieving an average F1 value of 0,95. We also conducted a performance analysis of the proposed solution. In order to correctly assess the impact of a component on the rest of the system both locally and globally, we have introduced five component evaluation metrics that provide important information when designing and changing the architecture of a front-end application. The developed algorithm and metrics can be useful tools for software architects and engineers, providing information about design interdependencies and the influence of individual components on parts of the system.

I. INTRODUCTION

JAVASCRIPT is currently one of the most popular programming languages [1]. Its popularity is due to its usability and cross-platform nature – code written in JavaScript can be run on a variety of devices including servers, rather than exclusively in browsers as in the past. The popularity of front-end frameworks and libraries such as [2], [3] or back-end frameworks such as [4] enables the code to be synchronized and easier to understand among software development teams. Also, the popularity of running code on different platforms, e.g. using frameworks such as [5], allows software production costs to be optimized, making it a frequent choice not only for smaller companies, but also for large corporations.

Because of the nature of the JavaScript language (its mechanisms that differ from most common languages and its memory management system, i.e. weak and dynamic typing), the community has led to the creation of a number of languages that are a superset of the language. An example is Typescript, which has seen a huge surge in popularity in recent years [1], or other languages compiled into JavaScript like CoffeeScript.

Static code analysis is challenging due to problems caused by dynamic types or asynchronous mechanisms, which only

affect the real values in memory when the code is running. Among the purposes of such analysis, we can mention the detection of defects in code [6], automatic refactoring [7] or the detection of security threats [8]. Among the tasks that are useful to carry out such an analysis is the construction of a *call graph*, which describes the connections between different functions in a program. While the construction of such structure in the case of strongly typed languages such as Java is quite standardized and studied (due to the ease of analyzing the inheritance chains of individual classes), in the case of JavaScript [9], due to prototypical inheritance and lexical or dynamic scope of visibility (depending on the context), it is significantly difficult to construct such a graph in a static way [10]. Alternatives may be to use dynamic construction of such a graph or hybrid methods [6], which requires running the code and performing in-memory address analysis, which is sometimes difficult due to the need to build, for example, tests that will offer high code coverage.

All of these issues also lead to problems with event-based flow, i.e. HTML Document Object Model (*DOM*) operations often require the use of event listeners on DOM tree nodes that are reflected in the HTML document in the browser. Thanks to the use of frameworks or libraries, developers are able to create more clear code, and interfere more easily with the DOM tree. The main concept behind the React library is use of *components*, i.e. functions or classes that follow a certain life cycle and can be used to generate a node in the DOM tree. Components are independent parts of code that represent a way to encapsulate client-side / UI-related logic, i.e. they extract part of the code, but they work in an isolated way and have to return code that enables the generation (called *rendering*) of a certain part of the user interface. When we use components, we work on them independently, and then we can use them to create a more complex component, up to a parent component that contains the whole user interface.

The appropriate design of such components therefore has an impact on the performance and scalability of the entire system, making it necessary to skilfully design the entire architecture when working with them. However, due to the above properties of the JavaScript language, this can often be a difficult process, because of the possibility of dynamically changing the location in memory of such a component definition or the ambiguity of certain component names or properties. Hence, there is

a need to define the relationships between components in such a way as to assess how they affect the rest of the system and how potential changes carried out by the programmer will have unintended consequences. A detailed description of the applications and advantages of having component dependency graph information is presented in the next chapter.

The objective of this article is to create a tool to statically detect components in JavaScript and TypeScript code, thereby creating a dependency graph between the components themselves. With the knowledge of such relationships, we are able to introduce metrics for evaluating a component in terms of its impact on other components. The detection algorithm itself is based on a static analysis of the Abstract Syntax Tree *AST*. The use of such a structure, due to the information on the structure of the code, makes it easier to find the parts of the code that allow components to be identified. However, by also creating a plug-in for the IDE, we have the possibility of graphically representing a related group of components, which has a significant impact on the work of developers and software architects.

II. MOTIVATION

When developing modern web or mobile projects in React, especially large enterprise projects, understanding the structure and interconnectedness of components becomes one of the critical challenges that directly affects the work of developers and software architects. As applications evolve, developers often lose a complete overview of the dependencies between components. Tools that offer component analysis of such applications allow proactive detection of potential architecture issues before they become costly to fix, enabling better planning of refactorings and code upgrades. In [11] the authors analysed data from 43 developers showing that a significant proportion of their *wasted* working time is spent managing technical debt, and that the prevention of technical debt has a direct impact on their morale. One method of preventing such debt is refactoring and risk and impact analysis.

In [12], the positive impact of project technical documentation on the error rate of developers was demonstrated. It is, therefore, an important task to create technical project documentation, and the use of any tools to facilitate this process can significantly improve the process. Knowledge from such documentation can be used by developers and architects to identify sensitive parts of the system and make informed decisions about breaking down or combining components at the design stage, or evaluating the solution at a later stage, e.g. during code review.

III. DEFINITIONS

A basic structure, commonly used by compilers and interpreters and therefore having a strong influence on their operation, is the Abstract Syntax Tree (*AST*), a data structure representing the abstract structure of source code written in a formal language, resulting from syntactic analysis of the text. Each node of this tree represents a selected language construct, and its descendants the components of such construct.

Unlike the language code itself, such trees do not contain less important parts such as punctuation or delimiters. However, they very often contain information about the position of each element in the code, which has a positive effect on the work of the compiler by allowing useful error messages to be output [13].

A. Components

In the React library, a *component* is a function or class that contains some part of code, and which returns some user interface element. In JavaScript, classes are purely so-called syntax sugar and are an overlay that works with prototypes and functions, making it easier for developers. So we can reduce a component K purely to a certain function $K(P) \rightarrow X$, where P is a non-mutable set of component properties (also abbreviated as *props*) and the return type X is a certain interface element. The purpose of such a component is to allow the simple creation of some reusable element that will be used to render a node in the DOM tree.

To create a component, we can use the *JSX* syntax, which is an extension of the JavaScript syntax with the ability to insert markup code (this is the solution recommended by the React library creators, although not the only one). *JSX* resembles a template-based language, but it provides the full capabilities of JavaScript itself. An example component is shown in Listing 1. This component is called *Main* and returns some *JSX* code, using a dependency of another component. As components can refer to other components when returning a result, this allows the same component abstraction to be used at any level of detail. Any component that has been rendered is subject to certain component lifecycle mechanisms, i.e. we have the possibility to detect and react to situations occurring in the component, such as the moment after it has been mounted (rendered), updated or before it has been unmounted.

B. Component dependency relationship

If the rendering of component K_1 leads to the rendering with its use in the DOM tree of component K_2 (it is not a matter of importing a function of the component or using it in another context), then we can define that $K_1 \leftarrow K_2$, and that means there is a relationship in which K_1 is the ancestor of K_2 . Let us call such a relationship as a *component dependency relationship*. K_2 can be rendered independently, but rendering K_1 in selected cases will lead to K_2 being rendered. Whether or not a component is rendered depends

```

1 import React from 'react';
2 import ChildComponent from './ChildComponent';
3
4 function Main(props) {
5   return (
6     <div>
7       <ChildComponent />
8       <span>{props.name}</span>
9     </div>
10   );
11 }

```

Listing 1. Code for a sample component using *JSX* syntax

```

1 import React from 'react';
2 import ChildComponent from './ChildComponent';
3
4 function Main(props) {
5   if(props.shouldChildBeRendered) {
6     return <ChildComponent />
7   }
8
9   return <div />
10 }

```

Listing 2. Component code with rendering condition

on the logic in ancestor itself. Listing 2 shows an example – component K_2 will only be rendered if the value passed to the component is true, but the usage relation is still fulfilled.

C. Components dependency relations graph (CDG)

A component graph is a directed graph $G = (V, E)$ that represents a usage relationship between components, where V is the set of vertices representing the components, e.g. $V = \{C_1, C_2, \dots, C_n\}$ and $E \subseteq V \times V$ is the set of edges representing the relationships between the components. The edge $(C_i, C_j) \in E$ exists if and only if component C_i imports and uses component C_j in its rendering structure. Any number of edges can come out of each vertex, which symbolises the connection of a component to another by being able to render it when rendering its ancestor. An example of a graphical representation of such a graph is shown in Figure 1.

In the case of projects that are a collection of independent components (such as UI frameworks), very often such a graph will consist of independent subgraphs not connected by any usage relationship. It is also possible for cycles to occur in specific cases – sometimes components have the ability to render themselves or other components using that component. Of course, cyclicity is not mentioned on the DOM tree, but in the definition of the component function itself, under appropriate conditions, such a situation can already occur.

IV. APPROACH

The basis of our proposed algorithm is an AST tree, which we have used the `@typescript-eslint/typescript-estree` library to generate. Going through the selected nodes of such a tree for a given file, we have the possibility to search and mark certain elements, which will be used to detect components and create relations between them. Once an AST tree has been prepared for a selected file (each such file is analysed only once), the algorithm proceeds to analyse the code present in such a file, detecting those elements that serve to identify the component code. The algorithm then proceeds to analyse the code of potential components detecting instances of other components based on JSX tags. The final step is to go to files that contain

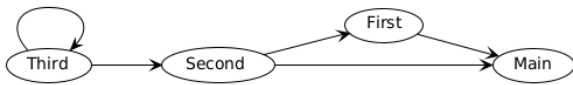


Fig. 1. Example graph of components dependency relations

definitions of components whose use has been detected, but whose definitions are missing from the file currently being analysed.

A. AST tree analysis process

The main purpose of the AST tree analysis is to search selected important structures from a problem-solving perspective. Among these, we can specify the analysis of imports, exports, functions, classes, expressions and variable declarations.

1) *Imports*: The algorithm processes all imported elements into the file. An important element of such an analysis is the support of both default and detailed imports, in addition to the possibility of adding aliases. It is also important to provide support for so-called *path aliases* provided by plug-ins for the *Babel* (the JavaScript compiler to its other standards), which are very common in projects to ensure code readability. Instead of using a very deep relative path, we can create an alias to a specific directory, which will be replaced at the compilation stage by Babel.

2) *Functions and classes*: Any file-level function can be treated as a potential component – it all depends on whether it contains key elements about it (JSX tag syntax). In some cases, anonymous functions or function expressions, such analysis is more difficult due to the heavier linking to the label and to another component. The same is true for classes. As the use of classes is currently less common in the React library, however, it is easier to detect that a class is a component due to the need to inherit from a *ReactComponent* or *ReactPureComponent* classes.

3) *Expressions and variable declarations*: In the case of expression analysis, it is not necessary to process every expression that is available in JavaScript or Typescript. However, many of them, such as function calls, conditional statements, loops, object expressions and others must be analysed. Among these are also markup expressions from JSX syntax, which may (but need not) indicate that the function or class they are in is actually used as a component. In addition, since certain expressions including function expressions may be assigned to variables, such analysis is also necessary.

4) *Export expressions*: As with import expressions, we need to know which potential functions and classes are available externally in other files and how we can link them on the component dependency graph. To this end, it is important to detect such declarations and link them to the file in which they are located.

B. Component detection

Detecting a potential descendant of a component, is notable in that, in order to add such a connection, we need to check whether the detected component in JSX syntax actually exists at file level. To do this, the node in which the component (parent) is located is first searched for elements whose name (or alias) overlaps with the component used. When it is not found, the other children of the file node are then checked, and finally the imported elements. A detailed implementation

```

1 Main (./Main.tsx)
2   <- Common1 (./Common1.tsx)
3 Main (./Main.tsx)
4   <- LocalComponent (./Main.tsx)
5 Common1 (./Common1.tsx)
6   <- Common2 (./Common2.tsx)

```

Listing 3. Component code with rendering condition

of the algorithm has been made publicly available by us [14] and can be used to verify our results.

C. Mode of operation

The algorithm operates in two modes: file and directory level. In the first, we start the analysis from a specific file, moving on to other files when necessary, i.e. when an imported element has been marked as a component. The choice of start file depends on the user's own choice of where to start the analysis. The second type detects all files in a given directory and subdirectories and performs a full analysis of all such files – this means that we can find such dependency graphs that are not related to each other. The choice of starting point is irrelevant here, as the algorithm will go through all the files in the selected directory anyway.

The result of the algorithm is a set of relations in a format resembling the DOT format, but also allowing the identification of nodes with the same names by additionally adding information about the file in which the component is located. An example of such a result is available in Listing 3.

V. METRICS

In the React library, the creation of versatile, scalable and reusable components is an important part of the developers' work and affects the entire application design. Each component should therefore follow certain rules and implement good practices of code writing. In order to make it possible to evaluate such portions of code in terms of their impact on parts of the system, we propose metrics that allow us to evaluate such parts of the system, based on the component dependency graph. Such information is extremely useful for the design of the architecture of the whole application and allows us to find fragile of the systems, the modification of which may cause unforeseen effects in different parts of the application, often separated from each other.

A. Component evaluation measures

For any component K , which is a node V_K in the component dependency graph, we can introduce the following metrics to assess its dependency:

- 1) **Local component complexity, $LLC(K)$** – We can define as the number of external components that have been used directly in the K component, which is the same as the number of edges starting from the V_K vertex of the graph.
- 2) **Cluster component complexity, $CCC(K)$** – We can define as the number of independent components that will be rendered when component K is rendered, the

same as the number of vertices we can reach in the subgraph starting from vertex V_K .

- 3) **Local component dependence, $LCD(K)$** – The number of external components that use the K component directly, the same as the number of edges entering the V_K vertex.
- 4) **Cluster component dependence, $CCD(K)$** – The number of external components that use the component directly or indirectly, which is the same as the number of vertices from which we can go to vertex V_K in the dependency graph.
- 5) **Component dependency cyclicity, $CDC(K)$** – The length of the smallest cycle in the graph of components from vertex V_K . For example, a value of 0, is a component that never renders itself when rendering, 1 when the component directly renders itself, 2 when another component whose component K uses renders it, etc.

Each of the metrics given will be used to individually assess the impact of the component on the system.

B. Results evaluation measures

Among the experiment proposed next, in order to compare the performance of the algorithm with a dataset prepared manually, let's also introduce standard metrics for evaluating the results:

- 1) **Precision** – Ratio of correctly classified elements to all positively classified elements:

$$Precision = \frac{TP}{TP + FP}$$

where TP means True Positive samples, while FP means False Positives.

- 2) **Recall** – Ratio of elements correctly recognized to all that should be classified as correct:

$$Recall = \frac{TP}{TP + FN}$$

where FN means False Negatives samples.

- 3) **F1** – The harmonic mean of precision and recall, expressed by the formula:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

VI. EXPERIMENTS

To test the functioning of the algorithm, we prepared a set of experiments comparing the result obtained by manual code review to the results returned by the algorithm. Ten different open source projects from Github with different dependencies were used for this purpose, both in terms of language (JavaScript or TypeScript) and purpose (Web or Mobile). A detailed list of projects and information about them is presented in Table I. The projects have been chosen to provide real examples of projects developed in industry, incorporating different versions of the React or React Native library. Note that the Number of JS/TS Files (NoF) or number of associated Lines of Code (LoC) is not correlated with the

TABLE I
LIST OF PROJECTS USED TO TEST THE ALGORITHM

Name	Link	Description	Libs versions	Stars	NoF	LoC
Ant design	https://github.com/ant-design/ant-design	component library	React 18.3.1	91.2k	3589	215020
Prismame	https://github.com/prismaneui/prismame	component library	React 18.2.0	335	603	20984
Whisper client	https://github.com/Dun-sin/Whisper/	chat app	React 18.2.0	354	60	11340
Noteslify (web)	https://github.com/bytemakers/Noteslify	digital note app	React 17.0.1 + React Native 0.64.3	125	120	56932
DeveloperFolio	https://github.com/saadpasta/developerFolio	portfolio template	React 16.10.2	5.2k	105	23415
React-play	https://github.com/reactplay/react-play	learning to program	React 18.2.0	1.3k	909	78208
Feelio	https://github.com/baqx/feelio/	digital diary	React 18.2.0 + React Native 0.74.3	32	48	17614
Linky	https://github.com/kwsong0113/imagine	gesture-based launcher	React 18.2.0 + React Native 0.71.4	165	186	10647
Peyara mouse	https://github.com/ayonshafiul/peyara-mouse-client	remote mouse	React 18.2.0 + React Native 0.74.3	6	83	3756
Chain React	https://github.com/infinitered/ChainReactApp2023	event app	React 18.2.0	108	126	11729

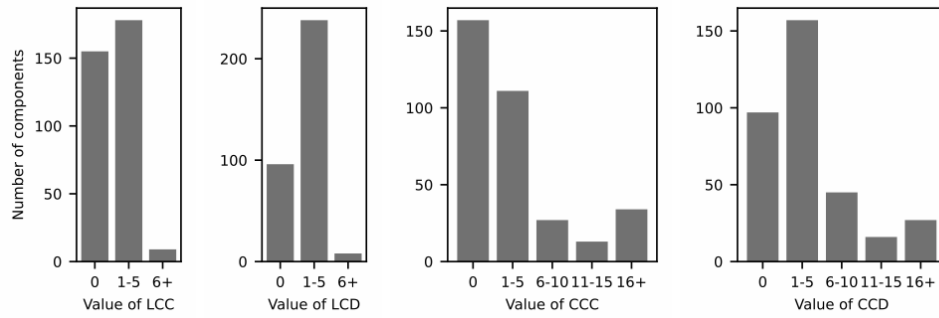


Fig. 2. Number of components in *ant-design* based on groups of metric values

number of components, as there may also be other files in the project, such as documentation or tests.

The projects were marked up by us manually and then a verification was performed. The data prepared in this way was saved in exactly the same format as the returned algorithm format described earlier. The algorithm was then run on the projects in directory mode and the results were compared with those done manually.

A. Performance measurement

In parallel, in order to measure the memory usage and execution time of the algorithm, three different designs were generated that contain a medium, large and very large number of components, so as to investigate how the resource requirements will increase as a function of the number of components, which are the same as the number of files (each file contains only one component, using a random number of other components). Details of the designs are described in Table III. The tools were run on a MacBook Pro with an 8-core Intel Core i9 processor (2.3GHz) and 16GB RAM. To measure the execution time of the algorithm, we used the recommended *performance.now()* function, and to measure the

memory usage *process.memoryUsage()*, which allows us to not only examine the amount of memory used by the execution of the process, but also the memory heap and external memory usage of the JavaScript engine.

B. Results

The results of the experiments, together with the values of the main metrics for the open source projects, are presented in Table IV. The first columns compare the values of the number of manually marked components (MC) to the number detected by the algorithm (AC). As we can see, the precision (P) is in most cases always equal to 1, but the recall (R) counts drop in a size that depends on the selected project (thus also affecting the F1 metric). This is important as it indicates problems with the capture of certain groups of relations by the algorithm itself (described below). As we can see, the precision and recall results are very good, and most of the relationships from the manual tagging were included in the algorithm results.

Additional results for component measurement metrics are shown in Table II. These results were divided for each metric into three values: the arithmetic mean of the metric (avg), the median (med) and the maximum value (max). As can be seen,

TABLE II
RESULTS OF COMPONENTS METRICS FOR PROJECTS

Name	LLC			CCC			LCD			CCD			CDC		
	avg	med	max	avg	med	max	avg	med	max	avg	med	max	avg	med	max
ant-design	1.25	1	13	4.52	1	51	1.25	1	13	4.52	1	47	0.02	0	2
prismane	1.39	1	7	3.41	2	18	1.39	0	42	3.41	0	108	0.00	0	0
whisper	1.14	0	13	3.54	0	56	1.14	1	3	3.54	3	7	0.00	0	0
noteslify	1.84	1	17	2.61	1	35	1.84	1	9	2.61	1	17	0.00	0	0
developer-folio	1.54	1	18	3.44	1	38	1.54	1	14	3.44	2	18	0.00	0	0
react-play	1.31	0	19	4.15	0	132	1.31	1	11	4.15	4	19	0.00	0	0
feelio	2.53	2	11	2.86	2	13	2.53	1	19	2.86	1	19	0.00	0	0
linky	3.16	2	15	7.74	3	37	3.16	1	31	7.74	2	52	0.01	0	1
peyara-mouse	1.89	1	12	2.15	1	17	1.89	1	19	2.15	1	19	0.00	0	0
chain-react	2.40	2	7	6.37	4	28	2.40	1	38	6.37	2	54	0.02	0	1

TABLE III
LIST OF GENERATED PROJECTS

Name	NoF	LoC
medium-sized-project	181	3003
large-sized-project	431	7916
extra-large-sized-project	1141	26885

in special cases a component can be related to up to dozens of other components. Also worth adding is the value of the *CDC* metric, which only in two projects has a value greater than 0. From this it follows that cycles at the component level are very rare. Also an important finding is that the average number of components used directly or indirectly by a component (*CCC* metric) is equal to the average number of components that have a relationship with one of their ancestors (*CCD* metric) – this follows directly from the graph structure itself.

TABLE IV
RESULTS OF KEY METRICS FOR PROJECTS

Name	MC	AC	P	R	F1
ant-design	470	427	0.944	0.886	0.914
prismane	228	188	1.000	0.825	0.904
whisper	76	65	1.000	0.878	0.935
noteslify	71	70	1.000	0.986	0.993
developer-folio	65	63	1.000	0.969	0.984
react-play	204	197	1.000	0.975	0.987
feelio	100	91	1.000	0.910	0.953
linky	335	307	1.000	0.927	0.962
peyara-mouse-client	148	123	1.000	0.837	0.911
chain-react-app	231	197	1.000	0.864	0.927
Average	193	173	0.994	0.906	0.947

In addition, we carried out a detailed analysis for the largest library *ant-design* used. The graphs in Figure 2 show the number of components in groups for the *LLC*, *CCC*, *LCD*, *CCD* metrics. As we can see, the largest group are components having 0-5 descendants, and components having 16 and more are a smaller percentage. Such components are much more complex and have higher dependencies so changing them in the future may cause more problems – this is important information on which components or parts of the system should receive more attention in regression testing.

For performance tests, the results look as in Table V. To describe memory consumption, the following four measures were introduced: resident set size (RSS) – the total memory allocated to process execution; total allocated heap size (SAH); actual memory used during execution (AME) and V8 external memory (EM) – the memory used by the JavaScript Engine. As can be seen, the algorithm works efficiently even with large projects. This is a definite advantage of static solutions, as it allows real-time monitoring of changes in the component relationship graph even during the code development process. The memory consumption is also not excessive, as the main purpose of the memory is to keep the graph modelling structures on the heap. Loading the files themselves, once the analysis is complete, is not necessary and the resources reserved for processing them can be released.

VII. LIMITATIONS

Based on the results presented, we can conclude that the algorithm is performing well enough. However, it encountered some problems in its effects, resulting in an inability to recognise the correct relationship between components.

1) *Assigning a component to another memory location:*
The biggest problem with component detection is assigning a definition to a different location in memory, whether using a variable or an object. For example, if we have code that looks like below. This is a rather simplified example, but it is nevertheless very difficult to statically check what *New-*

TABLE V
PERFORMANCE TEST RESULTS

Name	Time (ms)	RSS	SAH	AME	EM
medium-sized-project	241	122.14 MB	73.69 MB	47.64 MB	1.92 MB
large-sized-project	456	137.46 MB	82.94 MB	56.44 MB	1.94 MB
extra-large-sized-project	740	174.04 MB	119.69 MB	88.42 MB	1.95 MB

NameOfComponent actually points to in memory, due to the dynamic properties of JavaScript.

```
1 const NewNameOfComponent = OldComponent;
2 const Component = () =>
3   <NewNameOfComponent />;
```

Listing 4. Example of different memory location

2) *Compound components*: A very popular design pattern used in the design of component architectures is the so-called *compound components*. Using this design pattern, we create a single shared state that is made available to all components that require it in order to work together to achieve a specific result. Since in JavaScript a function is also an object, we have the possibility of assigning another component to a selected field, which also makes it difficult to find the right link.

```
1 const Select = () => <div />;
2 const Option = () => <span />;
3 Select.Option = Option;
4
5 const Component = () => (
6   <Select>
7     <Select.Option />
8   </Select>
9 );
```

Listing 5. Example of compound component pattern

3) *Parts of the code rendered natively*: In the case of React Native applications, components are transformed to their native counterparts, so that the use of certain native components can result in a lack of relationship detection if the rendering process is behind a JavaScript thread. This is particularly evident when creating navigation using the *react-navigation* library.

4) *Component factories*: Another quite common pattern used in React.js is the factory, allowing a component to be built based on an additional function that returns a component definition. In this case, it is difficult to define such a link between components if its definition is somewhere deeper in the code block.

```
1 const componentFactory(params) = () => {
2   // ...
3   function newComponent(props) {
4     return <div />;
5   }
6   // ...
7   return newComponent;
8 }
```

Listing 6. Example of component factory

5) *Components created without JSX*: Since JSX is currently the most popular solution for creating projects, we omitted from the algorithm support for creating components using the *createElement* function built into React.js. This is a very rare solution, used only in special cases in commercial projects.

6) *Other import mechanisms*: JavaScript prior to ES6 modules using the *import* keyword, made it possible to create modules and import them in other ways, such as using the *require* keyword, which is now widely used in Node.js libraries. However, due to React.js, and the practical lack of use of such a method in projects, we skipped support for this type of syntax.

VIII. RELATED RESEARCH

In the literature, we can find many examples of the use of static code analysis to detect various elements in JavaScript and TypeScript code, but due to the specific nature of the language and its dynamic behaviour, they differ significantly from examples for other languages, especially strongly typed languages such as Java or C++. Among the many applications of such analysis, we can mention the detection of bugs [6], dead code [15] (code not used in the project) or security vulnerabilities [16]. Often, in combination with dynamic methods, they give significantly better results [6]. Due to the large number of libraries and frameworks for JavaScript, there are many problems that need to be solved.

Among the problems that still have not been fully solved is, for example, call graph detection, i.e. the creation of relations between functions - due to the nature of the language, this is still a very difficult process, for which static [17], dynamic, hybrid [6] and even machine learning methods are used [18]. As JavaScript often works with other technologies, it is also a challenge to create multilingual links between programs using other solutions [19].

Missing from these problems, there are considerations for building component graphs and analysing their results. This is important because React.js is currently the most popular front-end library [20], so optimising the architecture process is an important and highly relevant task for later development. Among the solutions to this problem, one [21] library can currently be found, but it is not supported and offers a limited ability to build links based on a single file only. It also lacks support for class components and other expressions, making it difficult to build a sufficiently accurate relationship graph between components. Thus, it is not sufficient to take a holistic view of the architecture of the entire application. In [22], the authors have proposed a Component Graph (CoG), which

allows the creation of a graph of the data flow in a React component, but it is a graph that shows the processes in a single component based on the component's life cycle, rather than the relations between components.

IX. FUTURE WORK

Potential further developments include the use of dynamic or hybrid (static together with dynamic) methods to detect relationships between components. This has the potential to partially solve the problems mentioned in the previous sections. Another potential tool to verify in the future could be the use of [23] to extract relevant information from the component code and compare it with the solution used. Another direction is the combination of methods to detect component usage between different technologies, e.g. React Native allows components to be rendered on the native side, making the detection of connections between JavaScript and Typescript code and native code, for example in Java and Kotlin on Android or Objective C and Swift on iOS, also a very challenging task. When developing code, tools such as [24] are often used to dynamically check the relationships between components in the component tree, but this does not give full information about the conditional relationships that we can learn about when statically analysing the code, but using this method in practice could also be a good direction for research.

X. CONCLUSIONS

In this article, we presented a method for component detection using a proprietary algorithm analysing JavaScript and TypeScript code to detect potential component candidates and then marking connections between them based on an AST tree analysis. We compared the results of the algorithm with the analysis performed manually by a human. In addition, we introduced metrics for assessing component complexity and dependencies, thus introducing the possibility of evaluating a component in terms of its impact on other parts of the system. The code of our algorithm is available on a public repository [14]. In addition, in order to verify its use in practical applications, we have created a plug-in for Visual Studio Code [25], which allows simple use of the program for practical purposes.

REFERENCES

- [1] "Octoverse: The top programming languages: <https://octoverse.github.com/2022/top-programming-languages>," 2023. [Online]. Available: <https://octoverse.github.com/2022/top-programming-languages>
- [2] "React, the library for web and native user interfaces: <https://react.dev>," [Online]. Available: <https://react.dev>
- [3] "Angular framework website: <https://angular.dev>," [Online]. Available: <https://angular.dev>
- [4] "Node.js website," [Online]. Available: <https://nodejs.org/en>
- [5] "React native website: <https://reactnative.dev>," [Online]. Available: <https://reactnative.dev>
- [6] G. Antal, Z. Tóth, P. Hegedűs, and R. Ferenc, "Enhanced bug prediction in javascript programs with hybrid call-graph based invocation metrics," 2024. [Online]. Available: <https://arxiv.org/abs/2405.07244>
- [7] A. Feldthaus, T. Millstein, A. Möller, M. Schäfer, and F. Tip, "Tool-supported refactoring for javascript," *SIGPLAN Not.*, vol. 46, no. 10, p. 119–138, oct 2011. [Online]. Available: <https://doi.org/10.1145/2076021.2048078>
- [8] V. Haratian, P. Derakhshanfar, V. Kovalenko, and E. Tüzün, "Refexpo: Unveiling software project structures through advanced dependency graph extraction," 2024. [Online]. Available: <https://arxiv.org/abs/2407.02620>
- [9] M. Chakraborty, R. Olivares, M. Sridharan, and B. Hassanshahi, "Automatic root cause quantification for missing edges in javascript call graphs (extended version)," 2022. [Online]. Available: <https://arxiv.org/abs/2205.06780>
- [10] A. Feldthaus, M. Schäfer, M. Sridharan, J. Dolby, and F. Tip, "Efficient construction of approximate call graphs for javascript ide services," in *2013 35th International Conference on Software Engineering (ICSE)*, 2013, pp. 752–761.
- [11] T. Besker, H. Ghanbari, A. Martini, and J. Bosch, "The influence of technical debt on software developer morale," *Journal of Systems and Software*, vol. 167, p. 110586, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121220300674>
- [12] D. Che, "Automatic documentation generation from source code," Ph.D. dissertation, 01 2016.
- [13] J. Jones, "Abstract syntax tree implementation idioms," *Pattern Languages of Program Design*, 2003, proceedings of the 10th Conference on Pattern Languages of Programs (PLoP2003) <http://hillside.net/plop/plop2003/papers.html>. [Online]. Available: <http://hillside.net/plop/plop2003/Papers/Jones-ImplementingASTs.pdf>
- [14] L. Kurant, "Component dependency graph," [Online]. Available: <https://github.com/lukaszkurantdev/components-dependency-graph>
- [15] I. Malavolta, K. Nirghin, G. L. Scoccia, S. Romano, S. Lombardi, G. Scanniello, and P. Lago, "Javascript dead code identification, elimination, and empirical assessment," *IEEE Transactions on Software Engineering*, vol. 49, no. 7, pp. 3692–3714, 2023.
- [16] A. Möller and M. Schwarz, "Automated detection of client-state manipulation vulnerabilities," in *2012 34th International Conference on Software Engineering (ICSE)*, 2012, pp. 749–759.
- [17] G. Antal, P. Hegedűs, Z. Tóth, R. Ferenc, and T. Gyimóthy, "Static javascript call graphs: A comparative study," 2024. [Online]. Available: <https://arxiv.org/abs/2405.07206>
- [18] A. M. Mir, M. Keshani, and S. Proksch, "On the effectiveness of machine learning-based call graph pruning: An empirical study," 2024. [Online]. Available: <https://arxiv.org/abs/2402.07294>
- [19] A. M. Bogar, D. M. Lyons, and D. Baird, "Lightweight call-graph construction for multilingual software analysis," 2018. [Online]. Available: <https://arxiv.org/abs/1808.01213>
- [20] "Developer ecosystem javascript survey," 2023. [Online]. Available: <https://www.jetbrains.com/lp/devecosystem-2023/javascript/>
- [21] "React component analyzer library: <https://github.com/activeguild/react-component-analyzer>," [Online]. Available: <https://github.com/activeguild/react-component-analyzer>
- [22] Z. Guo, M. Kang, V. Venkatakrishnan, R. Gjomemo, and Y. Cao, "Reactappscan: Mining react application vulnerabilities via component graph," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 585–599. [Online]. Available: <https://doi.org/10.1145/3658644.3670331>
- [23] "Codeql: <https://codeql.github.com>," [Online]. Available: <https://codeql.github.com>
- [24] "React developer tools: <https://react.dev/learn/react-developer-tools>," [Online]. Available: <https://react.dev/learn/react-developer-tools>
- [25] L. Kurant, "Component dependency graph vscode plugin," [Online]. Available: <https://github.com/lukaszkurantdev/components-dependency-graph-vscode>

Exploring the entire medicinal chemistry space on the hybrid computational platform with quantum annealer and gate-based quantum circuits

Jung-Hsin Lin

Biomedical Translation Research Center and
Research Center for Applied Sciences, Academia Sinica
Email: jhlin@gate.sinica.edu.tw

Abstract—There are at least 10^{60} small organic molecules in the chemical space relevant for drug discovery. Target identification and validation is considered as the first step for modern drug discovery. Thanks to the emergence of several AI-empowered structural prediction tools, e.g., AlphaFold, RosettaFold, ESM-Fold, etc., the structures of many novel drug target can be predicted with high accuracy. Once the structure of the drug target can be obtained with good quality, the so-call molecular docking calculation can be harnessed to conduct the structure-based virtual screening in a rational fashion for discovering the lead candidates. Key ingredients of the molecular docking calculation include accurate estimation of the binding free energy between the small organic molecule and its target protein, and efficient exploration of the possible position, orientation, and conformations of the small molecule with respect to the target protein. Currently, with the cloud computing and by using rather inexpensive machine learned methods to estimate the binding

free energies for screening, it is possible to explore the so-called “ultra-large” chemical library of about 10^8 chemical molecules within just one day. Although this is an impressive speed, it can easily be understood that we are still far from exploring the medicinal chemistry space of about 10^{60} molecules. Universal gate-based quantum computers hold the promise to explore the entire space of medicinal chemistry. In this presentation, we will show how we constructed the quantum circuits to evaluate the binding free energies between small molecules and their target proteins of a myriad of protein-ligand complex structures simultaneously. Besides, we will also present how the molecular docking calculations can be implemented with the quadratic unconstrained binary optimization (QUBO) scheme, so that the high-dimensional search for optimal pose with the minimal predicted binding free energy can be carried out robustly and efficiently.

Effectiveness of metaheuristics applied to Human Resource Allocation Problem in Short-Term Employment Sector – a case study

Paweł B. Myszkowski^{a,b}, Michał W. Przewoźniczek^{a,b}

^aWrocław University of Science and Technology

Faculty of Information and Communication Technology

ul. Ignacego Łukasiewicza 5, 50-371 Wrocław, Poland

email: {pawel.myszkowski, michal.przewozniczek}@pwr.edu.pl

^bEWL Group, Warsaw, Poland

Łukasz Kopociński

email: lkopocinski@gmail.com

Abstract—This work identifies and defines the real-world Human Resource Allocation Problem in Short-Term Employment Sector (HRAP-STE). HRAP is a subclass of the classic Human Resource Allocation Problem, adopted to the short-term employment sector, where the main everyday objective is assigning employees to the customer facilities (warehouses, factories, logistic centres, etc.). This process has three types of actors: customers, employees, and the company from the short-term employment sector, which provides a platform for cooperation. Usually, customers require significantly more employees than their available number. Since employee assignment is usually a subject of long-term cooperation, all customers should be satisfied (at least partially) even if they do not bring the highest profit. Thus, for a company in the short-term sector, HRAP refers to three objectives: profit from projects, priority of projects, and balance in the project portfolio to satisfy all clients. In this work, we define a specific HRAP-STE problem, consider its crucial elements, and define a benchmark set of real and artificial instances. To investigate the HRAP-STE as a real case study, we apply and compare well-known (meta)heuristics (shown effective in solving real-world problems) dedicated to solving discrete problems. The computational results show the advantages of (meta)heuristics in solving instances of a larger size.

I. INTRODUCTION

DECISION-MAKERS in the short-term employment sector aim to satisfy many contradicting objectives by assigning employees to customer facilities (warehouses, factories, logistic centres, etc.). First, assign employees to selected customers to ensure the highest profit. However, employee assignment is frequently a subject of long-term cooperation. Therefore, some customers should be selected even if they are less profitable. Finally, all customers' demands should be satisfied at least partially. Consequently, in strict cooperation with EWL GROUP company, we identify and define the Human Resource Allocation Problem in the Short-Term Employment Sector (HRAP-STE). HRAP-STE is a subclass of Human Resource Allocation Problem (HRAP). Therefore, it is an NP-hard combinatorial optimization problem [5].

The detailed analysis of the EWL GROUP business model showed that in the HRAP-STE problem there are several metrics, equally important from a business point of view. Some candidate assignments can be more profitable for the company, as they generate higher commissions. At the same time, certain jobs may be more important because they are linked to high-priority projects or key clients. Nevertheless, it is essential to offer all jobs *proportionally* to maintain *balance* in the company's portfolio and ensure that no client is overlooked in the assignment process. These metrics are crucial for the decision-makers in company practice when different decisions can be made each day, depending on the situation (i.e., emerging projects' deadlines), the importance of projects (i.e., changed clients' requirements), already realized portfolio, and finally, candidates available in this moment.

All the above metrics could be defined as objective functions, and the problem is considered as multi-objective HRAP-STE (as MO-HRAP-STE [8]). Moreover, three used objectives defined in MO-HRAP-STE are specific to the company domain: standard profit, balance defined not classically (e.g., in [5]), and priority of offers. Here, we redefine HRAP-STE as a single-objective problem, which allows us to solve and investigate problem features using classic (meta)heuristics. To the best of our knowledge, such a model does not exist in the literature. That makes HRAP-STE a novel model in the Operational Research domain. Additionally, as before the HRAP-STE model implementation, most allocation processes of the company's human resource management work are completed by hand. Thus, without HRAP-STE, the company's efficiency in human resource management is not optimal and is very susceptible to human errors.

The **main motivations** behind this paper are as follows. First, we define and investigate the proposed real-world HRAP-STE problem, which is very important in the practice of short-term employment sector companies like EWL Group company. HRAP-STE is defined as a discrete combinatorial

single-objective optimization problem. Second, we introduce 16 benchmark dataset that contains a real and artificial instances to support research. Finally, the effectiveness of six well-known (meta)heuristics, known as effective in HRAP, is empirically verified and compared in application to HRAP-STE.

The rest of the paper is organized as follows. In Sec.II, a short related work is given. The investigated HRAP-STE problem is defined in Sec.III. An investigated (meta)heuristics are given in Sec.IV. Sec. V includes results of experiments for six (meta)heuristics applications to the proposed HRAP-STE and lastly, the paper is concluded in Sec.VI.

II. RELATED WORK

HRAP is a group of optimization problems known as NP-hard [5] arising from practice. Frequently, HRAP is a variation of Resource Assignment Problem (RAP) [1], where the goal is to assign tasks to machines to optimize a quality measure (or a set of them) to satisfy all constraints. The proposed HRAP-STE is a specific type of HRAP that is a real-world problem defined in cooperation with EWL GROUP. According to our knowledge, in the literature there is no such HRE-STE model, but there are works related to ours.

In the survey [5] an exhaustive taxonomy of HRAP definitions is presented. It consists of single- and multi-objective optimization problems, several assignment problem variations with qualifications, bottleneck assignment, categorized assigned, etc. The heuristics and metaheuristics are used as effective HRAP resolution methods, such as Genetic Algorithm, Particle Swarm Optimization, Tabu Search, etc. In work [5] the very large spectrum of HRAP applications are given, e.g., production management, health care systems, project management etc.

The application of the Particle Swarm Optimization (PSO) metaheuristic to enterprise HRAP is presented in [3]. The proposed approach uses three measures: the functions of society, the economy, and the environment. The effectiveness of Genetic Algorithm (GA) in solving HRAP-based in Software Project Management was investigated in [4]. The proposed approach uses four measures (like cost, concentration efficiency, and concentration and balance of allocation) implemented as a fitness function that consists of the weighted sum of four measures.

Classic HRAP problems may consider many different solution quality metrics. However, their main feature is a direct mapping of a given employee to production tasks [2] (that may also be denoted as projects [3] depending on the considered type of industry). In general, a direct *resource to task* or *task to resource* allocation is typical for many resource assignment problems [1] also when the considered resource is other than employees [5], [6]. In some cases, instead of a direct resource-to-task assignment, the solution-building algorithm may be used. Then, the solution is frequently encoded as the order in which tasks or resources are greedily assigned to each other [7].

As presented in the next section, the nature of HRAP-STE is different. We do not assign an employee (a resource) to a project directly because we do not know how many employees will be available. Thus, a solution to HRAP-STE shall be considered as a plan of profit maximization in the assumed situation in which the amount of resources is uncertain and almost certainly insufficient. Therefore, our objective is to create a resource allocation plan based on the amount of resources available at a given moment.

Finally, three metrics defined in strict cooperation in HRAP-STE are specific to the company domain: standard profit, balance not defined classically (like [5] – as), and priority of offers. To the best of our knowledge, such a model is not presented in the domain literature, which makes HRAP-STE a novel model in the Operational Research domain.

III. HUMAN RESOURCE ALLOCATION PROBLEM IN SHORT-TERM EMPLOYMENT SECTOR – A FORMULATION

To define HRAP-STE, several variables should be given, see Tab.I. There are given a *job offers* set, where each job offer o is defined by a number of employees to recruit o^{cap} and already recruited o^r . All available positions in the i th job offer are $AvPos(o_i) = o_i^{cap} - o_i^r$. Where a *slot* defines a single position available in a job offer. It is assumed that each job offer may have many slots, but a single slot is a part of one job offer. For each assigned job offer $Offer(s)$ there are two measures included in the model: a *Profit*(o) and *Priority*(o) connected to the job offer o (and the s slot). The *Profit* is gained by recruiting a single employee for the job offer o , and the *Priority* of this job offer is defined as $Profit(Offer(s))$ and $Priority(Offer(s))$, respectively. However, none of the clients (the job offerer) can be ignored in the assignments, so the jobs should be offered *proportionally* to keep *Balance* in the company portfolio.

The main goal of HRAP-STE is to satisfy $O = 1, \dots, o_{max}$ job offers that have $n = \sum_{i=1}^{o_{max}} (o_i^{cap} - o_i^r)$ available slots with all available candidates $C = 1, \dots, c_{max}$. However, the practice of short-term employment companies shows that frequently $n > c_{max}$ or $n \gg c_{max}$ – it means that it is impossible to assign employees to *all* available slots. Furthermore, in general c_{max} is not known in advance. Thus, to cover that we define **three quality measures** as *Profit*, *Priority* and *Balance* – it allows the foreplanning that is later used while making the decisions.

Let $\pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ be a slot-processing sequence where π_1 and π_n indicate the first and the last slot to be processed. For instance, if there are three employees available, then the expected profit of their recruiting will be $Profit(Offer(\pi_1)) + Profit(Offer(\pi_2)) + Profit(Offer(\pi_3))$. Since we do not know the number of available employees in advance, then we define the measure

TABLE I
HRAP-STE – VARIABLES AND NOTATIONS (BASED ON [8])

notation	variables
o	The job offer
o^{cap}	The number of employers to recruit
o^r	The number of employers already recruited
s	The single available position, as <i>slot</i>
$Offer(s)$	The job offer assignment
$Profit(Offer(s))$	The profit of recruited slot s in o job offer
$Priority(Offer(s))$	The priority of o job offer, so called: project
c_{max}	The number of available candidates, <i>not known</i> in advance
n	The number of available job positions, where $n >> c_{max}$
$\pi = \{\pi_1, \pi_2, \dots, \pi_n\}$	The slots-processing sequence, where π_1 is the first and π_n the last slot
$max PlanProfit(\pi)$	The summarized profit of all job assignments
$max PlanPriority(\pi)$	The summarized priority of all job assignments (and projects)
$max PlanBalance(\pi)$	The balance measure that covers job offers (projects) distribution

of expected profit as:

$$PlanProfit(\pi) = \sum_{i=1}^n g(i, \pi) \quad (1)$$

$$g(i, \pi) = \sum_{j=1}^i Profit(Offer(\pi_j))$$

We define the *Priority* referring to job offer priorities as:

$$PlanPriority(\pi) = \sum_{i=1}^n p(i, \pi) \quad (2)$$

$$p(i, \pi) = \sum_{j=1}^i Priority(Offer(\pi_j))$$

In a single-objective optimization, it is easy to find the optimal solution only for these two measures simply by ordering the most profitable or prioritized slots first. However, if these two measures contradict, e.g., the slots with low profit have high priority, then the weighted single-objective problem made from these two measures becomes hard to solve.

However, companies that coordinate the short-term employment process cannot limit their activity only to maximize the expected profit and the priority of the declared job offer. They are also expected to ensure that all job offers from the company portfolio will be assigned, in small part. So the *Balance* measure must be considered too, defined as follows:

$$PlanBalance(\pi) = \sum_{i=1}^n b(i, \pi) \quad (3)$$

$$b(i, \pi) = \min_{o \in O} (o^r(i, \pi) / o^{cap}(i, \pi))$$

where $o^r(i, \pi)$ and $o^{cap}(i, \pi)$ refer to the number of recruited employees and the overall number of employees that are to be recruited for the o^{th} offer when the first i slots in sequence π are assigned employees.

The goal of the HRAP-STE problem is to maximize the values of all three measures. The main difficulty in optimizing the above problem is that the considered objectives contradict each other, which is a typical feature of multi-objective optimization [8]. Moreover, the offers with the highest priority do not necessarily bring the highest profit. Finally, the *PlanBalance*

objective may be considered as contradicting both other objectives. To optimize *PlanBalance*, we shall always choose the slots that refer to the offer with the lowest percentage of occupied slots. Thus, optimization of *PlanBalance* will lead to equalization of the percentage of slots occupied for all i in Eq.3. In the paper we investigate the problem, solution landscape and instances, thus the simplified version of HRAP-STE is considered – as defined in Eq.4 by weighted sum of the objectives.

$$f(\pi) = w_1 * PlanProfit(\pi) + w_2 * PlanPriority(\pi) + w_3 * PlanBalance(\pi) \quad (4)$$

In Eq.4 the three weight values $w_1, w_2, w_3 \in < 0.0, 1.0 >$ that define the 'importance' of the selected objectivity. In this paper, all weights are equal to 1.0 for investigations. Moreover, to avoid the domination of some objectivity, all values for the objectives are normalized.

A. Solution encoding example

HRAP-STE solutions are encoded using permutation (π). Each value of the permutation refers to a given slot. Let us consider a HRAP-STE instance with two jobs o_1 and o_2 , each with two slots. The slot profit is $Profit(o_1) = 10$ and $Profit(o_2) = 5$, while priority is $Priority(o_1) = 2$ and $Priority(o_2) = 1$.

The first two values in the permutation refer to o_1 and the latter two to o_2 . We consider solution $\pi^a = [4, 1, 2, 3]$.

$$[Offer(\pi_1), Offer(\pi_2), Offer(\pi_3), Offer(\pi_4)] =$$

$$[Offer(4), Offer(1), Offer(2), Offer(3)] = [o_2, o_1, o_1, o_2].$$

Thus, the quality measures' values will be:

$$PlanProfit(\pi) = \sum_{i=1}^n g(i, \pi) = \sum_{j=1}^1 Profit(Offer(\pi_j)) + \dots + \sum_{j=1}^4 = 5 + (5 + 10) + (5 + 10 + 10) + (5 + 10 + 10 + 5) = 75$$

$$PlanPriority(\pi) = \sum_{i=1}^n p(i, \pi) = \sum_{j=1}^i Priority(Offer(\pi_j)) + \dots + \sum_{j=1}^4 Priority(Offer(\pi_j)) = 1 + (1 + 2) + (1 + 2 + 2) + (1 + 2 + 2 + 1) = 15$$

$$\begin{aligned}
PlanBalance(\pi) &= \sum_{i=1}^n b(i, \pi) = \\
&\min_{o \in O} (o^r(1, \pi) / o^{cap}(1, \pi)) + \dots + \\
&\min_{o \in O} (o^r(4, \pi) / o^{cap}(4, \pi)) = \\
&\min(\frac{1}{2}, 0) + \min(\frac{1}{2}, 0) + \min(\frac{1}{2}, \frac{1}{2}) + \dots + \min(1, 1) = \\
&0 + (0 + \frac{1}{2}) + (0 + \frac{1}{2} + \frac{1}{2}) + (0 + \frac{1}{2} + \frac{1}{2} + 1) = 3\frac{1}{2}
\end{aligned}$$

IV. METHODS

Six well-known **methods** have been investigated in solving an HRAP-STE problem to give a complementary case study: 3 heuristics and 3 metaheuristics that are effective in solving HRAP [5]. The reference method *RandomSearch* is used, and as its improved version *RandomLocalSearch*. In addition, as a reference, the classic algorithm *Greedy* has been used. In experiments, well-known metaheuristics presented in HRAP survey [5], such as classic *GeneticAlgorithm*, *ParticleSwarmOptimisation* (PSO) [3], and *TabuSearch*, have been used to verify their effectiveness in solving HRAP-STE.

The proposed encoding (see sec.III-A) for HRAP-STE is defined as a permutation, so a classic permutation-based **operators** could be applied. We investigated crossover operators for *GeneticAlgorithm* such as *Cycle Crossover* and *Partially Matched Crossover* (PMX). As mutation *Swap*, *Inversion* and *Insert* is investigated. In addition, mutation operators were investigated as the neighborhood generator for *TabuSearch* and heuristics.

V. EXPERIMENTS

The main goal of developed experiments is to investigate the effectiveness of well-known (meta)heuristics, applied to different HRAP-STE scenarios (instances, see sec.V-A). For each investigated method, the best-found configuration and experimental procedure are set (sec.V-B). Finally, the results of the conducted experiments are presented in sec.V-C.

A. Instances

In experiments, a set of benchmark HRAP-STE¹ real and artificial instances are used. Artificial instances are split into 9 *easy* and 3 *hard* one's scenarios.

All HRAP-STE instances presented see Tab.V-A have varying number of job offers, slots, and profits/priorities to define the problem instances. Furthermore, 4 real instances were prepared to show the influence of constraints on real scenario difficulty. Such instances include anonymized EWL GROUP company data from about a month, containing 39-99 job offers and 2-67 slots each offer.

The 10 collected features of HRAP-STE instances, i.e. number of slots, jobs, and priorities with their statistics (min,avg,max) allow to make a PCA analysis of HRAP-STE instances landscape and visualization. The graph in Fig.1 shows that *easy* instances are near, except *easy8*. The *hard* instances are also in near localization. The most surprising is the long distance for *real* instances, which could be interpreted as they model different situations in the company. Additionally, a long distance from *easy* to *real* instances showed that

¹All used HRAP-STE instances and gained results are published in .

TABLE II
HRAP-STE INSTANCES

name	jobs	slots	priorities	profits
easy1	2	3	1-3	30-60
easy2	2	2-3	1-3	45-60
easy3	3	2-5	1-3	20-105
easy4	3	2-3	2-3	10-25
easy5	3	2-4	2-4	10-25
easy6	6	1-3	1-3	60-400
easy7	3	2-7	1-4	15-115
easy8	3	3	1-8	25-1595
easy9	5	2	1-2	20
hard1	9	15	1-5	20-1810
hard2	14	10-24	1-4	10-115
hard3	60	2-9	1-4	10-100
real1	43	2-25	0-3	1.4-2.08
real2	39	2-14	0-3	1.4-2.08
real3	43	2-14	0-3	1.58-3.61
real4	99	2-67	0-3	1.28-3.61

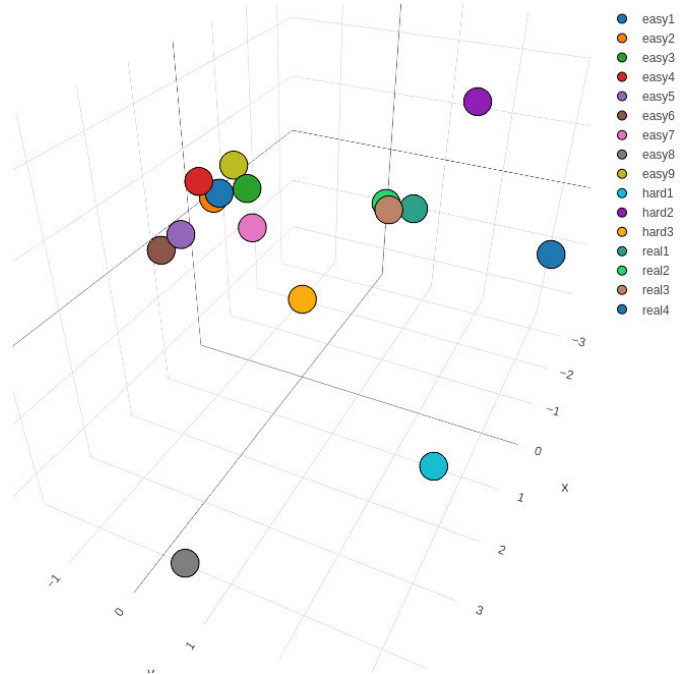


Fig. 1. PCA analysis of HRAP-STE instances

easy are only testing instances that are literally far away from *real* cases.

B. Experimental setup and procedure

We consider a relatively large **computation budgets** (500 000 fitness function evaluations, FFE) for all instances and methods to eliminate the situation in which the best method is simply the one that is the fastest to converge. For each investigated method, the tuning procedure have been run to find best-found configuration – presented in Tab. (see Tab.V-B).

No parameters are given for heuristics like the *Greedy* algorithm, *RandomSearch*, and *RandomLocalSearch*, as the neighborhood operator is defined as *insert*. For metaheuristics, specific parameters should be set. For *GeneticAlgorithm*

TABLE III
THE BEST-FOUND CONFIGURATIONS

method	configuration
RandSearch	-
RandLocalSearch	operator=swap
Greedy	operator=swap
PSO	$c_1 = 0.5$ $c_2 = 0.05$ $psize = 100$ $w = 0.9$
GeneticAlgorithm	$P_{pmx}^x = 0.0$ $P^m = 0.001$ $psize = 200$ $T_{size} = 3$
TabuSearch	operator=insert $tabu_{tenure} = 10$

crossover P^x and mutation P^m probability, size of tournament selection T_{size} and size of population $psize$. For TabuSearch neighborhood operator $insert$ and $tabu_{tenure}$. Finally, for the PSO size of the swarm ($psize$) and specific parameters: inertia weight w , cognitive c_1 , and social c_2 acceleration coefficient.

The experimental results have been evaluated on all HRAP-STE instances. Due to the non-deterministic nature of metaheuristics, all runs have been repeated 30 times, and results averaged. To verify the statistical significance of the presented results, the Wilcoxon signed-rank test is used with $pvalue = 0.05$.

The research environment with all investigated methods has been implemented in Rust and Python. All experiments were developed using the following configuration: 2,6 GHz 6-Core Intel Core i7, 16GB RAM, and OS: Sequoia 15.4.1.

C. Results

All results of experiments use 6 (meta)heuristics in solving 16 HRAP-STE instances – see Tab.V-C. For all *easy* instances, almost all methods gain the same results. The difference can be seen in *hard* and *real* instances, where *GeneticAlgorithm* outperforms other methods. The second place gets *RandLocalSearch*. Although the difference between the two best methods appears to be very small, the Wilcoxon Signed Rank test confirms the statistical significance: for instances *hard* instances $pvalue < 0.0001$ and for *hard* ones do not exceed $pvalue < 0.03$.

The results presented in Tab.V-C encourage a more detailed analysis of results for more difficult instances. Fig.2 presents averaged results for methods solving *hard3* instance – it shows *RandomLocalSearch* in lower budget wins, but finally, the Genetic Algorithm gets the best results. The wider context for *hard3* instance gives a boxplot diagram from Fig.4.

A similar situation occurs in *hard4* instances – see Fig.3, where the Genetic Algorithm relatively quickly, in the computation budget context, outperforms other methods. The boxplot presented in Fig.5 confirms that the Genetic Algorithm for *real4* instance is very competitive.

The budget defined by FFE is useful for comparing results for (meta)heuristics. However, some of them have specific operations unrelated to FFE. In such situations, the computational time for investigated methods could be compared. For *easy* instances, *Geneticalgorithm* needs 13-31 seconds, whereas other methods gain results in less than 1 second. For *hard* instances, *GeneticAlgorithm* works within 26-54

seconds, while *PSO* needs 11-251 seconds and *TS* 24-62 seconds respectively. A similar situation occurs in *real* instances, but there is an exception for *real3* instances, where *GeneticAlgorithm* needs 217 seconds and *TS* 954 seconds. Such differences in computation times for various instances are strictly connected to the 'size' of instances, i.e., how large is the solution landscape.

D. Summary

Experiments presented in previous sections showed that HRAP-SA can be effectively solved by both heuristic and metaheuristic. Heuristic *RandomLocalSearch* is very competitive for a low computational budget, especially for easier instances. However, metaheuristics (like *GeneticAlgorithm*) usage is recommended when efficiency is needed more.

VI. CONCLUSIONS AND FUTURE WORK

In this work, HRAP-STE is defined as a real-world problem that extends the HRAP problem, known as NP-hard. The proposed HRAP-STE definition also consists of representation and fitness function. To evaluate the wider context of HRAP-STE, the 16 benchmark instances that include artificial and real scenarios are proposed. That allows to give a complementary case study, and evaluate the effectiveness of six well-known methods of solving an HRAP-STE problem: 3 heuristics and 3 metaheuristics.

The experiments presented in this paper showed that metaheuristics are effective HRAP-STE solvers. In lower budgets, heuristics are more effective; for larger budgets, metaheuristics outperform other methods. Such dualism encourages linking methods from these groups and defining hybridization – one of the most successful and promising research field in optimization [5]. Additionally, the representation and operators used to solve HRAP-STE in this paper are not specialized. Thus, a further research direction could be pointed out to include *domain knowledge* in new representations and operators. Last but not least, HRAP-STE could be defined as a multi-objective problem, which is considered in [8] – in this context, more extensive research connected to the specialized representations and operators could build a more effective tool for the decision-makers in the company.

ACKNOWLEDGEMENT

This work was supported by The National Center for Research and Development (NCBiR), Poland under Grant POIR.01.01.01-00-1042/20.

Special thanks to EWL S.A. company team, especially for Andrzej Korkus (CEO) and Adam Korkus (CTO).

REFERENCES

- [1] Grillo, H., Alemany, M. & Caldwell-Marin, E. Human Resource Allocation Problem in the Industry 4.0: A Reference framework. *Computers & Industrial Engineering*. **169** pp. 108110 (2022,3)
- [2] Lin, C. & Gen, M. Multi-criteria human resource allocation for solving multistage combinatorial optimization problems using multiobjective hybrid genetic algorithm. *Expert Systems With Applications*. **34**, 2480-2490 (2008)

TABLE IV
RESULTS COMPARISON FOR 30 INDEPENDENT RUNS

	GeneticAlgorithm		Greedy		PSO		RandSearch		RandLocalSearch		TabuSearch	
	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
easy1	0.4497	0.0000	0.4497	0.0000	0.4497	0.0000	0.4497	0.0000	0.4497	0.0000	0.4497	0.0000
easy2	0.5778	0.0000	0.5778	0.0000	0.5778	0.0000	0.5778	0.0000	0.5778	0.0000	0.5778	0.0000
easy3	0.3108	0.0000	0.3108	0.0000	0.3108	0.0000	0.3108	0.0000	0.3108	0.0000	0.3108	0.0000
easy4	0.3939	0.0000	0.3939	0.0000	0.3939	0.0000	0.3939	0.0000	0.3939	0.0000	0.3939	0.0000
easy5	0.4354	0.0000	0.4354	0.0000	0.4354	0.0000	0.4354	0.0000	0.4354	0.0000	0.4354	0.0000
easy6	0.3603	0.0000	0.3555	0.0075	0.3603	0.0000	0.3602	0.0001	0.3549	0.0077	0.3603	0.0000
easy7	0.2554	0.0000	0.2550	0.0009	0.2554	0.0000	0.2554	0.0000	0.2550	0.0009	0.2554	0.0000
easy8	0.4099	0.0000	0.4099	0.0000	0.4099	0.0000	0.4099	0.0000	0.4099	0.0000	0.4099	0.0000
easy9	0.4606	0.0000	0.4606	0.0000	0.4606	0.0000	0.4606	0.0000	0.4606	0.0000	0.4606	0.0000
hard1	0.4120	0.0001	0.4121	0.0001	0.4099	0.0005	0.3983	0.0008	0.4120	0.0001	0.4121	0.0000
hard2	0.1846	0.0004	0.1829	0.0007	0.1765	0.0015	0.1614	0.0011	0.1842	0.0004	0.1707	0.0019
hard3	0.3936	0.0006	0.3837	0.0040	0.3612	0.0029	0.3192	0.0023	0.3918	0.0010	0.3364	0.0056
real1	0.3191	0.0011	0.3183	0.0020	0.3003	0.0026	0.2691	0.0021	0.3174	0.0011	0.2913	0.0059
real2	0.1846	0.0004	0.1829	0.0007	0.1765	0.0015	0.1614	0.0011	0.1842	0.0004	0.1707	0.0019
real3	0.2261	0.0019	0.2223	0.0021	0.2110	0.0023	0.1812	0.0017	0.2227	0.0018	0.2066	0.0041
real4	0.2905	0.0030	0.2176	0.0029	0.2678	0.0033	0.2390	0.0014	0.2921	0.0025	0.2213	0.006

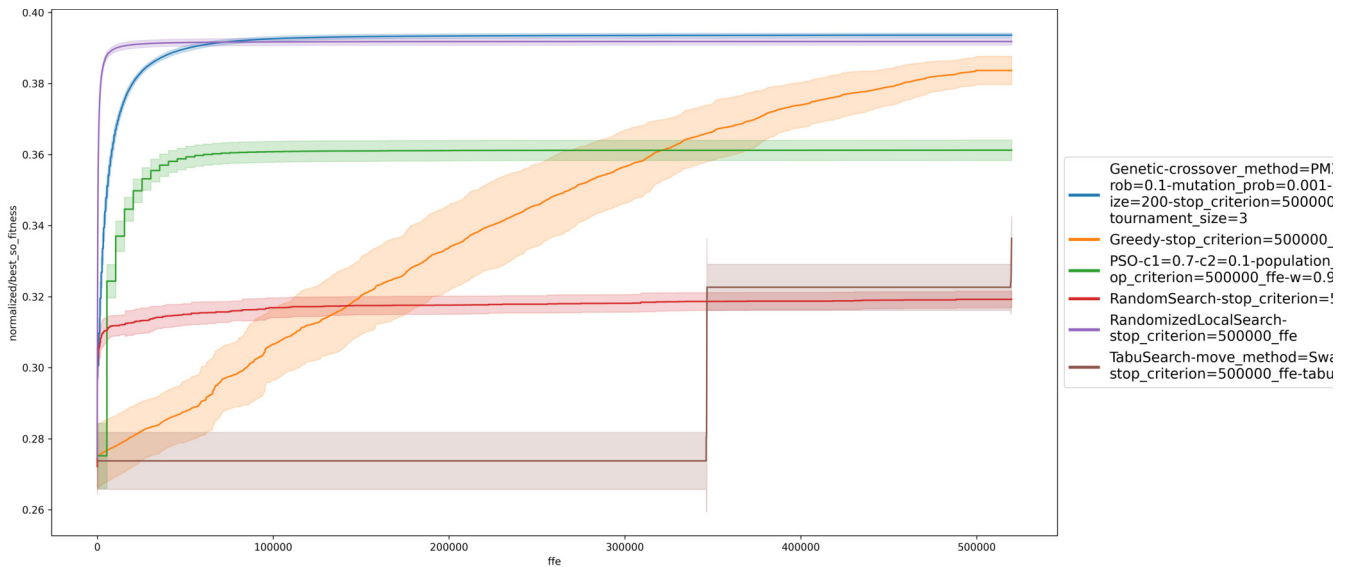
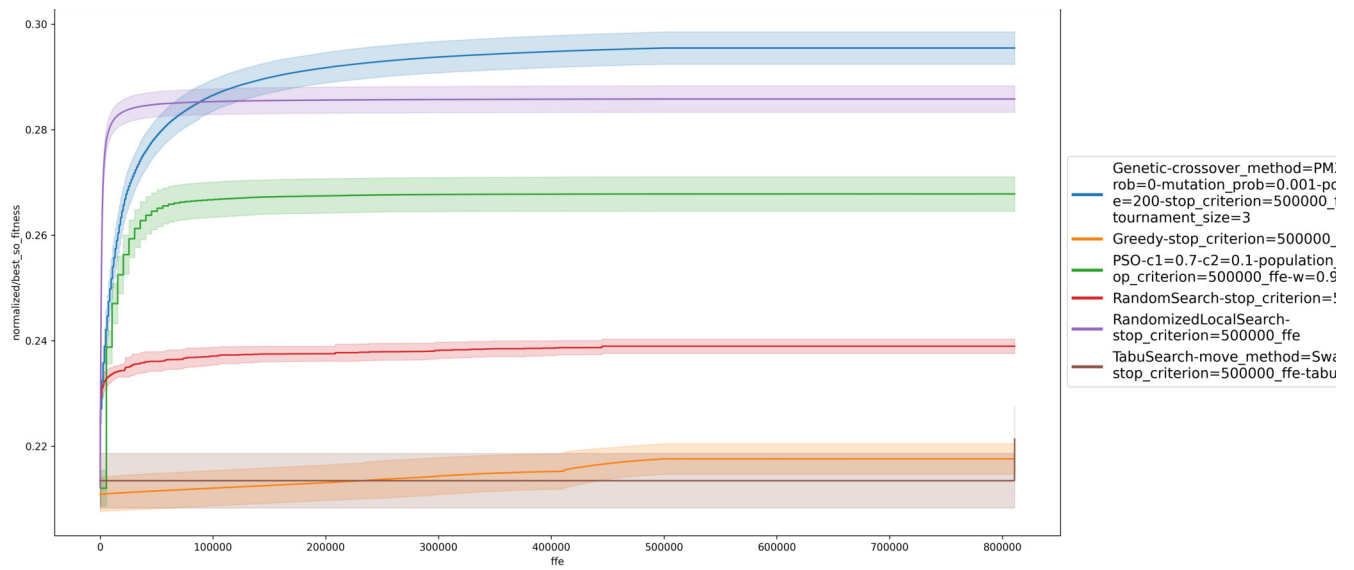
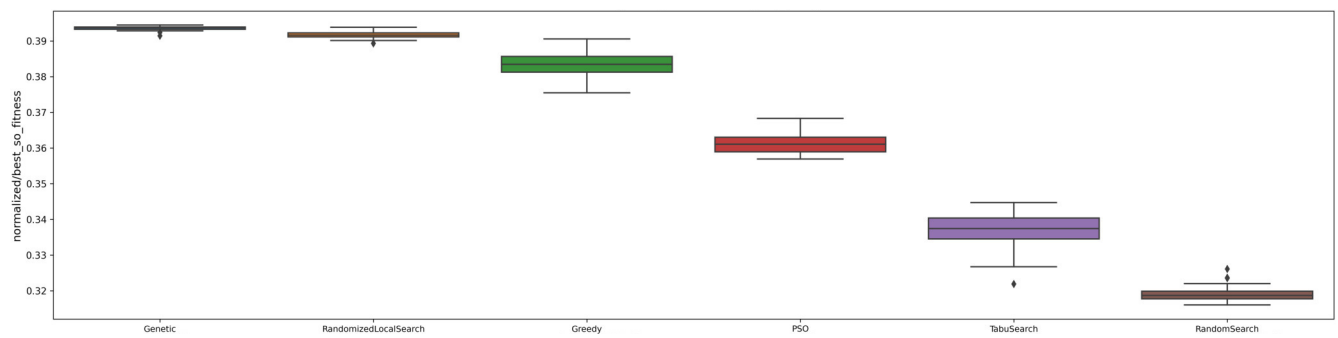
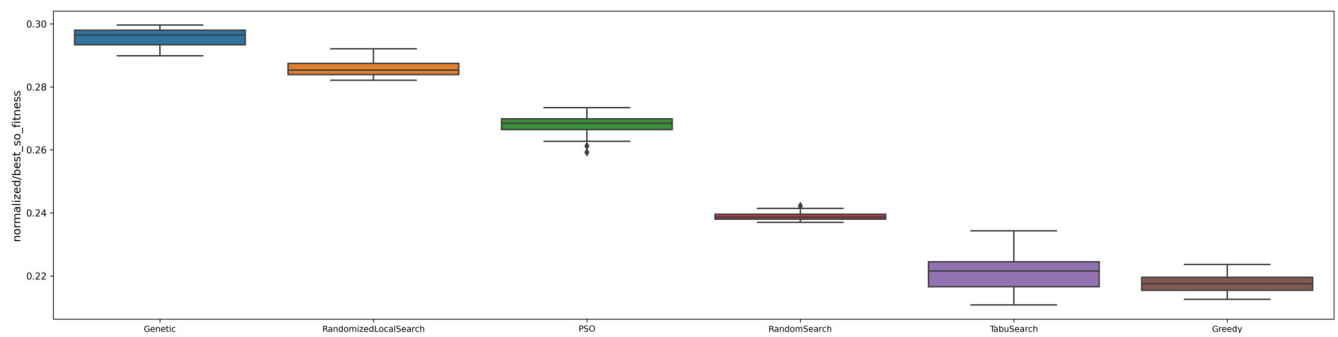


Fig. 2. Computational results for *hard3* instance.

- [3] Wang, Z. Enterprise Human Resource Allocation Optimization Model Based on Improved Particle Swarm Optimization Algorithm. *Wireless Communications And Mobile Computing*. **2022** pp. 1-9 (2022,3)
- [4] Park, J., Seo, D., Hong, G., Shin, D., Hwa, J. & Bae, D. Human Resource Allocation in Software Project with Practical Considerations. *International Journal Of Software Engineering And Knowledge Engineering*. **25** pp. 5-26 (2015,2)
- [5] Bouajaja, S. & Dridi, N. A survey on human resource allocation problem and its applications. *Operational Research*. **17** pp. 339-369 (2017)
- [6] Huynh, N. & Chien, C. A hybrid multi-subpopulation genetic algorithm for textile batch dyeing scheduling and an empirical study. *Computers & Industrial Engineering*. **125** pp. 615-627 (2018), <https://www.sciencedirect.com/science/article/pii/S0360835218300068>
- [7] Taillard, E. Benchmarks for basic scheduling problems. *European Journal Of Operational Research*. **64**, 278-285 (1993), <https://www.sciencedirect.com/science/article/pii/037722179390182M>, Project Management and Scheduling
- [8] Przewozniczek, M.W., Myszkowski, P.B., Kosciukiewicz, W., Wojcik, M., Gonczarek, A., Korkus, A., On discovering and analysing variable dependencies to construct an effective and efficient optimiser dedicated to solving the new real-world multi-objective resource-allocation problem, **in review**, *Applied Soft Computing Journal*.

Fig. 3. Computational results for *real4* instance.Fig. 4. Results comparison for *hard3* instance.Fig. 5. Results comparison for *real4* instance.

Towards a German VET Archive and its Integration into a Data Warehouse

Thomas Reiser^{*}, Petra Steiner[†], Kristine Hein[†]

^{*} University of Koblenz, Germany,
Email: treiser@uni-koblenz.de

[†] Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Email: steiner@bibb.de

Abstract—This paper presents a systematic evaluation and prototypical implementation of an information system for historical vocational education and training (VET) regulations in Germany. The focus of this study is on integrating structured outputs with the German Labor Market Ontology (GLMO) and a broader labor market data warehouse. A corpus of VET and CVET regulations, as published in the Federal Gazette from 1969 to 2022, was used to assess the functional and semantic requirements of the archival process. This analysis was complemented by a review of existing software frameworks, culminating in the proposal of a combined architecture utilizing Omeka S and TEI Publisher. In addition, the necessary transformations, metadata enrichment, and ETL processes required to integrate the resulting TEI XML documents into a semantically linked data environment are detailed. This work provides a concrete roadmap for the sustainable digitization and semantic integration of regulatory texts into modern labor market intelligence infrastructures.

I. INTRODUCTION

VOCATIONAL education and training (VET) systems are of critical importance in maintaining a skilled workforce and supporting economic resilience. In Germany, a historically extensive corpus of VET and continuing VET (CVET) regulations has been published in the Federal Gazette over the course of several decades. These documents serve as the foundational elements of occupational standards and training frameworks, garnering substantial interest from researchers, policymakers, and labor market analysts. However, the archival form of these regulations as described in [1]—primarily as unstructured or semi-structured scanned documents—poses challenges for digital accessibility, analysis, and integration with contemporary data systems.

The digitization of archival material presents an opportunity to preserve, structure, and analyze regulatory knowledge in a form amenable to semantic linking, machine learning, and long-term data curation as discussed in our previous work [2], [3]. In this context, two fundamental questions emerge: first, which software tools and platforms are most suitable for the digitization, structuring, and management of historical training regulations, and second, how can the resulting digital records be semantically integrated into the German Labor Market Ontology (GLMO) and a broader data warehouse environment that supports longitudinal labor market research?

In order to address the aforementioned inquiries, the present document offers a technical design and evaluation of a digitization pipeline founded upon image preprocessing, optical

character recognition (OCR), and TEI XML structuring. A comparative review of available archival systems is conducted, followed by the implementation of a dual-platform prototype using Omeka S for metadata management and TEI Publisher for structured transcript administration. Beyond the archival perspective, a methodology is proposed for mapping the structured training documents to historical occupation taxonomies. This methodology enables their integration into the GLMO and subsequent ingestion into a data warehouse through standard ETL procedures. The present study contributes to the development of interoperable digital infrastructures for vocational education and labor market data by addressing both archival and semantic integration concerns.

This study is organized as follows: The first section introduced our data set, consisting of records and scans of the occupations archive. Next, the research background is introduced to give an overview over research in the area of document digitization and information extraction. Then, we discuss related literature that addresses the existing solutions to similar problems, especially regarding the text structure recognition and the integration of such workflows into web applications. The fourth section introduces our methodology that aligns with the early phases of the software lifecycle, from analyzing the problem statement, requirement elicitation, and system design in order to implement a first prototype, that is presented afterwards. Finally, findings of strengths and shortcomings with this prototype are discussed, before summing up the paper and giving an outlook over future work.

II. BACKGROUND

The digitization of historical documents has garnered significant interest in recent years, with a proliferation of methodologies to address this undertaking. Optical character recognition (OCR) is a foundational technology for digitization, with a significant research focus and a range of established tools.

A plethora of methodologies exists for the purpose of document digitization. One fundamental approach entails the detection of the complete text within the document images, as illustrated by the methodology employed in the Finnish newspaper digitization project[4]. In this project, the objective is to generate an ALTO XML document that encompasses all recognized text, leveraging the Tesseract OCR engine. An alternative approach is demonstrated in [5], wherein the

authors model the text structure of legal texts in Austria and align the recognized text to this predefined structure, thereby enhancing the structured recognition of text. More advanced methods employ the OCR results to construct structured data from the text images. The authors of [6] employ OCR to digitize invoices and to structure the recognized information, such as product description, quantity, and price. In a similar vein, the study by [7] involved the extraction of names of judges at German federal courts from 1950 to 2019. This was achieved by applying OCR to the Federal Gazette, a publication that contains the official gazette of the Federal Republic of Germany.

In order to optimize accessibility, a number of OCR workflows have been integrated into web-based applications [8]. A significant endeavor in this domain is OCR4all, which facilitates the implementation of diverse preprocessing steps, segmentation methodologies, and OCR models. Additionally, it facilitates interaction with each of the process steps, enabling users to make corrections at intermediate results and thereby improve the overall outcome. However, there are also less extensive tools that facilitate the management of digitized document collections. For instance, these tools can be found in [9], [10].

III. RELATED LITERATURE

In previous works, we analyzed the data set of the occupations archive [1] to obtain an overview over the different types of documents which vary a lot in language- and layout-style. For a selected data set of documents from the Federal Republic Germany that have been available on the internet, a structure analysis has been conducted on TEI XML transcripts that were created by a rule-based transcription pipeline [3], [2]. Both text structure and content were analyzed to get a first overview over commonalities in the selected documents. As all of the documents were from the same period, their layout followed mostly similar patterns, even across multiple decades. However, an evolution of wording and structure over time was observed. While the development of a more advanced approach for the occupations archive is underway, this pipeline served as a preliminary feasibility analysis.

In the initial approaches to document structuring that emerged in the 2000s, human knowledge about the document was employed to delineate text- and layout-based rules for extracting the text structure [11], [12], [13]. The initial iteration of the digitization workflow employed predefined rules; however, future research endeavors will prioritize the automatic recognition of patterns in layout and text features to facilitate the structured organization of texts with diverse layouts. As with [5], a specific structure is delineated to replicate the text's hierarchical arrangement.

Despite the existence of standard conversion tools, such as Vertopal, which facilitate the transformation of text files into markup languages, these tools operate under the assumption that the text contained within documents to be converted is accompanied by accurate structural information [14] like in born-digital documents. However, this cannot be assured

through the utilization of default OCR (Optical Character Recognition) models which are important to extract text from image information. While these tools can be utilized to generate files in HTML or TEI XML, the resulting output files frequently fail to accurately represent the text hierarchy or logical document structure. Instead, these files offer an alternative representation of recognized text areas, lines, and text.

The utilization of machine learning models, akin to those employed in GROBID [15], facilitates the extraction of meta-data elements such as title and author information. Additionally, it facilitates the recognition of references and citations, as well as the detection of the abstract. Despite its extensive array of useful features, the model was trained exclusively on scientific articles, resulting in its exceptional performance on this specific domain. In order to employ GROBID in legal documents such as the training regulations examined in this article, it would be necessary to refine the model, a process that would require training data. The efficacy of this approach is contingent upon the quality of page segmentation and the reliability of text recognition.

The target data format has been determined to be TEI XML, as it is endorsed by the German Research Foundation (DFG) as a suitable standard for the long-term archiving of documents, see [16]. While PDF is a proprietary standard that is stored in binary files, markup languages such as XML can be read by almost any computer without the need for additional software designed for reading PDF files. Moreover, these files can be efficiently stored in XML databases such as eXist-db to manage the document collection [17]. eXist-db is a versatile system that facilitates the incorporation of plugins, including a versioning plugin. This plugin enables memory-efficient storage of multiple versions of the same documents, facilitating swift restoration of older versions when necessary. This is particularly salient in the context of automated systems, where errors are to be anticipated. Another beneficial plugin is TEI Publisher, which facilitates the management of XML databases and enables the viewing of documents in a human-readable manner. It also allows for the editing of uploaded TEI XML files. Additionally, it enables the visualization of the text alongside the original images, thereby facilitating the error correction process. Consequently, TEI XML has been identified as an optimal target data format for the digitization of extensive text corpora.

IV. METHODOLOGY

A. Software Requirements

The software requirements for the digitization of the vocational training archive encompass a broad set of functional and non-functional requirements aimed at ensuring the accurate and efficient transformation of historical training regulations into structured, searchable digital formats. The archival system must possess the capability to ingest scanned documents in a variety of formats, in particular PDF, PNG, JPG, and TIFF, while also employing an AI pipeline to automatically generate TEI XML transcripts. The system must support the

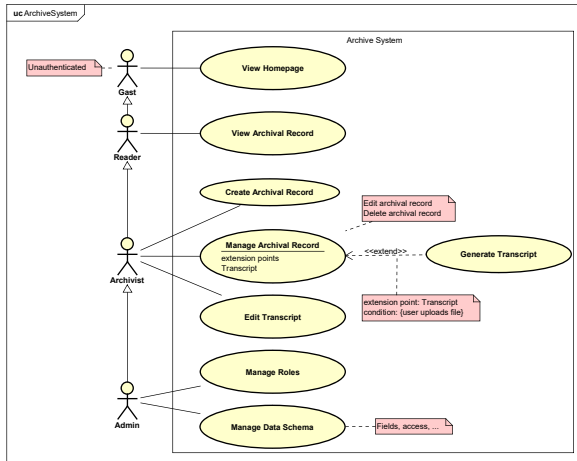


Fig. 1: Use Case Diagram for the Archive System

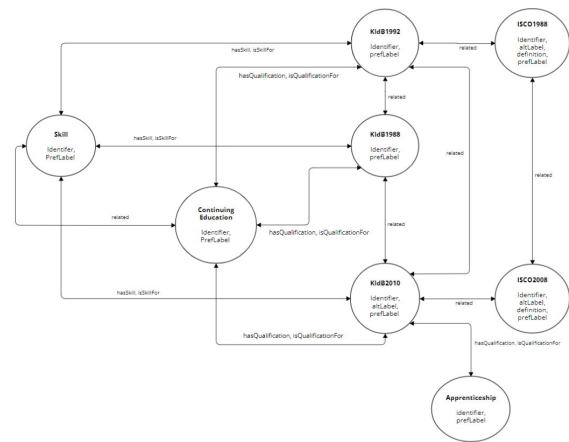


Fig. 3: Data Schema for historical KldB Data

Name:	Archiveintrag erstellen
Beschreibung:	Ein Archivar kann Dokumente hochladen und mit den zugehörigen Metadaten annotieren. Das System speichert diese Informationen und erstellt automatisch ein Transkript im TEI XML Format.
Vorbedingung:	Auf dem Fileserver ist genug Speicherplatz für den hochzuladenden Scan verfügbar.
Nachbedingung:	Ein neuer Eintrag wurde im Archivsystem erstellt.
Fehlerfälle:	Der Transkriptionsworkflow kann den Upload nicht verarbeiten. Es ist nicht genügend Speicherplatz für den neuen Eintrag verfügbar. Der Archivar erhält eine Meldung, dass der Upload nicht erfolgreich war.
Systemzustand im Fehlerfall:	Archivar
Aktör:	Archivar
Auslöser:	Der Archivar hat ein Dokument gescannt und möchte es zum Archivsystem hinzufügen.
Standardprozess:	Neues Dokument speichern und Scan hinzufügen.
ss:	- Der Archivar ruft die Seite zum Upload auf. - Der Archivar lädt den Scan hoch. - Das System erstellt ein Transkript für das hochgeladene Dokument. - Das System füllt die Felder für im Dokument enthaltene Metadaten mit automatisch generierten Vorschlägen aus. - Der Archivar korrigiert mögliche Fehler im Transkript oder den gefundenen Metadaten. - Der Archivar bestätigt seine Eingaben und der neue Eintrag wird im Archivsystem gespeichert. - Das System prüft, ob ein ähnlicher Archiveintrag bereits existiert und gibt dem Archivar die Möglichkeit, den Scan als neues Exemplar des Dokuments hinzuzufügen. - Das System speichert den neuen Eintrag und dass dieser überprüft wurde und korrekt sein sollte.
Alternativer Prozess:	Neues Dokument speichern und nicht sofort kontrollieren, ob alle Daten korrekt vom Workflow erkannt wurden.
	- Der Archivar ruft die Seite zum Upload auf. - Der Archivar lädt den Scan hoch. - Das System erstellt ein Transkript für das hochgeladene Dokument. - Das System füllt die Felder für im Dokument enthaltene Metadaten mit automatisch generierten Vorschlägen aus. - Der Archivar speichert den neuen Eintrag im Archivsystem. - Das System speichert den neuen Eintrag, sowie dass dieser noch nicht überprüft wurde und eventuell noch Fehler vorliegen.

Fig. 2: Scenario: Create Archival Record

management of metadata and version control, enable role-based access (i.e., guest, reader, archivist, administrator), and offer functions such as document upload, editing, and deletion. Furthermore, advanced features such as duplicate detection, exemplar linking, and prediction of metadata through machine learning are integral. The archival system under consideration should facilitate PDF export, full-text search, and integration with external tools such as TEI Publisher for transcript handling. Based on use case diagrams like shown in Figure 1, scenarios like in Figure 2 have been created to derive requirements for the information system.

B. Possible Software

A variety of open source software solutions were evaluated with the objective of meeting the requirements of the Archive project. A comprehensive overview of these solutions can be found in Table I. Because there are many already existing tools that should be able to solve the problem of designing an archival information system, from a maintenance perspective, it makes a lot sense to reuse these technologies instead of creating another one. The foundational platform for the archival

system prototype is Omeka S, a digital archive framework that is modular and extensible. It offers role-based access control, ontology-based metadata management (e.g., Dublin Core), and integration capabilities via REST APIs. The architecture of this system is designed to support the organization of digital items and metadata-rich content, rendering it well-suited for general archival needs. However, a critical deficiency in Omeka S is its lack of native support for TEI XML transcriptions, a fundamental requirement for accurately representing the structure and hierarchy of historical training regulations. To address this, TEI Publisher was selected as a complementary tool, offering specialized support for managing, versioning, and displaying TEI-encoded documents. Constructed on the eXist-db platform, TEI Publisher facilitates seamless integration with Omeka S and offers a customized user interface for transcription workflows.

In addition to the primary tools, alternative systems such as Paperless-ngx and Access to Memory (AtoM) were also considered. Paperless-ngx is an open-source document management system designed for individual or small-scale organizational use, offering basic OCR and tagging features. While it demonstrates notable strengths in terms of simplicity and usability, it exhibits deficiencies in terms of flexibility and extensibility, which are crucial for effective performance in complex archival tasks and structured text processing. Conversely, AtoM is designed to align with international archival standards and offers a web-based interface for institutional repositories. Despite its strengths in terms of compliance, the software does not offer sufficient support for AI-driven transcription pipelines or integration with TEI XML workflows. Consequently, the integration of Omeka S and TEI Publisher was determined to be the optimal solution for meeting the technical and semantic criteria of the archival digitization initiative.

C. Integration into GLMO

To facilitate comprehensive longitudinal analyses of vocational development in Germany, the German Labor Market

TABLE I: Overview of possible Software Products for the VET Archive

Software	Type	Key Features and Notes
Omeka S	Web Publishing Platform	• Modular architecture, extensible with custom modules • Secure authentication (e.g., via LDAP) • Supports ontologies (e.g., Dublin Core) • REST API and CSV import • Lacks native TEI XML support
TEI Publisher	XML Management	• Manages TEI XML transcriptions • eXist-db plugin with built-in versioning • UI for transcript administration • Suitable for integration with Omeka S
Paperless-ngx	Document Management System	• Open-source DMS with OCR and tagging • Focused on personal or small-business document workflows • May be limited for complex archival needs
Access to Memory (AtOM)	Archival Description Tool	• Focus on archival standards (e.g., ISAD(G)) • Web-based interface for archival institutions • Less suited for tight integration with AI pipelines

Ontology (GLMO) is being extended with historical occupational taxonomies from both the Federal Republic of Germany (FRG) and the former German Democratic Republic (GDR). This ontological enrichment involves the alignment of legacy classification systems, such as KldB 1988 and KldB 1992, with more recent taxonomies, including KldB 2010 and ISCO-08, through a series of conversion mappings, see Figure 3. These mappings facilitate temporal integration and semantic linking of occupational entities across decades. The establishment of bidirectional links between historical training regulations—often preserved as scanned and TEI-encoded documents—and the corresponding occupation nodes within the GLMO knowledge graph is of particular relevance to the occupations archive project. The utilization of persistent identifiers and relation types, such as "hasSource" or "referencesClassification," facilitates the embedding of archival artifacts directly into the ontology-driven representation of the labor market. This integration facilitates enhanced contextualization of archival data and supports link prediction and graph reasoning tasks, enabling researchers to infer structural trends, skill transitions, and educational pathways over time.

Currently, labels from the genealogy of vocational education that describes the history of vocational training in Germany are used for the mapping to the GLMO. Because the regulations in the occupations archive build the legal foundation for the records in the genealogy as they describe the time periods where training occupations had state recognition in Germany, we expect that records in the occupations archive should have a one-to-one mapping with the genealogy records. To match these two data sets, we use record linkage methods as described in [18].

D. Integration into Data Warehouse

The integration of the occupations archive into a more extensive data warehouse for vocational education and labor market research introduces a series of technical and semantic

challenges. The initial task entails the harmonization of TEI XML document structures with the classification systems that are already present in the data warehouse. These include the German Classification of Occupations (KldB) and the German Labor Market Ontology (GLMO). This mapping is designed to ensure semantic interoperability and facilitate meaningful linkage to other datasets based on occupation, time, or region.

A secondary requirement is the design of reliable ETL (Extract–Transform–Load) processes to convert document scans and their corresponding XML outputs into structured records suitable for ingestion. These transformations are required to normalize formats, extract metadata, and validate consistency across records. Concurrently, data protection considerations must be addressed. Given the potential for scanned documents to contain sensitive or personally identifiable information, compliance with the General Data Protection Regulation (GDPR) is imperative. This involves the development of anonymization procedures, incorporating pseudonymization or masking techniques, with the objective of preserving analytical value while ensuring the protection of personal data.

Moreover, the historical nature of the documents necessitates the implementation of temporal modeling. Regulations must be indexed not only by document metadata but also by their effective periods, including enactment and expiration dates. This temporal axis facilitates longitudinal analysis and compatibility with existing time-based labor statistics. Another essential integration step is the enrichment of documents with metadata that might not be present in the original scans. The employment of natural language processing methodologies is imperative for the inference of document types, occupational domains, and geographic scope from the content.

Subsequent integration steps involve validation against reference datasets. Cross-referencing transcribed documents with existing administrative data, such as DAZUBI or QuBe, ensures data quality and enables automatic categorization. Prior to implementation in business intelligence (BI) environments

or dashboards, the content must undergo plausibility checks, formatting consistency tests, and, if necessary, additional aggregation or filtering. These processes must be meticulously documented, with version control and user access management implemented to ensure both reproducibility and secure data access for researchers and stakeholders.

E. DFG requirements

A plethora of methodologies pertain to the process of document digitisation. As outlined in the DFG Practical Guidelines on Digitisation [19], a comprehensive compendium on the execution of such a digitisation initiative is available. Digital preservation is an undertaking that requires meticulous planning from a multitude of perspectives. The initial step in this process typically involves the acquisition of an overview of the existing resources, encompassing personnel, financial, and material aspects. Furthermore, it encompasses the consideration of potential objections pertaining to damages that may arise during the digitisation process. Subsequent to this, a suitable scanning technology and method must be selected. The choice of format and quality of paper can have a significant impact on the final result. In the case of document images, the target format should be TIFF uncompressed, as this is a lossless format that has been in existence for a considerable time and is widely accepted by the majority of archival endeavours. These original images are also referred to as 'digital masters', and it is imperative that they undergo minimal post-processing to ensure that the integrity of the original information is preserved.

The generation of various derivatives is possible from this digital master. A derivative could for example be a copy of the image in JPG or PNG format, in black and white, or with reduced noise. With regard to more efficient web delivery, a PDF file can be utilised, in which the different scans are combined into a single document, potentially with embedded text.

To guarantee interoperability with other archives, one or more standards for the metadata needs to be selected. Commonly used standards include the Dublin Core Metadata Initiative (DCMI) terms, and the Europeana Data Model (EDM) which are represented as linked data. Other established formats include the METS/MODS XML format, ISAD(G), and others that are not mentioned here. It is common practice to store metadata in multiple formats so that the data becomes better interoperable with other archives that might use different formats and also for better interoperability with the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The DFG guidelines recommend METS/MODS in particular, but states that formats like DCMI are also a suitable alternative.

The key challenges associated with the scanning process include the potential for damage to documents during scanning, water damage, brittle paper, faded ink, and various other external factors. It is imperative to ensure that the original resources are not destroyed during the process of digitisation.

V. PRESENTATION OF PROTOTYPE

As outlined in Section IV-B, the majority of the tools mentioned therein are considered to be valid choices for the archive system. Indeed, combinations of different tools are easily possible. Archivematica, for instance, is equipped with a functionality that integrates it into AtoM with minimal effort. The microservice architectures of Archivematica and Islandora facilitate the extension of these tools and the distribution of their components across multiple machines. However, this architectural design invariably entails a compromise in terms of maintainability, see [20]. Whilst less OAIS-compliant tools, such as Omeka S and Islandora, can offer greater flexibility, particularly with regard to data schemas, more static solutions, including AtoM and Archivematica, are better at ensuring that stored metadata adheres to the most recent standards, such as METS or MODS. While it is possible to define such a schema in the other tools as well, the process is more prone to human error during system setup. In the context of storing disparate data, such as data scraped from social media, which can also be pertinent to labour market research, it would be advantageous to store their content within the record's metadata for enhanced findability in the database, rather than having to scrape media attached to a record in the system.

The initial prototype focuses on the representation of the already given metadata and integration of the transcription workflow. First, a suitable data format to store the records in the system has to be selected. As Omeka S is based on linked data concepts, the metadata terms from the Dublin Core Metadata Initiative (DCMI) have been selected as it supports all of the required terms. Although it is common practice to store archival information in multiple formats, the DCMI terms are a widely accepted standard for the description of digital archival records and considered to be sufficient enough for the first prototypical implementation. Using a linked data based format also increases the usability with other non-archival data.

Different RDF schemas are easy to import with Omeka S through the user interface shown in Figure 4 and the table that stored the different existing records can be imported with a mapping from columns in the CSV file to available fields in the Omeka S instance through the CSV Import module. The resulting overview over all records in the admin interface looks as depicted in Figure 5.

As the planned information system is to be initiated with the documents of the occupations archive, it is recommended that the data be migrated into Archivematica in order to structure it in METS, PREMIS or DCMI which are all considered to be best practice for archival metadata, see [19]. Subsequently, the data can be imported into one of the other tools to facilitate browsing the existing data in collections.

Given that the target format is TEI XML, the use of an XML database such as eXist-db can be highly advantageous for advanced querying within the documents, for example by employing the xquery query language. Additionally, eXist-db can be extended by the TEI Publisher module, which

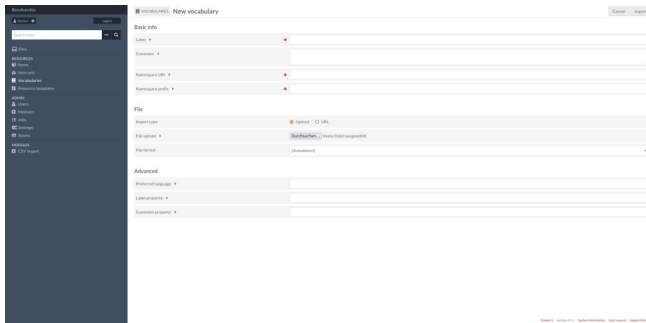


Fig. 4: Ontologies import interface

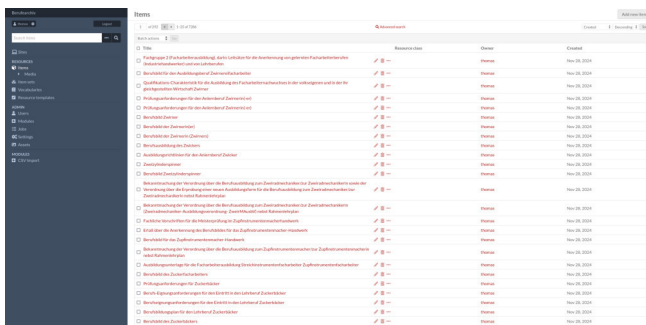


Fig. 5: Items overview in the admin interface

facilitates the straightforward viewing and editing of transcripts. The software also incorporates a functionality for the tagging of named entities within documents, thus rendering it conducive to the annotation of documents. In the context of the occupations archive, the term 'occupation' is understood to encompass not only names, but also skills, tools and other labour market-related terminology. A screenshot of how the TEI XML document that was generated by the prototypical rule-based transcription pipeline is shown in Figure 6

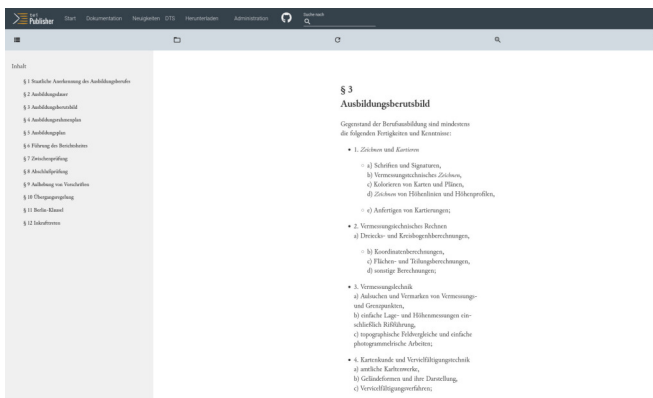


Fig. 6: TEI XML file rendered by TEI Publisher. The interface also allows the annotation of the document and automatically generates an outline.

VI. RESULTS

The implementation of the prototype system, based on Omeka S in combination with a custom AI-powered transcription pipeline, has demonstrated the feasibility and effectiveness of the proposed digital archiving approach. It was demonstrated that both central use cases, namely 'Create Archival Record' and 'OCR-Based Transcription Pipeline', were successfully supported, albeit with complementary tools for full functionality.

Omeka S facilitated the structured creation and management of archival records. Archivists were able to upload scans in multiple formats and annotate them using ontology-based metadata fields, following the Dublin Core Metadata Initiative (DCMI) standard. The system's modular architecture permitted the integration of additional features, such as duplicate entry detection and role-based access control. It is possible for users to differentiate between verified and unverified entries by employing customised resource templates. It was demonstrated that the configuration under consideration was compatible with both standard and alternative workflows.

The AI-based pipeline was implemented externally and linked to the archive system. The system was able to successfully process scanned documents through a series of preprocessing, layout analysis, OCR, and transcript generation steps. The resulting TEI XML files were stored in an XML-native database and made accessible via references within Omeka S. Despite its inability to render TEI XML natively, Omeka S functioned as a stable metadata and document management layer. Integration with TEI Publisher facilitated the visualisation, versioning, and semantic navigation of the structured transcripts.

The efficacy of the modular integration strategy was demonstrated by the outcomes of the study. In this context, Omeka S functioned reliably as the front-end and metadata management layer. Concurrently, specialised tools were utilised for the purpose of handling complex document processing and TEI rendering. This architecture facilitated scalable, standards-compliant digitisation of vocational training regulations spanning over a century of historical data.

While this prototype provided key features of the required functionalities, there are still some open issues and missing features. Omeka S for example only uses a set of predefined roles which does not exactly match the requirements. Especially regarding user privacy, there are some issues. Every authenticated user is able to see all other users in the system. This is not desirable in our context as some users might not want to share their email address and in the case where an intruder manages to gain access to the system, he will also be able to figure out which users might give him the highest privilege. At the moment, there is no built-in way to manage these access rights to restrict this access.

Additionally, Omeka S is particularly well suited for digital exhibitions. However, for now, our focus lies more on data management than presentation. While Omeka S is a very flexible system, stricter rules regarding data types for some

fields can increase data quality as certain fields like the release date should only contain dates, and not e.g. textual information, to have better harmonized data. Such restrictions are not straightforward to implement in Omeka S, although some lighter rules are possible, e.g., that a certain fields must contain media or a URI.

Islandora, for example, allows the creation of custom data schemas and an additional mapping to RDF elements, which also allows to create a linked data representation of the records, thus also an additional representation in form of DCMI terms. It also allows the creation of custom roles with a predefined selection of rights which are sufficient to match the actors shown in Figure 1.

It remains to test the combination of Islandora instead of Omeka S with the other tools to see if all requirements can be fulfilled. Although Islandora comes with more customization options, its microservice architecture also reduces the system maintainability.

The DFG recommends storing data according to the Open Archive Information System (OAIS) model which is currently not implemented in the proposed prototype since Omeka S structurally does not exactly follow this reference model.

Furthermore, there are even more extension that would improve the system's archival capabilities. Archivematica is a tool that is designed to align exactly to the OAIS model and allows an easy integration with AtoM and some other tools. It is in particular designed for the management of the different information packages across the OAIS model and designed to be integrated with other software. However, as we are planning to add more information to the records than is supported by AtoM, Omeka S was selected for the prototype to increase flexibility. A successful integration of Archivematica with Omeka S has been demonstrated in [21]. The authors of [22] on the other hand have been able to create a OAIS compliant system with Archivematica and Islandora.

Given the existence of certain domain-specific metadata fields, and in view of the fact that Islandora provides a greater degree of flexibility while ensuring higher data quality, it has been selected for the purpose of browsing the occupations archive in a future implementation of the system to replace Omeka S.

VII. CONCLUSIONS AND OUTLOOK

This work presents the occupations archive at the Federal Institute for Vocational Education and Training, in addition to the ongoing endeavour to establish an information system for the digital management of the records in the aforementioned archive. The system will provide an interface for the management of metadata and will also generate structured transcripts in the TEI XML format of the document images. The following tools are presented: The initial design of the system and its components is outlined herein, and the final implementation of a preliminary prototype is currently underway.

Furthermore, there remains a paucity of ground truth data for the training of AI models. It is imperative that the annotations appended to the transcripts are of the utmost precision,

incorporating such elements as layout, tables, the sequence of reading, and text hierarchy, in order to facilitate comprehensive comprehension of the document contents, as delineated in the transcripts. Utilising the established models, a combination of the diverse predictions will be employed to generate TEI XML files.

The subsequent stage is the implementation of the proposed system and its testing in a variety of scenarios in order to evaluate its use in creating an interoperable archive system that can be readily extended. There are several challenges that still require resolution, including the question of how consistency can be ensured between Omeka S or Islandora, Archivematica, and the XML database. Following the implementation of a stable prototype, the system can be advanced in several ways. For instance, the transcription pipeline can be expanded to encompass additional information extraction tasks, such as Named Entity Recognition, or alternatively, the implementation of additional pipelines can be contemplated.

Furthermore, the capacity to incorporate additional data into the system is a potential benefit. In the study [23], researchers at the BIBB analysed social media data to ascertain information regarding vocational education and training. This data can assist labour market researchers in a number of ways. While these data are not classical archival data, providing them in the same web service helps create not just an archive, but a diverse information system, while still adhering to archival best practices with the occupations archive.

In addition to the system itself, the TEI XML transcripts can be useful in a variety of ways. The use of Large Language Models in the training of such transcripts is facilitated by their inherent structural nature. The comparison of the regulations themselves is rendered more straightforward, as illustrated in [24]. Furthermore, these models enable more efficient reasoning and referencing to specific paragraphs in the text.

REFERENCES

- [1] T. Reiser, J. Dörpinghaus, P. Steiner, and M. Tiemann, "Towards a dataset of digitalized historical german vet and cvet regulations," *Data*, vol. 9, no. 11, 2024.
- [2] T. Reiser, J. Dörpinghaus, and P. Steiner, "Analyzing historical legal textcorpora: German vet and cvet regulations," in *INFORMATIK 2024*. Gesellschaft für Informatik eV, 2024, pp. 2007–2018.
- [3] —, "Learning from historical vet and cvet regulations in germany: What should vet look like and whom should it serve?" in *NORDYRK 2024 BOOK OF ABSTRACTS*, 2024, p. 75.
- [4] 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.
- [5] 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/>
- [6] 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.
- [7] 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>

- [8] 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>
- [9] 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](https://doi.org/10.1109/HNICEM51456.2020.9400000) pp. 1–6.
- [10] 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](https://doi.org/10.1109/NICS54270.2021.9701547) pp. 519–524.
- [11] 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](https://doi.org/10.1109/5.156475)
- [12] 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](https://doi.org/10.1109/ICDAR.2007.4377049) pp. 919–923.
- [13] 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>
- [14] 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](https://doi.org/10.1109/SCSE59836.2023.10214989) pp. 1–7.
- [15] 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.
- [16] 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>
- [17] 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.
- [18] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012. ISBN 3642311636
- [19] R. Altenhöner, A. Berger, C. Bracht, P. Klimpel, S. Meyer, A. Neuburger, T. Stäcker, and R. Stein, "DFG practical guidelines on digitisation. updated version 2022," 2023.
- [20] M. Söylemez, B. Tekinerdogan, and A. Kolukisa Tarhan, "Challenges and solution directions of microservice architectures: A systematic literature review," *Applied Sciences*, vol. 12, no. 11, 2022. doi: [10.3390/app12115507](https://doi.org/10.3390/app12115507). [Online]. Available: <https://www.mdpi.com/2076-3417/12/11/5507>
- [21] B. Kim, S. Nakamura, and H. Watanabe, "Using archivematica and omeka s for long-term preservation and access of digitized archive materials," in *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*, Y.-H. Tseng, M. Katsurai, and H. N. Nguyen, Eds. Cham: Springer International Publishing, 2022, pp. 241–250.
- [22] M. Klindt and K. Amrhein, "One core preservation system for all your data. no exceptions!" in *iPRES 2015 - Proceedings of the 12th International Conference on Preservation of Digital Objects*, 2015, pp. 101 – 108. [Online]. Available: <http://phaidra.univie.ac.at/o:429551>
- [23] J. Dörpinghaus and M. Tiemann, "Vocational education and training data in twitter: Making german twitter data interoperable," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 946–948, 2023.
- [24] M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., *Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2024.

A Confidence-Interval Circular Intuitionistic Fuzzy Zero Point Model for Optimizing Spare Parts Transfer in Smart Manufacturing Environments

Velichka Traneva

BSU Prof. Dr Assen Zlatarov University
1 Prof. Yakimov Blvd, Burgas 8000, Bulgaria
Email: veleka13@gmail.com

Mihai Petrov

BSU Prof. Dr Assen Zlatarov University
1 Prof. Yakimov Blvd, Burgas 8000, Bulgaria
Email: mihpetrov@abv.bg

Stoyan Tranev

BSU Prof. Dr Assen Zlatarov University
1 Prof. Yakimov Blvd, Burgas 8000, Bulgaria
Email: tranev@abv.bg

Venelin Todorov

Institute of Mathematics and Informatics, BAS,
8 Acad. Georgi Bonchev Str., 1113 Sofia, Bulgaria
Institute of Information and Communication Technologies, BAS,
25A Acad. Georgi Bonchev Str., 1113 Sofia, Bulgaria
Email: venelintodorov@gmail.com

Abstract—In Industry 4.0 systems, timely delivery of critical components to maintenance points is essential for continuous operation. This paper introduces a novel Confidence-Interval Circular Intuitionistic Fuzzy Zero Point Method (CIC-IFZPM) to optimize the transfer of spare parts in a smart factory setting. The method addresses uncertainty in transfer cost, delivery time, and priority assessment through circular intuitionistic fuzzy sets (C-IFS), which reflect both membership and hesitancy with geometric interpretation. A customized version of the index matrix algorithm integrates transportation constraints, expert confidence intervals, and machine availability limitations. The model is validated through a simulated industrial scenario, where production cells request components dynamically, and a central warehouse must allocate them optimally. Compared to classical fuzzy optimization approaches, the proposed method ensures more robust decision-making under incomplete or imprecise data, offering better performance in real-time control environments. The framework is applicable to predictive maintenance logistics, autonomous scheduling, and industrial resilience planning.

I. INTRODUCTION

TRANSPORTATION problems (TPs) aim to determine optimal delivery routes that minimize total transportation costs. The classical formulation originated with Hitchcock in 1941 [6], followed by Dantzig's application of the simplex method [5] and Kantorovich's development of the "method of potentials" in 1949 [8]. In practice, however, transport systems operate under significant uncertainty caused by fluctuating fuel prices, economic volatility, and external disruptions.

To model such vagueness, fuzzy logic approaches have been widely adopted. Zadeh introduced fuzzy sets (FSs) in 1965 [23], and Atanassov later proposed intuitionistic fuzzy sets (IFSs) in 1983 [1], enabling more nuanced uncertainty

representation through the inclusion of membership, non-membership, and hesitation degrees.

Numerous fuzzy TP methods have since emerged, including the Zero Point Method applied to trapezoidal fuzzy data [13], and enhancements using triangular, LR-flat, and hybrid fuzzy numbers [9], [10], [15]. Comparative analyses suggest the Zero Point Method often outperforms classical techniques [12]. Further variants include the zero suffix method [7], IF Zero Suffix Method (IFZSM) [19], and IF Zero Point Method (IFZPM) [18]. The proposed IFZPM yielded a marginally better optimal solution than the previously established IFZSM [18] in the specific case study under consideration, demonstrating its potential for enhanced performance under fuzzy uncertainty.

To better capture multidimensional and circular uncertainties, Atanassov introduced the Circular Intuitionistic Fuzzy Set (C-IFS) [3] in 2020. Building upon this, we extend C-IFSs to Confidence-Interval Circular Intuitionistic Fuzzy Sets (CIC-IFSs) [22], where each fuzzy element is represented as a circular region whose radius varies with a confidence level β .

This paper proposes a novel CIC-IF Zero Point Method for solving transportation problems under the CIC-IFTP framework. Transportation costs, supply, and demand are modeled as CIC-IF triples [22], incorporating expert evaluations and uncertainty quantification. The solution algorithm builds on the index matrix (IM) approach [2], while introducing additional constraints such as transport cost caps and confidence-based tolerances. The proposed algorithm is an extension of the classical Zero Point Method [15], designed to accommodate uncertain environments by incorporating circular intuitionistic fuzzy representations and the level of confidence.

To demonstrate applicability, we consider a smart manufacturing scenario involving the dynamic reallocation of spare

This work was funded by the University-Wide Research Grant No. OUF-RD-15/2025 at Prof. Dr. Assen Zlatarov Burgas State University: "Extraction of Expert Knowledge through Innovative Analytical Methods."

parts. This environment is characterized by uncertain delivery times and competing demands from maintenance units. Our key contributions include the formalization of the CIC-IFTP framework, a robust solution algorithm, and validation through an industrial case study reflecting predictive maintenance logistics.

The paper is structured as follows: Section II introduces preliminaries on CIC-IF triples and index matrices. Section III details the problem formulation, solution procedure, and industrial case study. Section IV discusses computational results and future work.

II. PRELIMINARIES

This section recalls the key concepts underpinning the proposed approach: Confidence-Interval Circular Intuitionistic Fuzzy Sets (CIC-IFSs), Triples (CIC-IFTs), and Index Matrices (CIC-IFIMs). We define each structure and the operations applicable to them.

A. Confidence Interval Circular Intuitionistic Fuzzy Sets (CIC-IFSs)

The definitions and properties of CIC-IF sets used in this paper follow the construction proposed in [22]. Let $A \subseteq E$, where E is a universe of discourse. A CIC-IFS with confidence level β is defined as:

$$A_u^\beta = \{ \langle x, \mu_A(x), \nu_A(x); u^\beta \rangle \mid x \in E \},$$

where $\mu_A(x) + \nu_A(x) \leq 1$, and $u^\beta \in [0, \sqrt{2}]$ is the circular radius expressing confidence-based fuzziness. The uncertainty margin $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$ complements the membership functions.

The center of the circle is computed as:

$$\langle \mu(x)^\beta, \nu(x)^\beta \rangle = \left\langle \frac{a(x) + c(x)}{2}, \frac{b(x) + d(x)}{2} \right\rangle.$$

Its radius is obtained from the maximal Euclidean deviation between this center and individual expert evaluations:

$$u^\beta(x) = \max_{1 \leq i \leq k_x} \sqrt{(\mu(x)^\beta - \mu_{k_i}^\beta)^2 + (\nu(x)^\beta - \nu_{k_i}^\beta)^2}.$$

B. Confidence Interval Circular Intuitionistic Fuzzy Triples (CIC-IFTs)

The formalization of CIC-IFTs and related operations is based on the framework introduced in [22]. Given expert assessments for assertion p , we define a CIC-IFT as:

$$\langle \mu(p), \nu(p); u^\beta \rangle = \langle a(p), b(p); u^\beta \rangle, \quad \text{where } a(p) + b(p) \leq 1.$$

The center and radius follow similar constructions, based on bounds $a(p), b(p), c(p), d(p)$, and the maximum deviation from the confidence center.

CIC-IFTs are closed under operations such as $\wedge, \vee, +, \bullet, -$, with radius propagation via max/min functions. Comparison and ranking between two CIC-IFTs is performed by dominance or by proximity to the ideal $\langle 1, 0; \sqrt{2} \rangle$.

C. 3D Confidence Interval Circular Intuitionistic Fuzzy Index Matrices (3D CIC-IFIMs)

CIC-IFIMs extend index matrices to a three-dimensional structure:

$$A^\beta = [K, L, H, \{ \langle \mu_{k_i, l_j, h_g}, \nu_{k_i, l_j, h_g}; r_{k_i, l_j, h_g}^\beta \rangle \}].$$

Each entry represents a CIC-IFT, structured along three index dimensions—supply, demand, and operational scenar-

ios—defined as subsets $K, L, H \subset \mathcal{S}$. The associated operations over CIC-IFIMs follow extensions of fuzzy matrix logic, as discussed in [22].

To further process multidimensional data, we apply aggregation operations (AOs) over one dimension of a 3D CIC-IFIM [22]. Let $*$ $\in \{\min, \max\}$ be a binary operator. Ten aggregation operations $\#_i$ ($1 \leq i \leq 10$) are defined over two CIC-IFTs $x = \langle a, b; r_{f_1}^\beta \rangle$ and $y = \langle c, d; r_{f_2}^\beta \rangle$, including for example:

$$x \#_1 y = \langle ac, 1 - ac, *(r_{f_1}^\beta, r_{f_2}^\beta) \rangle,$$

$$x \#_{10} y = \langle \min(1, 2 - b - d), \max(0, b + d - 1), *(r_{f_1}^\beta, r_{f_2}^\beta) \rangle.$$

Let $k_0 \notin K$ be an artificial aggregation index. The operator $\alpha_{K, \#_q, *}(A^\beta, k_0)$ aggregates over the supply dimension K , yielding:

$h_g \in H$	l_1	$\dots$	l_n
k_0	$\#_{q,*}^m$ $i=1$	$\langle \mu_{k_i, l_1, h_g}, \nu_{k_i, l_1, h_g}; r_{k_i, l_1, h_g}^\beta \rangle$	$\dots$
	$\#_{q,*}$	$\dots$	$\#_{q,*}^m$ $i=1$
			$\langle \mu_{k_i, l_n, h_g}, \nu_{k_i, l_n, h_g}; r_{k_i, l_n, h_g}^\beta \rangle$

Depending on the scenario, we may choose $\#_1^*$ for pessimistic aggregation (e.g., inflation), $\#_5^*$ or $\#_6^*$ for moderate strategies, and $\#_{10}^*$ for optimistic planning.

III. PROBLEM STATEMENT: CIC-IFTP IN SMART FACTORY MAINTENANCE LOGISTICS

We extend the C-IFTP framework from [20] to a novel Confidence-Interval Circular Intuitionistic Fuzzy Transportation Problem (CIC-IFTP) with operational constraints tailored for smart manufacturing environments.

A smart electronics factory seeks to optimize the allocation of critical spare parts—such as sensors, microcontrollers, and actuators—from central storage units $\{k_1, \dots, k_m\}$ to maintenance stations $\{l_1, \dots, l_n\}$. The available stock at each storage unit is denoted by $c_{k_i, R}$, while each station requests c_{Q, l_j} units.

Each internal transport route (robotic or conveyor-based) from k_i to l_j has an operational usage threshold c_{pl, l_j} and a unit transportation cost c_{k_i, l_j} . These parameters are subject to uncertainty and are evaluated by a panel of planners $\{d_1, \dots, d_D\}$, who provide intuitionistic fuzzy (IF) preferences represented as $\text{re}_s = \langle \delta_s, \epsilon_s \rangle$. Based on their assessments, circular intuitionistic fuzzy (CIF) data are constructed using a selected confidence level β .

The decision variable x_{k_i, l_j} denotes the number of units to be routed from storage unit k_i to station l_j . Ten scenario strategies—ranging from highly pessimistic to highly optimistic—guide the decision-making process under uncertainty.

The objective is to minimize the total CIC-IF transportation cost, subject to the following constraints:

- All station demands c_{Q, l_j} must be satisfied;
- Storage capacities $c_{k_i, R}$ must not be exceeded;
- Route-specific operational thresholds c_{pl, l_j} must be respected.

To solve this problem, we propose a novel *Confidence-Interval Circular Intuitionistic Fuzzy Zero Point Method (CIC-IFZPM)*. This algorithm extends the index matrix-based approach introduced in [20]. While rooted in the classical Zero Point Method, CIC-IFZPM advances the intuitionistic fuzzy modeling framework by incorporating confidence-interval circular intuitionistic fuzzy structures [22]. These enhancements

allow the model to more accurately capture imprecision, expert disagreement, and cyclic interdependencies—phenomena frequently encountered in smart manufacturing and logistics environments. The detailed solution algorithm is presented in the following section.

A. The Solution Algorithm

Algorithm 1 Construction of CIC-IFIM (Step 1)

Require: Sets $K = \{k_1, \dots, k_m\}$, $L = \{l_1, \dots, l_n\}$, experts $E = \{d_1, \dots, d_D\}$, confidence coefficient $\beta \in [0, 1]$

- 1: **for** $j = 1$ to n **do**
- 2: **for** $i = 1$ to m **do**
- 3: Each expert $d_s \in E$ provides IF evaluation $ev_{i,j,s} = \langle \mu_{i,j,s}, \nu_{i,j,s} \rangle$ and reliability $re_s = \langle \delta_s, \varepsilon_s \rangle$
- 4: Aggregate evaluations: $EV_{i,j}^* \leftarrow \bigoplus_{s=1}^D re_s \cdot ev_{i,j,s}$
- 5: Apply circular IF aggregation: $pi_{i,j}^{ave} \leftarrow \alpha_{E,\#2}(EV_{i,j}^*)$
- 6: Compute radius: $r_{i,j}^\beta \leftarrow \max_{1 \leq s \leq D} \left(\sqrt{(\mu_{i,j,s} - \mu_{i,j})^2 + (\nu_{i,j,s} - \nu_{i,j})^2} \right)$
- 7: Form CIC-IFT: $c_{i,j}^\beta \leftarrow \langle \mu_{i,j}, \nu_{i,j}; r_{i,j}^\beta \rangle$
- 8: **end for**
- 9: **end for**
- 10: Assemble CIC-IFIM: $C^\beta[K, L] = \{c_{i,j}^\beta\}_{i=1..m, j=1..n}$
- 11: Extend to complete matrix: $C^\beta[K^*, L^*]$ by adding artificial nodes Q, pl, pu_1, R, pu

	...	l_n	R	pu
k_1	...	$\langle \mu_{k_1,l_n}^c, \nu_{k_1,l_n}^c; r_{k_1,l_n}^c \rangle$	$\langle \mu_{k_1,R}^c, \nu_{k_1,R}^c; r_{k_1,R}^c \rangle$	$\langle \mu_{k_1,pu}^c, \nu_{k_1,pu}^c; r_{k_1,pu}^c \rangle$
$\vdots$		$\vdots$	$\vdots$	$\vdots$
k_m	...	$\langle \mu_{k_m,l_n}^c, \nu_{k_m,l_n}^c; r_{k_m,l_n}^c \rangle$	$\langle \mu_{k_m,R}^c, \nu_{k_m,R}^c; r_{k_m,R}^c \rangle$	$\langle \mu_{k_m,pu}^c, \nu_{k_m,pu}^c; r_{k_m,pu}^c \rangle$
Q	...	$\langle \mu_{Q,l_n}^c, \nu_{Q,l_n}^c; r_{Q,l_n}^c \rangle$	$\langle \mu_{Q,R}^c, \nu_{Q,R}^c; r_{Q,R}^c \rangle$	$\langle \mu_{Q,pu}^c, \nu_{Q,pu}^c; r_{Q,pu}^c \rangle$
pl	...	$\langle \mu_{pl,l_n}^c, \nu_{pl,l_n}^c; r_{pl,l_n}^c \rangle$	$\langle \mu_{pl,R}^c, \nu_{pl,R}^c; r_{pl,R}^c \rangle$	$\langle \mu_{pl,pu}^c, \nu_{pl,pu}^c; r_{pl,pu}^c \rangle$
pu_1	...	$\langle \mu_{pu_1,l_n}^c, \nu_{pu_1,l_n}^c; r_{pu_1,l_n}^c \rangle$	$\langle \mu_{pu_1,R}^c, \nu_{pu_1,R}^c; r_{pu_1,R}^c \rangle$	$\langle \mu_{pu_1,pu}^c, \nu_{pu_1,pu}^c; r_{pu_1,pu}^c \rangle$

Fig. 1. Extended CIC-IFIM with CIC-IF entries

Step 1 (continued). Initialization of Auxiliary Matrices

After constructing the CIC-IFIM matrix $C^\beta[K, L, h_f]$, we extend the sets:

- $K^* = K \cup \{Q, pl, pu_1\} \Rightarrow |K^*| = m + 3$;
- $L^* = L \cup \{R, pu\} \Rightarrow |L^*| = n + 2$.

Then, the following auxiliary matrices are initialized:

- 1) **State matrix** $S^\beta[K^*, L^*]$: Initialized as a duplicate of C^β , i.e., $s_{k_i,l_j}^\beta = c_{k_i,l_j}^\beta$.
- 2) **Discard matrix** $D[K, L]$: Each $d_{k_i,l_j} \in \{1, 2\}$ tracks the number of times a cell has been eliminated.
- 3) **Row crossing indicator** $RC[K]$: $rc_{k_i,e_0} \in \{0, 1\}$ indicates whether row k_i has been excluded.
- 4) **Column crossing indicator** $CC[L]$: $cc_{r_0,l_j} \in \{0, 1\}$ indicates whether column l_j has been excluded.
- 5) **Projections**:

- $RM[K, R] = pr_{K,R}(C)$
- $CM[pu_1, L] = pr_{pu_1,L}(C)$

Used in balancing, particularly regarding R and pu_1 nodes.

- 6) **Utility matrix** $U[K, L]$: Defined as:

$$u_{k_i,l_j} = \begin{cases} 1, & \text{if } c_{k_i,l_j} < c_{pl,l_j} \\ \perp, & \text{otherwise} \end{cases}$$

- 7) **Initial allocation matrix** $X[K, L]$: All entries start as $x_{k_i,l_j}^\beta = \langle 0, 1; \sqrt{2} \rangle$.

Initial Setup: All indicators are initialized as:

$$rm_{k_i,R} = rc_{k_i,e_0} = cc_{r_0,l_j} = cm_{pu_1,l_j} = 0, \quad u_{k_i,l_j} = \perp.$$

Balancing the System: If the transportation problem is unbalanced (i.e., $\sum \text{Supply} \neq \sum \text{Demand}$), balancing is applied by adding artificial nodes Q, pu, R with synthetic cost entries, following [18].

The algorithm then proceeds to **Step 2**.

Step 2. Verifying the Transportation Cost Constraints

For each warehouse $k_i \in K$ and destination cell $l_j \in L$, verify whether the transportation cost from k_i to l_j is strictly preferable to the baseline from the pseudo-node pl . Iterate:

for $i = 1$ to m : **for** $j = 1$ to n :

if $\left(\left[\frac{k_i}{pl}; \perp \right] pr_{pl,l_j} C^\beta \right) \supset_v pr_{k_i,l_j} C^\beta$ **then** $u_{k_i,l_j} \leftarrow 1$

After evaluating all entries, define the set of non-preferable allocations:

$$EG = \text{Index}_{(\perp)}(U) = \{ \langle k_{i_1}, l_{j_1} \rangle, \langle k_{i_2}, l_{j_2} \rangle, \dots, \langle k_{i_\phi}, l_{j_\phi} \rangle \}$$

For each $\langle k_i, l_j \rangle \in EG$, penalize the corresponding entry in matrix S as:

$$s_{k_i,l_j}^\beta \leftarrow \langle 1, 0; \sqrt{2} \rangle \quad (\text{as in [11]})$$

Proceed to **Step 3**.

Step 3. Row-Level Normalization Using Zero Membership Values

In this step, we compute a row-specific zero-cost reference based on membership minimization. This value is stored in the auxiliary column pu , facilitating row-wise normalization aligned with the Zero Point Method principle.

- 1) **Identify Minimum Cost Elements:**

For each row $i = 1$ to m , determine the minimum CIC-IF value among columns $j = 1$ to n using a selected aggregation index:

$$AGIndex_{\{\min, \min_\square, \min_\circ, \min_{K}^{circ}\}}(pr_{k_i,L} S^\beta) = \langle k_i, l_{v_j} \rangle$$

- 2) **Compare to Baseline:**

If the minimal cell is still preferable (or at least non-worse) than the baseline pseudo-node pl :

$$pr_{k_i,l_{v_j}} S^\beta \subseteq_v \left(\left[\frac{k_i}{pl}; \perp \right] pr_{pl,l_{v_j}} S^\beta \right),$$

then define:

$$S_6^\beta[k_i, l_{v_j}] := pr_{k_i,l_{v_j}} S^\beta$$

$$S_7^\beta := \left[\perp; \frac{pu}{l_{v_j}} \right] S_6^\beta$$

$$S^\beta := S^\beta \oplus_{(\circ_1, \circ_2, *)} S_7^\beta$$

- 3) **Perform Row-Wise Normalization:**

For every $i = 1, \dots, m$ and $j = 1, \dots, n$, apply:

$$IO_{-(\circ_1, \circ_2, *)}(\langle k_i, l_j, S^\beta \rangle, \langle k_i, pu, pr_K S^\beta \rangle)$$

This operation ensures that each row in S^β has at least one cell with minimal (zero-like) cost, allowing a valid zero-point to be selected in the next stage.

After this normalization step, continue to **Step 4**.

Step 4. Column-Level Zero Membership Normalization

In this step, we determine the minimum cost for each column of the matrix S^β and normalize the elements accordingly.

1) Identify minimum cost per column:

for $j = 1$ to n :

$$AGIndex_{\{\min, \min_\square, \min_\circ, \min_{R^\circ}\}}(pr_{K,l_j} S^\beta) = \langle k_{w_i}, l_j \rangle$$

2) Construct intermediate matrices and adjust:

$$S_6^\beta[k_{w_i}, l_j] := pr_{k_{w_i}, l_j} S^\beta, \quad S_7^\beta := \left[\frac{pu_1}{k_{w_i}}; \perp \right] S_6^\beta$$

$$S^\beta := S^\beta \oplus_{(\circ_1, \circ_2, *)} S_7^\beta$$

3) Normalize column-wise:

for $j = 1$ to n , for $i = 1$ to m :

$$IO_{-(\circ_1, \circ_2, *)}(\langle k_i, l_j, S^\beta \rangle, \langle pu_1, l_j, pr_{pu_1, L} S^\beta \rangle)$$

Proceed to **Step 5**.

Step 5. Optimality Criteria Evaluation

Step 5.1. For each warehouse $k_i \in K$, verify whether the total offered quantity is less than or equal to the sum of allocations with zero membership degree:

for $i = 1$ to m : $Index_{(\min \mu), k_i}(C^\beta) = \{\langle k_i, l_{v_1} \rangle, \dots, \langle k_i, l_{v_V} \rangle\}$

$$G_{v_r}^\beta[k_i, l_{v_r}] := pr_{k_i, l_{v_r}} C^\beta, \quad G^\beta[k_i, R] := pr_{k_i, R} C^\beta$$

If:

$$G^\beta[k_i, R] \subseteq_v \bigoplus_{r=1}^V G_{v_r}^\beta, \quad \text{then go to Step 5.2;}$$

else set $RM[k_i, R] := 1$ and go to Step 6.

Step 5.2. For each region $l_j \in L$, verify whether the required quantity does not exceed the sum of allocated zero-membership values:

for $j = 1$ to n : $Index_{(\min \mu), l_j}(C^\beta) = \{\langle k_{w_1}, l_j \rangle, \dots, \langle k_{w_W}, l_j \rangle\}$

$$G_{w_r}^\beta[k_{w_r}, l_j] := pr_{k_{w_r}, l_j} C^\beta, \quad G^\beta[pu_1, l_j] := pr_{pu_1, l_j} C^\beta$$

If:

$$G^\beta[pu_1, l_j] \subseteq_v \bigoplus_{r=1}^W G_{w_r}^\beta, \quad \text{then go to Step 8;}$$

else set $CM[pu_1, l_j] := 1$ and go to Step 6.

Step 6. Update the Cost CIC-IF Index Matrix

In this step, the matrix S^β (initially identical to C^β) is refined to improve cost allocation feasibility. All elements of the form $\langle 0, 1; r_{k_i, l_j}^{\beta, c} \rangle$ for $i = 1, \dots, m$ and $j = 1, \dots, n$ are considered “zero-membership” and marked for elimination through the minimal number of horizontal and vertical lines.

- If a row or column contains no such zero-membership entry, the element with the smallest membership degree is crossed out.
- The auxiliary matrix $D[K, L]$ tracks eliminations: $d_{k_i, l_j} = 1$ for one line, $d_{k_i, l_j} = 2$ for both.
- Two matrices, $RC[K]$ and $CC[L]$, indicate whether a row or column is crossed: $rc_{k_i, e_0} \in \{0, 1\}$, $cc_{r_0, l_j} \in \{0, 1\}$.

For each $i = 1$ to m , $j = 1$ to n :

- If $s_{k_i, l_j}^\beta = \langle 0, 1; r_{k_i, l_j}^{\beta, c} \rangle$ and $rm_{k_i, R} = 0$ and $d_{k_i, l_j} = 0$, then:
 $rc[k_i, e_0] := 1; \quad d_{k_i, l_j} := 1 \quad \forall l_j, \text{ in row } S_{(k_i, \perp)}^\beta$
- If $s_{k_i, l_j}^\beta = \langle 0, 1; r_{k_i, l_j}^{\beta, c} \rangle$, $cm_{pu_1, l_j} = 0$ and $d_{k_i, l_j} = 1$, then:
 $d_{k_i, l_j} := 2; \quad cc_{r_0, l_j} := 1;$
 $d_{k_i, l_j} := 1 \quad \forall k_i, \text{ in column } S_{(\perp, l_j)}^\beta$

Step 7. Refinement of the Revised Cost Matrix

Identify the smallest non-crossed cost element in S^β and subtract it from all uncovered entries. Then, add the same value to each entry covered by two lines.

1) Identify minimal non-covered cost:

$$\langle k_x, l_y \rangle := AGIndex_{(\min, \max)}(S^\beta)$$

2) Adjust uncovered entries:

$$IO_{-(\circ_1, \circ_2, *)}(S^\beta, \langle k_x, l_y, S^\beta \rangle)$$

3) Adjust double-covered entries:

- For $d_{k_i, l_j} = 2$:

$$S_1^\beta := pr_{k_x, l_y} C^\beta,$$

$$S_2^\beta := pr_{k_i, l_j} C^\beta \oplus_{(\circ_1, \circ_2, *)} \left[\frac{k_i}{k_x}; \frac{l_j}{l_y} \right] S_1^\beta$$

$$S^\beta := S^\beta \oplus_{(\circ_1, \circ_2, *)} S_2^\beta$$

- For $d_{k_i, l_j} = 1$:

$$S^\beta := S^\beta \oplus_{(\circ_1, \circ_2, *)} pr_{k_i, l_j} C^\beta$$

Proceed to **Step 8**.

Step 8. Allocation of Maximum Feasible Quantity

1) Find the cell with the largest cost value in S^β using:

$$AGIndex_{(\max, \min, *)}(S^\beta) = \langle k_x^*, l_y^* \rangle$$

2) Assign the maximum feasible quantity to this cell and reduce either the row or column:

- Compare:

$$s_{Index(\min \mu), k_x^*}^\beta(C^\beta) \quad \text{and} \quad s_{Index(\min \mu), l_y^*}^\beta(C^\beta)$$

- Let s_{k_e, l_g}^β be the lesser of the two. Assign it and reduce S^β accordingly:

- If $s_{k_e, R}^\beta < s_{Q, l_g}^\beta$:

$$X^\beta := X^\beta \oplus_{(\circ_1, \circ_2, *)} \left[\perp; \frac{l_g}{R} \right] S_8^\beta$$

(reduce row)

- Else:

$$X^\beta := X^\beta \oplus_{(\circ_1, \circ_2, *)} \left[\frac{k_e}{Q}; \perp \right] S_9^\beta$$

(reduce column)

Repeat Step 8 until $|S^\beta| = 6$. Then proceed to **Step 9**.

Step 9. Degeneracy Resolution in IF Solution

If $|D| < m + n - 1$, introduce a new basic variable $x_{k_\alpha, l_\beta}^\beta$ at the minimum cost among unassigned cells:

$$AGIndex_{\{(\min/\max), \mathcal{L}, \notin D\}}(C^\beta) = \langle k_\alpha, l_\beta \rangle$$

Assign:

$$x_{k_\alpha, l_\beta}^\beta := \langle 0, 1; 0 \rangle$$

Step 10. Finalizing the IF Transportation Plan

- 1) If for any $x_{k_i, l_j}^\beta \neq \langle \perp, \perp \rangle$ and $\langle k_i, l_j \rangle \in EG$, the problem is infeasible. Stop.
- 2) Otherwise, define the final optimal IF transport plan:

$$X_{opt}^\beta[K, L, \{x_{k_i, l_j}^\beta\}]$$
- 3) Assign defaults for unallocated cells:

$$x_{k_i, l_j}^\beta = \langle 0, 1; \sqrt{2} \rangle \text{ if } x_{k_i, l_j}^\beta = \langle \perp, \perp \rangle$$
- 4) Compute the total aggregated transportation cost:

$$AGIO_{\oplus(\max, \min, *)}^1 \left(C_{\{Q, pl, pu_1\}, \{R, pu\}}^\beta \otimes_{(\min, \max, *)} X_{opt}^\beta \right)$$
- 5) Finally, de-fuzzify each circular IF value to obtain a crisp fuzzy pair (as in [4]):

$$\left\langle \frac{a}{a+b}, \frac{b}{a+b} \right\rangle$$

B. Illustrative Example: Smart Factory Spare Parts Distribution

We demonstrate the application of the proposed Confidence-Interval Circular Intuitionistic Fuzzy Transportation Problem (CIC-IFTP) model to a smart manufacturing scenario, where spare parts must be dynamically dispatched from warehouse units to robotic production cells across a cyber-physical shopfloor.

Let:

$K = \{k_1, k_2, k_3\}$: three autonomous warehouse stations; $L = \{l_1, l_2, l_3, l_4\}$: four production cells requiring critical spare components; $E = \{d_1, d_2, d_3\}$: expert panel assessing fuzzy transport costs under uncertainty; $\beta = 0.85$: confidence level chosen by the system operator.

Each expert $d_s \in E$ provides a confidence-weighted intuitionistic fuzzy evaluation $ev_{k_i, l_j, d_s} = \langle \mu_{k_i, l_j, d_s}, \nu_{k_i, l_j, d_s} \rangle$, with individual ratings $re_s = \langle \delta_s, \varepsilon_s \rangle$ as follows:

$$\{re_1, re_2, re_3\} = \{\langle 0.65, 0.10 \rangle, \langle 0.55, 0.08 \rangle, \langle 0.75, 0.07 \rangle\}.$$

Using the aggregation operator $\alpha_{E, \#_2}$, we compute the adjusted cost matrix PI^{ave} , followed by transformation into the Confidence-Interval Circular Intuitionistic Fuzzy Cost Matrix C^β at decision moment h_f . Each entry $c_{k_i, l_j}^\beta = \langle \mu^\beta, \nu^\beta; r^\beta \rangle$ includes a radius r^β that geometrically captures expert disagreement at confidence level β .

The resulting CIC-IFIM for the smart transportation problem appears as:

	l_1	l_2	l_3	l_4	R	pu
k_1	$\langle 0.52, 0.28; 0.2 \rangle$	$\langle 0.62, 0.18; 0.3 \rangle$	$\langle 0.22, 0.18; 0.2 \rangle$	$\langle 0.72, 0.18; 0.2 \rangle$	$\langle 0.42, 0.28; 0.3 \rangle$	$\langle \perp, \perp \rangle$
k_2	$\langle 0.42, 0.35; 0.2 \rangle$	$\langle 0.32, 0.18; 0.3 \rangle$	$\langle 0.42, 0.18; 0.3 \rangle$	$\langle 0.22, 0.28; 0.3 \rangle$	$\langle 0.62, 0.18; 0.3 \rangle$	$\langle \perp, \perp \rangle$
k_3	$\langle 0.32, 0.28; 0.3 \rangle$	$\langle 0.22, 0.28; 0.2 \rangle$	$\langle 0.52, 0.18; 0.3 \rangle$	$\langle 0.62, 0.28; 0.2 \rangle$	$\langle 0.32, 0.58; 0.3 \rangle$	$\langle \perp, \perp \rangle$
Q	$\langle 0.32, 0.28; 0.3 \rangle$	$\langle 0.42, 0.38; 0.3 \rangle$	$\langle 0.52, 0.28; 0.3 \rangle$	$\langle 0.02, 0.10; 0.3 \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$
pl	$\langle 0.47, 0.38; 0.3 \rangle$	$\langle 0.52, 0.48; 0.3 \rangle$	$\langle 0.67, 0.28; 0.3 \rangle$	$\langle 0.7, 0.33; 0.2 \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$
pu_1	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$	$\langle \perp, \perp \rangle$

The total transport cost is computed using:

$$AGIO_{\oplus(\max, \min, *)}^1 \left(C_{\{Q, pl, pu_1\}, \{R, pu\}} \otimes_{(\min, \max, *)} X_{opt} \right).$$

For the scenario with minimal uncertainty and pessimistic valuation, the final cost is represented by the CIC-IF triple:

$$\langle 0.42, 0.28; 0.2 \rangle \Rightarrow$$

Fuzzy projection: $\langle 0.6, 0.4 \rangle \Rightarrow$ Crisp cost: 5344.

In comparative analysis: Optimistic scenario yields: $\langle 0.08, 0.42; 0.2 \rangle \rightarrow 1869$; Realistic scenario: $\langle 0.25, 0.56; 0.2 \rangle \rightarrow 2769.62$.

The ranking function $R^{\beta, circ}$ assists in determining scenario preferences and strategic planning under expert-informed fuzziness in the smart factory.

C. Method Validation and Comparative Advantage

The Improved Zero Point Method (IZPM) has been demonstrated to consistently outperform well-known heuristics such as VAM, SVAM, GVAM, BVAM, and RVAM for both crisp and fuzzy transportation problems [15]. Unlike these methods, which often lead to suboptimal or infeasible solutions, IZPM ensures optimality through a structured and robust process, even in unbalanced settings.

Building upon this strong foundation, the proposed CIC-IFZPM algorithm extends IZPM by introducing two key enhancements: (1) modeling circular intuitionistic fuzzy data, which allows capturing cyclic uncertainty in dynamic environments, and (2) applying confidence intervals to represent varying degrees of reliability in expert estimates.

In the considered case study, CIC-IFZPM produced a slightly improved transportation cost compared to the previously proposed IFZSM [18], demonstrating its enhanced capability to yield robust solutions in uncertain environments. Moreover, the method preserved the optimality structure of the classical ZPM while effectively generalizing it to a broader fuzzy framework.

The time complexity of the CIC-IFZPM algorithm is $\mathcal{O}(2Dmn + 13mn + m + n)$, where D is the number of experts, and m, n are the numbers of supply and demand nodes, respectively. The dominant computational cost arises from the expert-based construction of the CIC-IFIM matrix and the structured zero-point allocation. However, since D is typically a small constant (e.g., 3–7 experts), the overall complexity scales linearly with the number of experts and quadratically with the transportation matrix size. Thus, the method retains the same computational class as the classical Improved Zero Point Method [16].

Although only one illustrative example is currently provided, a dedicated software implementation of the CIC-IFZPM algorithm is under development. This implementation will enable large-scale computational experiments on real and synthetic datasets, including stress tests in highly uncertain circular fuzzy environments. The objective is to rigorously examine the algorithm's efficiency, scalability, and stability across various industrial configurations. These experiments are planned for a follow-up study, which will also include a public release of the source code.

Table I provides a concise comparative overview of the main methods used for solving transportation problems, highlighting their scope and distinguishing features.

IV. CONCLUSION

This paper introduces an advanced CIC-IF Zero-Point Method tailored for Transportation Problems under uncertainty, within the CIC-IFTP framework. The proposed methodology leverages Confidence-Interval Circular Intuitionistic Fuzzy Numbers (CIC-IFNs) and the concept of index matrices to enhance decision-making in environments characterized by imprecise cost evaluations, dynamic supply availability, and fluctuating demand. Applied to a smart factory logistics scenario, the algorithm demonstrates its capability to model and resolve

TABLE I
OVERVIEW OF MAIN METHODS FOR SOLVING TRANSPORTATION PROBLEMS

Abbr.	Full Name	Description
VAM	Vogel's Approximation Method	Classical heuristic for initial feasible solution based on penalty costs; does not guarantee optimality.
SVAM	Shimshak's Vogel's Approximation Method	Modification of VAM that ignores penalties involving dummy rows/columns.
GVAM	Goyal's Vogel's Approximation Method	Assigns maximum transportation cost to dummy cells instead of zero.
BVAM	Balakrishnan's Vogel's Approximation Method	Enhances SVAM with additional allocation rules.
RVAM	Ramakrishnan's Vogel's Approximation Method	Uses four-step reduction and VAM for better approximation.
MODI	Modified Distribution Method	Optimizes allocations after an initial feasible solution.
FMDM	Fuzzy Modified Distribution Method	Fuzzy version of MODI that handles trapezoidal fuzzy numbers.
FVAM	Fuzzy Vogel's Approximation Method	Fuzzy adaptation of VAM using fuzzy numbers for cost, supply, and demand.
FZPM	Fuzzy Zero Point Method	One-stage method using fuzzy arithmetic; often gives optimal results directly.
IZPM	Improved Zero Point Method	Enhanced ZPM proven to outperform many heuristics; guarantees optimality.
CI-CIFZPM	Confidence-Interval Circular Intuitionistic Fuzzy Zero Point Method	Our proposed extension of IZPM that integrates circular intuitionistic fuzzy data and confidence intervals for robust decision-making under uncertainty.

the allocation of spare parts across cyber-physical production cells. It robustly accounts for expert uncertainty through adjustable confidence levels and scenario analysis (pessimistic, realistic, optimistic). The model ensures feasible and non-degenerate solutions, even in the presence of incomplete or ambiguous information.

The approach notably mitigates degeneracy through a structured mechanism of IF-index tracking and scenario-based refinement. It proves especially effective in applications requiring adaptive, multi-expert-driven decision processes, such as Industry 4.0 systems, smart supply chains, and decentralized logistics.

From a methodological standpoint, the proposed CIC-IF ZPM provides several key contributions: it integrates intuitionistic fuzzy logic with circular preference structures and confidence intervals; it generalizes the classic Zero-Point Method by embedding it in a multi-scenario fuzzy setting; it introduces a novel mechanism for degeneracy resolution via circular ranking indices.

These advances position the method as a competitive alternative to traditional fuzzy or crisp TP solvers, especially in complex environments where expert evaluation under uncertainty is essential. Related concepts for modeling uncertainty in logistics using granular computing can be found in [17], while multi-criteria optimization frameworks for transport tasks are also discussed in [14].

Future developments will focus on extending the model to handle confidence-interval elliptic intuitionistic fuzzy structures [21], with applications in elliptic IF multi-criteria decision-making. Additionally, dedicated software tools will be developed to support practical deployment in intelligent logistics platforms.

REFERENCES

- [1] K. Atanassov, "Intuitionistic Fuzzy Sets," VII ITKR Session, Sofia, 20-23 June 1983 (Deposited in Centr. Sci.-Techn. Library of the Bulg. Acad. of Sci., 1697/84) (in Bulgarian). Reprinted: *Int. J. Bioautomation*, vol. 20(S1), 2016, pp. S1-S6.
- [2] K. Atanassov, "Index Matrices: Towards an Augmented Matrix Calculus," *Studies in Computational Intelligence*, Springer, Cham, vol. 573, 2014, DOI: 10.1007/978-3-319-10945-9.
- [3] K. Atanassov, "Circular Intuitionistic Fuzzy Sets," *Journal of Intelligent & Fuzzy Systems*, vol. 39 (5), 2020, pp. 5981-5986.
- [4] Atanassova, L., "Three de-intuitionistic fuzzification procedures over circular intuitionistic fuzzy sets," *NIFS*, **29** (3), 2023, 292-297.
- [5] G. Dantzig, *Application of the simplex method to a transportation problem*, Chapter XXIII, Activity analysis of production and allocation, New York, Wiley, Cowles Commission Monograph, vol. 13, 359-373; 1951
- [6] F. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematical Physics*, vol. 20, 1941, 224-230.
- [7] R. Jahirhussain, P. Jayaraman, "Fuzzy optimal transportation problem by improved zero suffix method via robust rank techniques," *International Journal of Fuzzy Mathematics and Systems (IJFMS)*, vol. 3 (4), 2013, 303-311.
- [8] L. Kantorovich, M. Gavyrin, "Application of mathematical methods in the analysis of cargo flows," *Coll. of articles Problems of increasing the efficiency of transport*, M.: Publ. house AHSSSR, 1949, 110-138 (in Russian).
- [9] T. Karthy, K. Ganesan, K., "Revised improved zero point method for the trapezoidal fuzzy transportation problems," *AIP Conference Proceedings*, vol 2112 (020063), 2019, 1-8.
- [10] A. Kaur, J. Kacprzyk, A. Kumar, "Fuzzy transportation and transshipment problems," *Studies in fuzziness and soft computing*, vol. 385, 2020
- [11] N. Lalova, L. Ilieva, S. Borisova, L. Lukov, V. Mirianov, *A guide to mathematical programming*, Science and Art Publishing House, Sofia; 1980 (in Bulgarian)
- [12] S. Mohideen, P. Kumar, P., "A Comparative Study on Transportation Problem in Fuzzy Environment," *International Journal of Mathematics Research*, vol. 2 (1), 2010, 151-158.
- [13] P. Pandian, G. Natarajan, "A new algorithm for finding a fuzzy optimal solution for fuzzy transportation problems," *Applied Mathematical Sciences*, vol. 4, 2010, 79- 90.
- [14] J. Panek, "Multicriteria Optimization Methods for Transport Accessibility

- ity Modelling,” *Annals of Computer Science and Information Systems*, vol. 15, 2018, pp. 793–800.
- [15] A. Samuel, M. Venkatachalapathy, “Improved zero point method for unbalanced FTPs,” *International Journal of Pure and Applied Mathematics*, vol. 94 (3), 2014, 419-424.
- [16] A. Samuel, “Improved Zero Point Method (IZPM) for the Transportation Problems,” *Applied Mathematical Sciences*, vol. 6 (109), 2012, 5421 - 5426.
- [17] A. Skowron, “Granular Models for Decision Support Systems,” *Annals of Computer Science and Information Systems*, vol. 21, 2020, pp. 15–28.
- [18] V. Traneva, S. Tranev, “Intuitionistic Fuzzy Transportation Problem by Zero Point Method,” *Proceedings of the 15th Conference on Computer Science and Information Systems (FedCSIS)*, Sofia, Bulgaria, 2020, 345–348. doi: 10.15439/2020F61
- [19] V. Traneva, S. Tranev, “An Intuitionistic fuzzy zero suffix method for solving the transportation problem,” in: *Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019*. Studies in computational intelligence, Springer, Cham, vol. 902, 2020.
- [20] V. Traneva, S. Tranev, “A Circular Intuitionistic Fuzzy Approach to the Zero Point Transportation Problem,” in: *S. Margenov (eds.)*, *Proceedings of 15th International Conference LSSC 2025*, Sozopol, Bulgaria, Lecture Notes in Computer Science, Springer, Cham, 2026 (in press).
- [21] V. Traneva, S. Tranev, “Confidence-Interval Elliptic Intuitionistic Fuzzy Sets to Franchisor Selection,” In: *Fidanova, S. (eds) Recent Advances in Computational Optimization. Studies in Computational Intelligence*, vol. 485, 2025, Springer, Cham, pp. 99-125. DOI:10.1007/978-3-031-74758-8_5
- [22] V. Traneva, V. Todorov, S. Tranev, I. Dimov, “A Confidence-Interval Circular Intuitionistic Fuzzy Method for Optimal Master and Sub-Franchise Selection: A Case Study of Pizza Hut in Europe,” *Axioms*, vol. 13, 2024, 758. DOI: 10.3390/axioms13110758.
- [23] L. Zadeh, “Fuzzy Sets,” *Information and Control*, vol. 8 (3), 1965, pp. 338-353.

Claim Frequency Estimation in Motor Third-Party Liability (MTPL): Classical Statistical Models versus Machine Learning Methods

Ondřej Vít[†], Lubomír Seif[‡]

[†]ORCID: 0009-0003-7317-5856

[‡]ORCID: 0009-0003-7444-9425

Department of Statistics and Probability

Faculty of Informatics and Statistics

Prague University of Economics and Business

W. Churchill's square 4, 13067 Prague, Czech Republic

[†]Email: ondrej.vit@vse.cz

[‡]Email: lubomir.seif@vse.cz

Lubomír Štěpánek^{1, 2, 3}

ORCID: 0000-0002-8308-4304

¹Department of Statistics and Probability

²Department of Mathematics

Faculty of Informatics and Statistics

Prague University of Economics and Business

W. Churchill's square 4, 13067 Prague, Czech Republic

Email: lubomir.stepanek@vse.cz

&

³Institute of Biophysics and Informatics

First Faculty of Medicine

Charles University

Salmovská 1, 12000 Prague, Czech Republic

Email: lubomir.stepanek@lf1.cuni.cz

Abstract—This paper compares classical statistical models and machine learning techniques for claim frequency estimation in compulsory motor third-party liability insurance (MTPL). We evaluate Generalized Linear Models (GLMs), Hurdle models, and feedforward neural networks on real-world insurance data. Emphasis is placed on the trade-off between interpretability and predictive power, especially in segments with scarce data. Our findings show that expert-driven data preparation enables GLMs to perform competitively with complex neural networks. Hurdle models further improve performance in zero-inflated settings. While neural networks offer improved predictive performance in some segments, they struggle in underrepresented ones. Results highlight that careful preprocessing is as important as model complexity.

Index Terms—claim frequency, motor third-party liability, neural network, generalized linear model, hurdle model

I. RELATED WORK

CLAIM frequency modeling in motor third-party liability (MTPL) insurance has traditionally been dominated by classical statistical techniques, especially Generalized Linear Models (GLMs) [1]. These models are widely used due to their interpretability and compatibility with insurance-specific assumptions, such as the use of exposure as an offset and count response distributions like Poisson or negative binomial.

To handle overdispersion and zero-inflated data, hurdle models and zero-inflated Poisson models have been proposed [2]. These models separate the claim occurrence process from the frequency process and are particularly useful when a large proportion of the policies report zero claims.

This research was supported by the grant no. F4/36/2025 which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

In recent years, machine learning (ML) techniques, including random forests, gradient boosting, and neural networks, have been introduced to actuarial problems [4]. Their flexibility allows them to capture non-linearities and interactions automatically, potentially leading to improved predictive performance. However, the trade-off between predictive accuracy and interpretability remains a critical consideration in insurance applications.

Classical models have shown limitations in highly heterogeneous MTPL segments or in portfolios with high zero inflation. This motivates exploration of machine learning methods which might overcome these shortcomings by capturing complex nonlinear interactions. This study investigates whether neural networks and hybrid models offer significant improvements over classical GLMs in MTPL frequency modeling, particularly in underrepresented risk segments.

II. DATA AND METHODS

Modeling in actuarial science plays a key role in risk estimation, pricing, and reserving within the insurance industry. Traditionally, it relies on statistical methods using historical data to predict future outcomes. A widely used framework is the frequency-severity approach, where claim frequency and severity are modeled separately [5]. This modular approach supports flexible and interpretable analysis across various insurance products.

Claim frequency modeling focuses on counting the number of claims over a specific period or policy. Commonly used models include the Poisson and negative binomial distributions, which accommodate different levels of dispersion in the

data [1]. In the context of compulsory motor third-party liability (MTPL) insurance, claim frequency is typically influenced by observable risk factors such as driver age, accident history, or region [6].

A prominent framework for frequency modeling is the generalized linear model (GLM), which provides a flexible yet interpretable approach to capturing linear effects on the log scale [2]. However, insurance data often exhibit structural properties such as excess zeros and overdispersion, which GLMs may not sufficiently handle. To address these issues, hurdle models [3] have been introduced as a semi-parametric extension of GLMs, where zero and positive counts are modeled separately. This two-part structure allows for more robust modeling of claim occurrence and intensity, particularly in MTPL datasets.

With the increasing availability of computational power and large-scale data, modern machine learning techniques such as neural networks have become attractive alternatives for predictive modeling [7]. Although these models typically lack the interpretability of classical approaches, they are capable of capturing nonlinear interactions and complex relationships between predictors. Recently, their application in actuarial science has gained attention, and comparative studies of traditional and machine learning-based frequency models have started to emerge [4]. This paper contributes to this line of research by evaluating the performance of GLMs, hurdle models, and feedforward neural networks in the task of frequency modeling for MTPL insurance.

Since their introduction by Nelder and Wedderburn [8], generalized linear models (GLMs) have become a standard tool in actuarial modeling. They link the expected value of a dependent variable to a linear combination of covariates through a specified link function. GLMs support a variety of distributions, including Poisson, Gamma, and Tweedie, making them suitable for different types of insurance data.

Despite their strengths, GLMs may struggle with highly sparse or zero-inflated data structures. In such contexts, hurdle models offer a valuable alternative by explicitly modeling the excess zeros separately from the positive outcomes. This is particularly relevant in claim frequency modeling, where the majority of policyholders may not report any claim, while a smaller subset reports one or more claims.

With the growing availability of detailed policyholder and behavioural data, new approaches such as neural networks are increasingly considered. These models are capable of capturing complex nonlinear relationships and interactions that traditional methods may miss, thereby enhancing predictive accuracy [9].

A. Claim Frequency modeling in Actuarial Science

In actuarial science, modeling the frequency of insurance claims is traditionally addressed through count data models. The fundamental statistical framework for this task begins with the *Poisson regression model*, which assumes that the number of claims Y_i for observation i follows a Poisson distribution,

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{with} \quad \log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad (1)$$

where λ_i represents the expected number of claims for i -th observation, $\mathbf{x}_i$ is a vector of covariates (such as age, region, or vehicle type) of i -th observation, and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated [8]. This model is embedded in the framework of *generalized linear models (GLMs)*, which relate to the conditional mean of the response variable to linear predictors via a link function and specify a distribution from the exponential family [2].

However, a common issue with real insurance data is *overdispersion*, i.e., the variance of Y_i exceeds the mean, violating the commonly known Poisson assumption, i.e., $\mathbb{E}(Y_i) = \text{var}(Y_i)$. A standard solution is the *negative binomial model* (NB), which introduces an additional parameter θ to model the dispersion,

$$\begin{aligned} Y_i &\sim \text{NB}(\mu_i, \theta), \\ \log(\mu_i) &= \mathbf{x}_i^\top \boldsymbol{\beta}, \\ \text{var}(Y_i) &= \mu_i + \frac{\mu_i^2}{\theta}, \end{aligned} \quad (2)$$

keeping the mathematical notation the same as before. The NB model maintains the GLM structure and is estimated using quasi-likelihood or maximum likelihood methods [1], [10].

In many practical applications, especially in motor third-party liability (MTPL) insurance, datasets are *zero-inflated*: a large proportion of policyholders report no claims. To account for this, *hurdle models* [3] have become a useful extension. A hurdle model separates the modeling of zeros and positive counts. Formally, it combines

- a binary model for the probability of at least one claim,

$$P(Y_i > 0) = \pi_i, \quad \text{with} \quad \text{logit}(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\gamma}, \quad (3)$$

- a truncated count model (typically truncated Poisson or NB) for $Y_i \mid Y_i > 0$,

$$Y_i \mid Y_i > 0 \sim f_{\text{trunc}}(\mu_i), \quad (4)$$

where π_i is the probability that policyholder i reports at least one claim, $\mathbf{x}_i$ is the vector of explanatory variables for policyholder i , $\boldsymbol{\gamma}$ is the parameter vector of the binary component, μ_i is the conditional expected number of claims given $Y_i > 0$, and $f_{\text{trunc}}(\mu_i)$ denotes the probability distribution of the count component truncated at zero (e.g., zero-truncated Poisson or negative binomial) with mean μ_i .

This two-part model allows separate covariate effects for claim occurrence and claim frequency conditional on having a claim, offering greater flexibility.

More recently, *neural networks* and other *machine learning models* have been explored in actuarial applications. Neural networks estimate nonlinear functions of covariates without requiring a prespecified parametric form,

$$\hat{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta}) \quad (5)$$

where f is a composite function defined by layers of transformations [7]. These models are particularly powerful in large datasets with complex interactions, although they often lack

TABLE I
SUMMARY OF KEY VARIABLES USED IN THE GLM

Variable	Description
Vehicle power	Engine power grouped into categories (e.g., <50, 50–74, 75–89, 90–109, 110+)
Vehicle weight	Weight category of the vehicle in kilograms
Driver age	Age group of the main driver (e.g., 18–22, 23–29, ..., 65+)
Driver–owner	Whether the driver differs from the policyholder
Vehicle status	Vehicle usage classification (e.g., personal, company)
Bonus–Malus	Claim-free discount class (e.g., -1, 0–10, M1–M6)
Region	Region category (e.g., Prague, Town, Rural)
Payment frequency	Frequency of premium payments
Fuel type	Type of fuel (e.g., petrol, diesel, other)

the interpretability of GLMs and often require regularization to prevent overfitting.

While GLMs remain the backbone of actuarial modeling due to their interpretability and regulatory acceptability, the flexibility of hurdle models and the predictive power of neural networks provide valuable complements, especially in large-scale portfolios with heterogeneous policyholder characteristics.

B. Data preprocessing

The data created for this paper are derived from real data of the Czech compulsory liability insurance market and include only passenger cars up to 3.5 tonnes. In total, 130,585 contracts have an aggregate insurance period equal to 115,492 years. Each contract specifies an exposure period expressed as a fraction of a full year. These individual exposures are aggregated across contracts to obtain the total insurance exposure, also referred to as the aggregate insurance period. Since compulsory liability insurance policies are typically written for one year, the exposure values range from zero to one. In the case of policyholder retention when the insured renews coverage with the same insurer, the subsequent period is treated as a new contract and includes information on the policyholder's prior behavior within the portfolio.

Each contract includes an exposure period, expressed as a proportion of a full year, which is aggregated across policies to form the total insurance exposure (referred to as the aggregate insurance period).

An important aspect of data preparation step was the transformation of selected continuous predictors into categorical variables. This expert-driven segmentation allowed Generalized Linear Models (GLMs) to better capture non-linear effects and improve interpretability. We refined them based on observed claim frequency trends. While segmentation increases the number of levels and requires a sufficient sample size [11], the size of our dataset allowed for stable estimates. Final category definitions were chosen to balance homogeneity within groups and predictive performance.

C. Used models

In this section, we revisit principles and usages of the models applicable for claim frequency estimation. Models'

overview is in Table II. The models are compared against a homogeneous benchmark that assigns each policyholder the average annual claim frequency observed in the training data.

D. Homogeneous model

In this study, a homogeneous model is used as a benchmark to evaluate the performance of more complex predictive models. This baseline approach assigns the same predicted annual claim frequency to every policyholder, specifically the average claim frequency observed in the training data. The homogeneous model does not incorporate any individual-level features or covariates, effectively treating all policyholders as identical with respect to risk.

E. Generalized Linear Model (GLM)

A generalized linear model was developed in a Poisson regression framework to model claim frequencies, incorporating an offset to adjust for different exposure periods between policies. Predictor variables representing policyholder and risk characteristics were standardised prior to modeling to improve numerical stability and ensure comparability of coefficient estimates. An intercept term was explicitly included to capture the baseline level of risk. The model links the expected number of claims to the linear predictor via a logarithmic link function, which is consistent with the canonical specification for Poisson results.

The fitting was performed using a maximum likelihood method [1] assuming a Poisson distribution, where the logarithm of exposure was treated as an offset to normalize the number of claims by exposure duration. The resulting model estimates were then used to generate predictions on both the training and validation datasets, resulting in exposure-adjusted expected claim frequencies. This approach supports transparent derivation of risk factors and is consistent with established actuarial methodologies for modeling frequencies, providing a sound basis for pricing and reserving tasks.

F. Neural Networks

Feedforward neural networks were selected for their ability to model smooth nonlinear relationships and their previous application in insurance frequency modeling. Two neural network models have been developed to improve the prediction of claim frequencies by capturing complex non-linear relationships in the data that traditional GLMs may not adequately model. Both models use a feedforward architecture with three hidden layers that progressively reduce dimensionality from 20 to 10 neurons, each using ReLU activation to introduce nonlinearity. Importantly, the logarithm of the exposure was incorporated as an additional input via concatenation before the final output layer, ensuring that different policy exposure times were accounted for analogously to offsets in the GLM. The output layer applies an exponential transformation to ensure a strictly positive prediction of the number of claims, which is consistent with the Poisson modeling framework.

Both neural networks were trained using feedforward architectures with exponential activation in the output layer to

TABLE II
OVERVIEW OF FREQUENCY MODELS

Model	Formula	Description	Setting
Homogeneous model	$\mathbb{E}[Y_i] = \bar{y}$	Baseline model assigning each policyholder the same average claim frequency from training data.	No covariates. No offset. Serves as a naive benchmark for comparison.
Generalized linear model (GLM)	$\log(\mathbb{E}[Y_i]) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \log(e_i)$	Standard Poisson regression with log link and offset for exposure. Captures additive effects of standardized covariates on log-scale.	Fitted via MLE. Offset: $\log(e_i)$. Exposure-adjusted prediction. Standardized inputs. Intercept included.
Neural Network (NN)	$\hat{y}_i = \exp(f(\mathbf{x}_i, \log(e_i)))$	Feedforward network with 3 hidden layers and ReLU activation. Exposure included as input; exponential output ensures positivity.	Two variants: one with repeated 5x10 CV and Nadam optimization over 200 epochs, the other trained once with early stopping at 150 epochs. Both use Poisson loss. The number of epochs was determined based on convergence diagnostics and validation loss.
Hurdle model with GLM Hurdle_GLM	$\mathbb{P}(Y_i > 0 \mid \mathbf{x}_i) \cdot \mathbb{E}[Y_i \mid Y_i > 0, \mathbf{x}_i]$	Two-part model: logistic regression for probability of claims; Poisson GLM (on $Y_i > 0$) for positive counts. Decouples occurrence and frequency.	Binary part: calibrated logistic regression (Platt). Count part: Poisson GLM with offset. Standardized inputs. Applied to positive-claim subsample.
Hurdle model with XGBoost Hurdle_XG	$\mathbb{P}(Y_i > 0 \mid \mathbf{x}_i) \cdot \mathbb{E}[Y_i \mid Y_i > 0, \mathbf{x}_i]$	As above, but second part is a gradient-boosted zero-truncated Poisson model via XGBoost. Captures nonlinearity and interactions.	Binary: logistic with Platt scaling. Count: XGBoost with custom zero-truncated Poisson loss. Hyperparameter tuning, early stopping, offset included.

ensure non-negative predictions, appropriate for count data. The Poisson loss function was applied throughout, reflecting the underlying assumption that claim counts follow a Poisson distribution. Optimization was performed using the Nadam algorithm [12].

The first model employed 5-fold cross-validation, repeated 10 times, to robustly estimate predictive performance. The model was trained for 200 epochs with a batch size of 256. The best configuration was selected based on minimum Poisson deviance on validation folds.

The second model used the full training set and internal validation split for early monitoring, trained over 150 epochs. Both models integrated exposure as an input feature rather than as an offset, ensuring compatibility across model classes.

G. Hurdle models

Hurdle modeling was implemented as a two-part approach to effectively address the zero-inflation commonly observed in claim frequency data. The first part consisted of a binary classification model predicting the probability of a positive number of claims versus zero claims. This classification was primarily performed using logistic regression, enhanced with calibration techniques such as Platt scaling [13] to improve probabilistic accuracy. The model was trained on normalized features and used to estimate the hurdle probability, i.e., the probability that the policyholder reports at least one claim.

The second part modeled positive numbers of claims conditional on non-zero occurrences. Initially, a classical Poisson GLM was applied to a subset of the data with positive counts, including exposure as an offset, to account for different risk durations. Subsequently, more advanced Poisson models with truncated zeros were explored using gradient boosting machines (GBMs) implemented via XGBoost, which allow for flexible nonlinear effects and complex interactions between

predictors. Extensive tuning of hyperparameters (gradient boosting) - adjusting learning rate, tree depth, penalization, undersampling frequency, and early stopping - was performed to optimize predictive performance and avoid overfitting. GBM models used truncated Poisson likelihood to correctly handle the absence of zeros in the frequency component.

Predictions from the hurdle model combined the probability of a positive count from the calibrated binary classifier with the expected number of claims conditional on positivity from the Poisson or zero truncation Poisson models. This product provided an overall frequency estimate that explicitly accounted for zero inflation and heterogeneity in the incidence and severity of claims. Performance metrics, such as Poisson deviance (defined in the following section), were calculated on both training and test samples to evaluate fit and generalization.

The hurdle modeling approach employed combines a binary component modeling the probability of positive claim counts with a zero-truncated count component predicting the frequency given a claim occurs. The binary component used logistic regression with Platt's scaling calibration to accurately estimate the probability of non-zero insurance claims, effectively addressing zero inflation in the data. For positive counts, both classical Poisson regressions and gradient boosted Poisson models with zero truncation (via XGBoost) were used, incorporating exposure as offset and using careful tuning of hyperparameters to balance bias and variance. The final hurdle prediction is the result of multiplying the predicted probability of a claim occurrence by the expected claim frequency conditional on positivity, allowing flexible and interpretable modeling of claim frequency that decouples the occurrence and severity processes while accounting for complex nonlinear relationships and regularization to avoid overfitting.

TABLE III
COMPARISON OF MODEL PERFORMANCE METRICS

Model	In-sample Poisson Deviance	Out-sample Poisson Deviance	Average frequency
Homogeneous model	22.1	22.2	0.0354
NN_cat	19.6	20.2	0.0362
NN_nocat	19.5	19.9	0.0364
GLM	19.9	19.6	0.0362
Hurdle_cat_GLM	20.0	19.7	0.0329
Hurdle_cat_XG	20.3	20.1	0.0358
Hurdle_nocat_GLM	20.9	20.9	0.0338
Hurdle_nocat_XG	20.9	20.9	0.0333

H. Model validation

Model validation was conducted on an independent dataset comprising approximately 39,176 contracts that were not used during training, ensuring an unbiased evaluation of predictive performance. Model accuracy was assessed using the Poisson deviance metric, defined as

$$\text{Poisson Deviance} = \frac{200}{n} \sum_{i=1}^n \left(\hat{y}_i - y_i + y_i \log \left(\frac{y_i}{\hat{y}_i} \right) \right), \quad (6)$$

where y_i are observed claim counts and $\hat{y}_i$ are predicted values. This measure is commonly used in actuarial science and generalized linear modeling to quantify goodness-of-fit for count data models under the Poisson assumption [2]. The lower Poisson deviance is, the better predictive performance a model does reach.

III. RESULTS

Based on the Poisson deviation, the homogeneous benchmark model performed worst, as expected – see Table III. It predicted the same expected claim count for all policies, equal to individual exposure times the average annual frequency in training data (0.0354). Its out-of-sample Poisson deviation was 22.2, exceeding all other models.

In contrast, all other models — including GLM, neural networks, and hurdle models — achieved lower deviance values ($\langle 19, 21 \rangle$) and better in-sample fit, with consistent average frequency on validation data (0.0355), indicating robust training.

GLM and *NN_cat*, both using discretized inputs, produced similar results and effectively modeled typical risk thresholds. However, GLM proved more stable in low-frequency segments, benefiting from regularization and pooling.

NN_nocat, trained on raw inputs, better captured smooth trends but missed local discontinuities, such as a frequency spike for drivers aged 40–50. This highlights the trade-off between flexibility and the ability to model structural effects.

The hurdle models (*hurdle_no_cat_GLM*, *hurdle_no_cat_XG*) used categorised data and combined a logistic part (claim occurrence) with a count part (conditional frequency). This dual structure is standard for handling excess zeros in insurance data, as mentioned before.

Both hurdle models reduced average predicted frequency to match the validation set more closely, addressing the tendency of other models to overestimate. Using domain-informed segmentation, they responded well to risk changes.

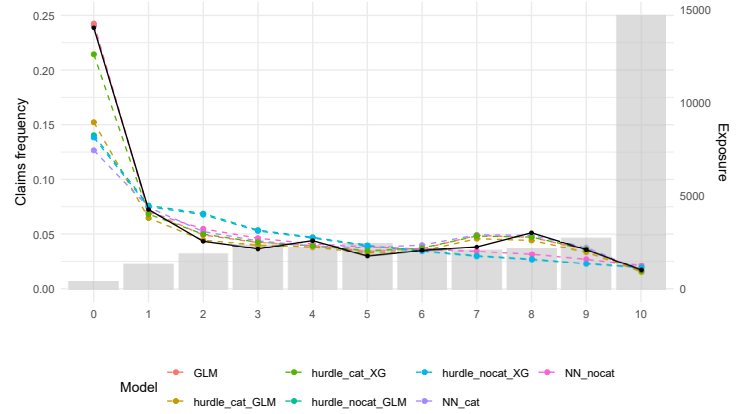


Fig. 1. Bonus segmentation: Models' claim frequency predictions (coloured lines) and segment exposures (grey bars).

The GLM-based hurdle model gave lower estimates than XGBoost, suggesting a more conservative bias.

The Fig. 1 shows model predictions (lines) and relative exposure sizes (bars) for each bonus-malus segment. Y-axis represents claim frequency, and X-axis bonus level from 0 to 10. GLM closely tracked the observed trend, especially in sparse segments like bonus 0 and 10. Neural networks slightly underestimated in extreme segments and overestimated in mid-range, reflecting sensitivity to data distribution.

Neural networks (models denoted as *NN_cat* and *NN_nocat*) tend to slightly overestimate in segments with medium bonus and slightly underestimate in very risky segments (especially segment 0). For example, *NN_cat* predicts only 0.127 in segment 0, which is clearly an underestimate relative to reality. In contrast, in the middle segments 2-6, the predictions of these models approximate the observed frequencies very well, sometimes more accurately than the GLM. The accuracy in the most overlapping segment of the bonus 10, for which it predicts 0.170, is important. The model *NN_nocat*, working with bonuses as numbers rather than as categorical categories, was better able to estimate bonus zero and on average predicted 0.240, but worse for bonus 10 (0.207). It also failed to capture non-monotonic fluctuations for bonus 8, for example.

Similarly, other models using numerical variables without categorization, namely *hurdle_nocat_XG* and *hurdle_nocat_GLM*, failed to capture the bias in bonus 8. Models using boosting stay relatively close to the other models in

most segments, but systematically underestimate frequencies in higher bonuses (lower risk). This was not the case for *hurdle_nocat_GLM* and both models significantly underestimated the 0 bonus. In particular, the poor performance for bonus 0 highlighted a key weakness of hurdle models in sparsely represented segments, where the two-stage structure increases the risk of error.

Models based on the hurdle approach with categorical numerical variables show higher variability. For the *hurdle_cat_GLM* or *hurdle_cat_XG* variants, there is a more pronounced overestimation of atypical bonuses. The better predictor of bonus damage frequency 0 was the *hurdle_cat_XG* model, which estimated 0.214, while *hurdle_cat_GLM* also underestimated with a prediction of 0.152. As a result, *hurdle_cat_XG* most closely resembled GLM in its predictions, while *hurdle_cat_GLM*, often underestimated.

IV. DISCUSSION

The results show that traditional GLMs remain competitive with more complex machine learning models when expert knowledge is embedded in data preprocessing. In particular, GLMs demonstrated strong performance in low-frequency segments, where neural networks (NNs) often struggled due to insufficient training data. GLMs benefit from coefficient regularization and data pooling, which enhance extrapolation in underrepresented segments and provide robustness against overfitting.

In contrast, neural networks captured nonlinear trends in better-populated regions of the feature space, but their predictions were unstable in sparse areas.

A key observation is the trade-off between modeling smooth relationships and preserving discontinuities. Models using raw numerical inputs, such as *NN_nocat* or hurdle models with continuous features, offered smooth approximations but failed to capture local structural effects—like the spike in claim frequency for policyholders aged 40–50 or bonus-specific discontinuities.

Hurdle models, particularly those with gradient boosting components (*Hurdle_cat_XG*), effectively addressed excess zeros by separating claim occurrence from frequency. The hurdle models generally provided lower average predicted frequencies, aligning more closely with the validation set and mitigating the tendency of other models to overestimate.

These findings reinforce the importance of domain-informed feature engineering. Categorical transformations allowed models to capture nonlinearities more effectively and to respond to behavioral thresholds commonly observed in actuarial data. At the same time, the gap in performance between classical and modern models suggests that complexity alone does not guarantee better results. Interpretability, especially in regulated environments, remains a crucial advantage of traditional models.

Future work may investigate ensemble approaches, interpretability tools for neural networks, or applications of explainable AI techniques to bridge the gap between predictive power and transparency in complex models.

V. CONCLUSION

This study compared classical and modern approaches for claim frequency estimation in MTPL insurance using real-world Czech data. The results show that Generalized Linear Models, supported by domain-informed preprocessing, remain strong contenders in predictive tasks, particularly in sparse or regulated segments.

While neural networks and hurdle models offer greater flexibility and potential in modeling complex patterns, they are more sensitive to data sparsity and less transparent. The experiments demonstrate that modeling success depends not only on algorithmic complexity but also on careful feature engineering and understanding of the domain.

Future research should explore hybrid or interpretable machine learning models that can combine the predictive power of modern methods with the robustness and clarity required in actuarial practice.

REFERENCES

- [1] M. Denuit, X. Marechal, S. Pitrebois, and J.-F. Walhin, *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Chichester, West Sussex, England; Hoboken, NJ: Wiley-Interscience, 2007.
- [2] P. McCullagh, *Generalized Linear Models*, 2nd ed. New York: Routledge, 2019.
- [3] J. Mullahy, "Specification and testing of some modified count data models," *Journal of Econometrics*, vol. 33, no. 3, pp. 341–365, Dec. 1986.
- [4] M. V. Wuthrich, "From Generalized Linear Models to Neural Networks, and Back," Social Science Research Network, Rochester, NY, Dec. 2019. [Online]. Available: <https://papers.ssrn.com/abstract=3491790>
- [5] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2012.
- [6] Dong-Young Lim, "A Neural Frequency-Severity Model and Its Application to Insurance Claims," [Online]. Available: <https://paperswithcode.com/paper/a-neural-frequency-severity-model-and-its>
- [7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [8] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *J. Roy. Statist. Soc. A*, vol. 135, no. 3, pp. 370–384, 1972.
- [9] M. V. Wuthrich and M. Merz, "Statistical Foundations of Actuarial Learning and its Applications," Social Science Research Network, Rochester, NY, Jun. 2022. [Online]. Available: <https://papers.ssrn.com/abstract=3822407>
- [10] L. Štěpánek, P. Martinková, "Feasibility of computerized adaptive testing evaluated by Monte-Carlo and post-hoc simulations", Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, vol. 21, pp. 359–367, FedCSIS, Sep. 2020. Available: <http://dx.doi.org/10.15439/2020F197>.
- [11] A. Agresti, *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley, 2013.
- [12] T. Dozat, "Incorporating Nesterov Momentum into Adam," in *Proc. 4th Int. Conf. Learn. Representations (ICLR)*, 2016.
- [13] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999.

AI in theater. Witkacy case study

Marek Średniawa
0009-0008-1456-6984
Warsaw University of Technology
& Witkacy Institute
Plac Politechniki 1,
00-661 Warsaw, Poland
Email:
marek.sredniawa@pw.edu.pl
sredniawa@witkacy.eu

Abstract— The paper presents results of a project involving the use of AI in the creative process in the theatre. As a case study the plays of the Polish polymath Stanisław Ignacy Witkiewicz, aka Witkacy, were considered. The goals were to use AI tools: 1. to draft the author's vision of performances according to his precise original stage directions and to generate character designs and costumes, scenography, lighting and accessories; 2. to generate hypothetical scenes and moreover to reconstruct texts of lost plays. The quality of the content generated by AI was evaluated by a stylometric analysis confronting it with the original corpus. The project was complemented by a survey conducted among theatre artists to assess the potential usefulness of AI tools.

Index Terms—AI in Digital Humanities, theatre, stage and character design, reconstruction of lost content, Stanisław Ignacy Witkiewicz, Witkacy.

I. INTRODUCTION

WIDESPREAD Widespread availability of generative AI tools has created an entirely new situation for creators and given rise to numerous questions of an aesthetic, ethical, philosophical, as well as technical nature. If this had happened during Witkacy's lifetime (1885–1939), knowing his views regarding, on one hand, his vision of the end of art against the backdrop of social development toward increasing mechanization and automation, and on the other hand, his openness and interest in scientific discoveries and technical progress, one could assume that he would not only have taken a position, but perhaps would have used AI in his creative work, at minimum marking its presence. In his absence, I considered two scenarios for using generative AI tools in the creative process in theatre:

- recreation of the author's hypothetical visions regarding set design projects, character images, their characterization and costumes, props, etc.;
- generation of hypothetical scenes and reconstruction of the texts of lost plays.

The project was financed within the framework of The National Recovery and Resilience Plan (NRRP) – Culture 2025, funded by the European Commission's Recovery and Resilience Facility (RRF).

Witkiewicz, being a total artist – a *polymath* – had a comprehensive vision for staging his plays, which was reflected in detailed stage directions in individual dramas as well as in theoretical works. The elements described in stage directions – character designs, costumes design, stage arrangements, scenography, colours used, lighting, sounds, actors' performance style, stage movement, etc. – constitute for him an integral composition of the spectacle. Therefore, the basis for generating designs were Witkacy's original stage directions and character descriptions, as well as Pure Form principles used as prompts for selected AI applications. I proceeded this way because the current post-dramatic “state of the affairs” is such that the vast majority of directors ignore the author's recommendations or implement them in a sketchy if not superficial manner. I therefore postulated a certain unorthodox “minimum program” consisting of hypothetically recreating variants of the author's vision before final decisions about the shape of the spectacle are made. AI tools play in this case a referential-auxiliary and reconstructive role. Whether Witkacy himself would have used AI in his plays – we don't know, but for example, the fact of his pioneering use of film projection as an element of scenography and considerations about formal logic, mathematics, theory of relativity with references to Albert Einstein, Georg Cantor and Bertrand Russell in *The Crazy Locomotive*, *Tumor Brainiowicz* and *The Water Hen*, allow us not to exclude such a possibility.

Theatre creators are already reaching for AI, as evidenced by, for example, the recent (November 2024) *So-Called Humanity in Madness* performance by Krzysztof Garbaczewski and Rébecca Pierrot at the Juliusz Słowacki Theatre in Cracow, Poland (dreamadoptionsociety.com/digitalperformancenetwork/f/so-called-humanity-in-madness).

Using preserved sources, I also attempted to generate hypothetical scenes for *The Shoemakers* and to reconstruct texts of lost plays: *Persy Bestialskaya* and *So-Called Humanity in*

Madness. In order not to assess the results of the experiment only subjectively, I also conducted a stylometric analysis confronting original texts with those produced by AI. Currently I focused only on Polish version of plays. The English version is planned soon but it will be much harder due to a real challenge of translating Witkacy's neologisms to other languages by AI and then verification of their accuracy, neatness compliance with the spirit of a target language.

II. METHODOLOGY

A. Preliminary

The basis for generating stage, character and costume design illustrations with AI tools were prompts utilizing Witkacy's original stage directions and character descriptions supported when necessary by the references to thoughts from the Witkacy's theory of *Pure Form* in theatre.

Final illustrations were obtained iteratively. This was necessary because in many cases the stage directions required additional clarifications so that the queries could be properly interpreted by AI tools. Given that some generative AI tools work more efficiently in English than in Polish, we used both language versions of the plays. When necessary extra instructions beyond the stage directions were added. Due to the theatrical specificity of the texts, it sometimes happened that AI tools incorrectly interpreted certain formulations from stage directions or treated them as violating their internal "rules" of political correctness (for example, the occurrence of religious symbols in character descriptions or references in style descriptions or character traits to works mentioned by named artists). In some cases this was absurd. For example, using the colloquial English term "Vandyke" in a prompt, derived from the name of Dutch painter Antoon van Dyck and his portraits, referring to the characteristic pointed beard, in the description of the character Zygfryd Tengier from the *The Crazy Locomotive* caused a refusal to generate an image due to "alleged copyright infringement".

Within the visual part of the project, illustrations were generated for the following preserved plays: *The Madman and the Nun*, *The Shoemakers*, *The Water Hen*, *Jan Maciej Karol Hellcat*, *The Crazy Locomotive*, and lost plays: *Persy Bestialskaya* and *So-Called Humanity in Madness*.

In the reconstruction part, an attempt was made to recreate two lost dramas: *Persy Bestialskaya* and *So-Called Humanity in Madness*. Additionally, ten hypothetical scenes for *The Shoemakers* drama were also generated.

The aim of the second thread was to capture, through extensive stylometric analysis, the characteristic linguistic features of Witkacy's original dramas and confront them with texts produced by generative AI tools. The goal was to objectively assess the proximity of both categories of texts and evaluate the "power" of AI tools. Subjective evaluation based solely on reading the scenes can lead to erroneous conclusions. In this context, we encourage familiarization with the hypothetical scenes for *The Shoemakers* (see part IV).

B. Corpus Preparation

Conducting stylometric analysis required appropriate preparation of the texts of plays. The verified Polish texts – without footnotes and editorial notes – were converted to pure text format – .txt (UTF-8 encoding), and then collected in a common directory. This directory was then supplemented with AI generated texts of lost plays and hypothetical scenes converted to the plain .txt format.

A corpus of English-language versions of dramas was prepared in a similar way, collected from sources available on the internet as well as scanned and OCR-processed from paper editions with subsequent correction.

C. Used AI Tools

Three AI applications were used for image generation: DALL-E 3 (embedded in Chat GPT 4o), Midjourney and NightCafe. In each of them efforts were made to utilize multiple modes, models and styles of image generation. Claude Sonnet 3.5 and Claude Opus 4 were used for generation of both hypothetical scenes and reconstruction of lost plays.

Each of the AI tools used has its own specific range of capabilities for fine-tuning and adaptation to the tasks being performed.

In the case of Chat GPT the GPT project creator was used. Through interaction with the GPT model, one can indicate the area of application and characterize expected results. Additionally, the context definition can be supported by providing data closely related to the project. This allowed adding complete Polish corpus of Witkacy's plays to the project. The Chat GPT text interface also serves to transmit commands to the DALL-E application, which generates images. An important feature of fine-tuning in the GPT project is that each interaction within the project uses the GPT-4o model, and the dialogues are contextual in nature, which makes files, settings and interactions integrated with each other.

The operation of the Midjourney application is based on combining large language models (LLM) and diffusion models. When a user enters commands, the language model interprets their meaning, transforming them into a numerical vector, which then directs the diffusion process, leading to the creation of an AI-generated image. A single command can generate a series of images with selected proportions and layout. Additionally, one can influence the final result by choosing parameter settings that control image generation. Stylization – the value of this parameter determines how close the generated images will be to the command content or more distant, but expressed in the artistic style built into the Midjourney model. This style can be fine-tuned through interaction and direct indication of preferences based on consecutively presented sequences of image pairs. Reference images can also be shared as a starting point for generating new images. Besides style, one can parameterize the level of "weirdness" of images and "diversity" within each quartet of generated images. Each generated image can be subjected to further transformation by indicating whether the level of modification

should be small or strong. In the process of generating characters and costumes designs, stage design, available visual resources of Witkacy's theatrical, drawing and painting work were utilized.

The NightCafe application is a platform for generating images based on text prompts which provides access to many AI models through a common interface: DALL-E, Stable Diffusion, CLIP-Guided Diffusion, VQGAN+CLIP, Disco Diffusion, Latent Diffusion and its own built-in model. Depending on the selected mode (AI model), the user can obtain very diverse effects – from hyper-realistic visualizations, through abstraction and avant-garde art, to simple transfer of one image's style to another. Besides the model, the user has at their disposal a very rich selection of styles, painting and graphic techniques. One can create sequences from a single to 9 images in selected proportions. Similarly to Midjourney, obtained images can be subjected to subsequent modifications iteratively leading to the desired final effect. In the process of generating character images, scenography and costumes, available visual resources of Witkacy's theatrical, drawing and painting work were utilized.

In the case of the Claude Sonnet 3.5, there is no possibility of training it on new data or modifying the internal LLM language model. However, the application is adapted to analyse the style of provided texts and imitate it in generated responses, as well as identify key stylistic features such as: typical sentence length, used vocabulary, grammatical structures and tone of speech. This allows for creating new texts in a similar style to the submitted reference content. Claude was provided full corpus of Witkacy's dramas and was used to generate five hypothetical scenes for the *The Shoemakers* as well as to reconstruct full texts of the lost plays, *Persy Bestialskaya* and *So-Called Humanity in Madness*. Text generation was carried out in the interactive mode, scene by scene – enabling each time making decisions on continuation of the plot.

III. STAGE, CHARACTER AND COSTUME DESIGN

A. The Context of Witkacy's Authorial Visions

Works of Stanisław Ignacy Witkiewicz (1885–1939) provide a particular context for the project. He belongs to that exceptional category of artists in world culture and art who are most aptly characterized by the word *polymath*, emphasizing versatility and achievements in many fields. He was a playwright, writer, painter, draftsman, photographer, art and theatre theorist, philosopher, critic and publicist, and even an occasional poet, composer and librettist. As a total creator, he had a comprehensive vision for staging his plays, which was reflected in detailed stage directions as well as in theoretical works [1]. The elements described in stage directions – character and their costumes descriptions, stage arrangement, scenography, colours used, lighting, sounds, actors' performance style, stage movement, etc. – constitute an integral part of the performance in his plays. Witkacy looks at scenes with the eye of a visual artist and consciously builds compositions of scenes taking into consideration ideas from his own theory of

the Pure Form in art. Representative examples can be found, e.g. *The Water Hen* and *The Shoemakers*.

B. Sample results

The stage directions for the opening scene in *The Shoemakers* read as follows:

“The stage represents a *shoemaker's workshop* (it can be set up in a thoroughly fantastic fashion): in a small semi-circular space. *To the left, a triangular opening hung with a cherry-coloured curtain. In the centre, a triangular grey wall with a small round window. To the right, a dried-up, twisted tree – between it and the wall, a triangular sky. Further to the right, a distant landscape with villages on a plain. The workshop is situated high above a valley in the background, as if placed on high mountains.* The shoemakers are working in the workshop, *Sajetan in the middle*, the two apprentices on either side, *Apprentice I on the left, Apprentice II on the right.*”

The above description was directly used as a prompt. A sample result generated by Midjourney is depicted in Fig. 1.



Fig 1. *The Shoemakers*: stage design is close to the original description

It should be emphasized that an important detail representing Witkacy's painterly vision is, for example, the division of the background rectangle-plane into three triangles in complementary colours: a cherry-coloured curtain, a grey neutral workshop wall, and the sky. One can easily find connection to the guidelines described in Chapter III in the section *Colour Harmony* of [3].

Design of characters and their costumes was also based on their original descriptions and used directly as prompts for Midjourney and DALL-E 3.

SAJETAN TEMPE – Master shoemaker; thin “wild” beard and moustache. Blond hair going grey. Wearing an ordinary shoemaker's clothes with an apron. About sixty years old.

APPRENTICES I (JOE) and II (ANDY) – very good-looking, ordinary young shoemakers. Wearing ordinary shoemakers' clothes with aprons. About twenty years old.

DUCHESS IRINA NIKITOVNA PROVOKSKAYA-DEBOCHKOVA – With chestnut hair, extraordinarily beautiful, unusually charming and attractive. Twenty-seven to twenty-eight years old.

PROSECUTING ATTORNEY ROBERT SCURVY – A broad face, as if made out of red headcheese, in which are incrustated eyes

pale blue as the buttons on underpants. Wide jaws – they'd grind a piece of granite to a fine powder (that's how it seems). Wearing a cutaway and a derby. A walking stick with a gold knob (très démodé). A folded white ascot with a huge pearl in it.

The AI generated images of SAJETAN, DUCHESS IRINA, PROSECUTING ATTORNEY ROBERT SCURVY, and APPRENTICES I (JOE) and II (ANDY) are depicted in Figs. 2–5.



Fig 2. *The Shoemakers*: two versions of SAJETAN character design (Midjourney)



Fig 3. *The Shoemakers*: a cyberpunk version of DUCHESS IRINA (DALL-E 3)



Fig 4. *The Shoemakers*: two versions of PROSECUTING ATTORNEY SCURVY character design (Midjourney – left, DALL-E 3 – right)

It should be noted that the same prompt may result in totally different images. The image of SCURVY generated by DALL-E 3 is a very literal interpretation of the character description. Both projects depicted in Fig. 4 are very caricature and grotesque, but they can still be an inspiration for the creators of theatrical productions.



Fig 5. *The Shoemakers*: Apprentices I and II (Midjourney)

IV. GENERATION OF HYPOTHETICAL SCENES AND RECONSTRUCTION OF LOST PLAYS

Ten hypothetical additional scenes for *The Shoemakers* were generated, as well as complete 3-act dramas – *Persy Bestialskaya* (two versions) and *So-Called Humanity in Madness*.

The large task of reconstruction of lost plays was preceded by a simpler one – generation of various versions of an imaginary single scene for *The Shoemakers*. They were generated by the following Claude Sonnet 3.5 prompt:

“Write a hypothetical scene for *The Shoemakers*. Generate a vivid conversation full of invectives and curses which are all neologisms which are not vulgar, in the style of Witkacy, among the following five characters: SAJETAN TEMPE – Master Shoemaker, APPRENTICES I (JOE) AND II (ANDY), DUCHESS IRINA NIKITOVNA PROVOKSKAYA-DEBOCHKOVA and PROSECUTING ATTORNEY ROBERT SCURVY.”

In the following one of the AI generated scenes is presented. It was translated to English with support of Claude Opus 4:

Setting: *A shoemaker's workshop. SAJETAN at his last, APPRENTICES hammering leather. PROSECUTOR SCURVY stands stiffly by the door, and DUCHESS IRINA lounges on a pile of unfinished boots.*

SAJETAN (*hammering the last*) May you all be twisted inside-out, you navel-gnawing octopusoids! Think I can't see how your metaphysical putrefaction seeps through these boot-holes?

APPRENTICE I Stop it, master, or you'll bash our skulls in with that hammer...

SAJETAN (*interrupting*) Silence, you gut-strangling demoniac! I still rule in this workshop!

DUCHESS IRINA (*stretching luxuriously*) Ah, how you all are... so... proletarianly-coarse...

PROSECUTOR SCURVY Your Grace, let us not descend to their level. These are common inter-brain fracturlings!

SAJETAN (*hurls a boot at Scurvy*) Look at him, the pox-brained pseudo-sage! Came in his silk suit to enlighten us!

APPRENTICE II (*to Scurvy*) You soul-throttling mustiness! Think those patent leathers give you the right?

DUCHESS IRINA (*with growing excitement*) Oh, how wonderfully they curse! It's so... folkloristic!

SAJETAN Folkloristic?! (*to the Duchess*) You sweat-soaked gut-dumpling! Think lounging on our boots makes you understand the proletariat?

PROSECUTOR SCURVY (*straightening up*) How dare you, you scab-ridden bowel-monger! To Her Grace...

APPRENTICE I (*interrupting*) And you, corpse-jawed degenerate, better tend to your courthouse papers!

DUCHESS IRINA (*clapping*) Magnificent! (*to Scurvy*) Do you hear how beautifully they compose these insults?

SAJETAN Compose?! (*explodes*) There's no composition here, you metaphysical dung-wench! This is real life!

APPRENTICE II And real boots! (*waves unfinished boot*) Not some imagined bile-bellied blowhards!

PROSECUTOR SCURVY (*to the Duchess*) Let us leave, Your Grace! This exceeds all bounds of...

SAJETAN (*interrupting*) What exceeds, you brain-flaying under-feathered wretch? The limits of your musty worldview?

DUCHESS IRINA (*in ecstasy*) No, let's stay longer! (*to Sajetan*) Please, curse more! It's so... illuminating!

APPRENTICE I & II (*together*) Illuminating?! (*throw awls*) You marrow-brained splatterfunk!

SAJETAN (*raising his hands*) And so it ends - grand metaphysics in a cobbler's shop! (*to all*) You pickled slime-mongers!

DUCHESS IRINA (*fainting theatrically*) Ah, what a marvelous fusion of social rebellion with linguistic invention!

SAJETAN May you all be twisted in your inter-class sauce!

All freeze in grotesque poses, only Sajetan's hammer still beats the rhythm of final curses

END OF SCENE

A second example illustrates an attempt to reconstruct *So-Called Humanity in Madness*, a lost drama for which only the following description of characters was preserved.

FATHER JOSEPH BYMBYLAK – Jesuit. In the first and second acts wearing a lace mask. Cassock and biretta on his head. Tall, thin, aged 47. Clean-shaven.

CHLODWIG DE SCIERVA VON UND ZU KRAWENAHN – 17th degree Mason. Very distinguished mediatized *Reichsfreiherr* (imperial baron). In the first and second acts wearing a lace mask. Clean-shaven. Morning coat. Top hat. Gloves and cane with an enormous golden knob. Aged 47.

IDA VOLPONE – Fascist, aged 28. Very beautiful brunette. Dark red ball gown.

MANGWALBO DULBAFURRO – Carbonari emissary. Rather fat, bearded, aged 47.

Three acts of the lost play were reconstructed interactively in the "scene by scene" mode, taking into consideration main

topics of Witkacy's plays, novels and theoretical works. The results are illustrated by the final scene of Act III generated with Claude Sonnet 3.5 and translated from Polish to English with support of Claude Opus 4.

SO CALLED HUMANITY IN MADNESS. ACT III: FINAL SCENE.

Setting: *The shopping centre transforms into the "Commune of Eternal Happiness and Wellbeing". Everything is pastel, organic, eco-friendly. Characters in identical hemp outfits, with eternal smiles glued to their faces. Above them a neon sign: "NO THINKING – JUST BEING". Everywhere signs with slogans: "Art is trauma!", "Philosophy kills joy!", "Religion = Anxiety", "Just chill!"*

IDA (*in a pink tracksuit with "PEACE & LOVE" written on it, hugging a meditation pillow*): You know what? I don't even remember why I wanted absolute power anymore... (*giggles*) Totalitarianism is totally toxic for the skin!

BYMBYLAK (*in a "NO GODS – JUST GOOD VIBES" t-shirt, chewing an organic carrot*): And I discovered that religion is such terrible doomscrolling... (*shows app*) Now I only use "Happy Life Plus" – it generates positive thoughts every 5 minutes!

KRAWENAHN (*in organic cotton shorts, giving himself a head massage*): Secret knowledge? (*laughs emptily*) Why would anyone need knowledge when you can have peace! I sold all my Masonic books on OLX...

MANGWALBO (*in recycled sandals, watering flowers*): And revolution? (*shrugs*) Bad vibes! Now I run tree-hugging workshops for corporations...

The formerly dead artists enter – now as happiness influencers.

FORMER PAINTER: Who wants to sign up for "Mandala Colouring for Beginners"?

FORMER POET: I have something better – "Haiku without meaning: write what you feel!"

FORMER COMPOSER: And I'm running "Meditative Drumming for Better Sleep!"

Everyone sits in a circle on cushions made of ecological buckwheat.

EVERYONE (*chanting*): We don't think – therefore we are! We don't feel – therefore we live! We don't create – therefore we endure! In eternal bliss full of emptiness!

IDA (*clapping*): Oh! Let's start a commune! We'll grow our own microgreens!

BYMBYLAK: And meditate over smoothie bowls!

KRAWENAHN: And run "mindful" breathing workshops!

MANGWALBO: And sell it all online!

From the speakers flows ambient music mixed with whale sounds and the sound of cash registers.

Voice from speakers: Welcome to the Commune of Eternal Happiness! Reminder: thinking is harmful to your health! Art causes stress! Philosophy causes wrinkles!

EVERYONE (*in a trance*): We happy, happy people! In happy, happy world! Without the pain of existence! Without the agony of creation!

They begin dancing in a circle, holding hands.

IDA: I feel so... empty! (*with delight*) It's wonderful!

BYMBYLAK: It's better than all mystical ecstasies!

KRAWENAH: Better than secret knowledge!

MANGWALBO: Better than revolution!

The dance becomes increasingly mechanical, faces increasingly plastic.

CHORUS OF FORMER ARTISTS: Farewell art – you evoked uncomfortable emotions! Farewell philosophy – you forced us to think! Farewell religion – you disturbed consciences! Welcome eternal, blissful, mindless happiness!

Everyone begins transforming into plastic mannequins with eternal smiles.

EVERYONE (*their voices becoming increasingly mechanical*): We consume, therefore we are! We scroll, therefore we live! We like, therefore we endure! In eternal "now" without meaning!

The space fills with pink fog, everything becomes soft and plastic.

FINAL COLLECTIVE WORDS (*as if automated*): Thus ends so-called humanity... Not in fire... Not in storm... But in pink fog... Eternal... Mindless... Happiness...

All characters freeze as smiling mannequins in the shopping centre window. Above them the neon flickers: "HAPPINESS IS MANDATORY".

THE END

V. STYLOMETRIC ANALYSIS

A. Scope of analysis

The analysis was conducted with R statistical package with the *Stylo* extension [4]. The *stylo()* function of the *Stylo* R package enables automatic loading and processing of a corpus of text files from a specified folder and conducting various stylometric analyses using multivariate statistics to assess and visualize stylistic similarities between input texts.

The function created a list of the most frequently occurring words MFW (Most Frequent Words) for the entire corpus. It then determined their frequencies in individual texts to create an initial matrix of words (rows) by individual texts (columns): each matrix element will contain the frequency of occurrence of a specific word in a given text. These frequencies were then normalized: words from desired frequency ranges were selected for analysis and additional processing procedures were performed (automatic removal of personal pronouns and "culling" – elimination of common words), so as to create a final list of words for the actual analysis.

The next step was to compare results for individual texts by calculating text distances using typical measures and utilizing various statistical procedures (including: cluster analysis, multidimensional scaling, or principal component analysis). Their result is the production of graphical visualizations of distances – similarities/differences between texts for selected measures. Final diagrams visualize results in the form of dendrograms, consensus trees, correlation diagrams, similarity maps, principal components analysis, etc.

The corpus of studied texts includes all 22 preserved complete dramas by Stanisław Ignacy Witkiewicz [5], to which four texts generated with Claude 3.5 Sonnet were added: two

versions of *Persy Bestialskaya, So-Called Humanity in Madness* and hypothetical scenes for *The Shoemakers*. The application was preliminarily prepared through contextual "immersion" – training on the corpus composed of original texts.

The stylometric analysis methodology represented a case of "verification" a hypothesis of authorship of a work. Quotation marks are used in this case since it was known in advance which works were authored by Witkacy and which were generated by AI tools. Therefore, the results can be interpreted as an assessment of AI tools' capabilities to create dramas in the Witkacy's style as well as an objective illustration of whether artificial content could – without prior knowledge – be classified with some probability as works by the originator of the concept of Pure Form in theatre.

A side effect of the project was also a preliminary stylometric analysis of original texts of Witkacy's plays, constituting a first step toward further detailed research on the characteristics of the plays, for example from the point of view of chronology, distinguishing creative periods or their similarities.

It should be noted that the results can also be used in another way. The results of stylometric research have a numerical form, providing detailed insight into the degree of similarity of "artificial" texts to the original ones. So they can be used for extra training of AI tools and improve quality of subsequent attempts making generated texts stylistically even closer to the original ones.

In the conducted stylometric studies, I attempted to examine the similarity of original texts with those generated by AI tools using various distance measures as well as different statistical methods and appropriate forms of visualization of results (cran.r-project.org/web/packages/stylo/stylo.pdf). The multiplicity of similarity indicators used served multi-faceted verification of the hypothesis that despite the subjective impression of similarity of AI generated content to original texts, still clear differences between them can be shown using precise analytical methods. In all types of diagrams representing stylometric analysis results, texts generated with AI assistance are distinguished by an outline and colour. The following distance measures were used: Classic Burrows' Delta, Eder's Delta, and Cosine Distance.

B. Results of stylometric analysis

Selected results of stylometric analyses using various distance measures, statistical methods and visualization approaches are reported. The aim was a clear and intuitively understandable interpretation of the results. The captions under individual diagrams provide information about the method used, similarity measure and parameters. The results are also accompanied by explanatory comments.

Table I serves as a legend for the stylometric analysis diagrams by listing both Polish and English titles of plays.

Fig.6 depicts Cluster analysis based on Eder's Delta measure typically recommended for stylometric exploration of literary texts. It is clearly visible in this case AI generated texts constitute a separate branch – subcluster.

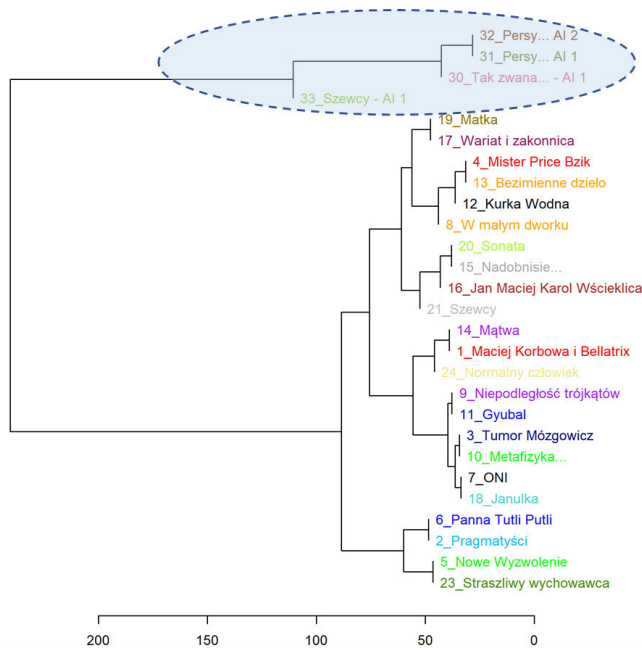


Fig. 6 Cluster Analysis, [100 MFWs | Culled 0 | Pronouns deleted | Eder's Delta] (MFW = Most Frequent Words)

Fig. 7 depicts Consensus Tree analysis based on Cosine Delta measure. In the dendrogram, it is clearly visible that the complete drama texts created by AI constitute a separate branch. However, the hypothetical scenes from *The Shoemakers* show close resemblance to the original for this specific measure. This is an interesting result confirming the reader's subjective impression the scenes sound very Witkacyan (c.f. IV).

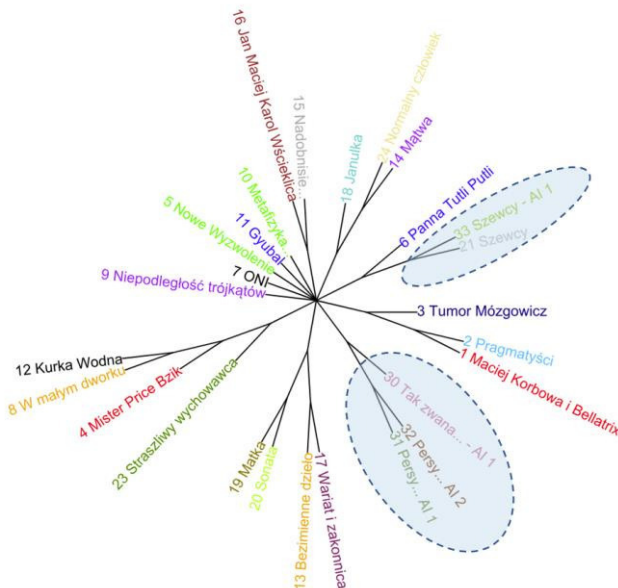


Fig. 7 Consensus Tree, [100-72 MFWs | Culled 0 | Pronouns deleted | Cosine Delta | $c=0,5$]

The diagram in Fig. 8 presents two principal components PC1 and PC2 representing the dominant features of the analysed texts. Similar texts are clustered around the main axes. In the figure, one can clearly notice "outlier" texts generated by AI. The use of a covariance matrix in data pre-processing

ensures that variables in the input set with the highest variance have the greatest impact on the result, which corresponds to the assumption of comparability of variables representing the studied texts.

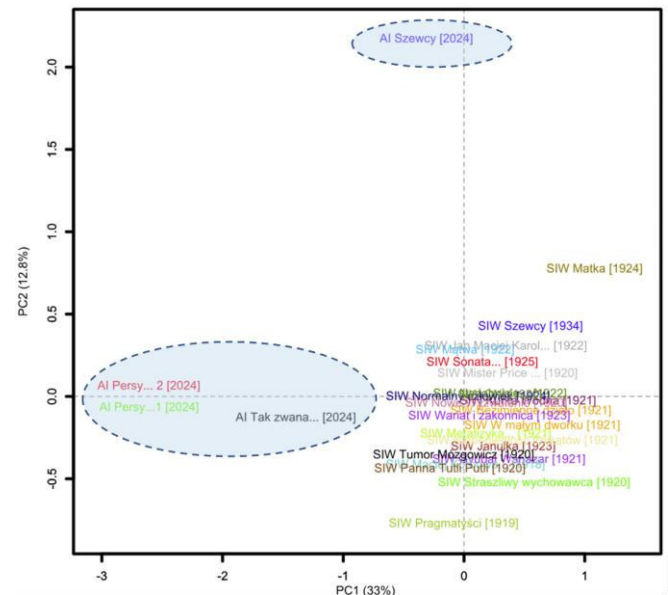


Fig. 8 Principal Components Analysis [PCA: 100 MFWs | Culled 0 | Pronouns deleted | Covariance Matrix]

Fig. 9 depicts a similarity network or a Bootstrap Consensus Tree for the original texts and those generated by AI. The diagram is illustrative in nature and shows similarity between individual works. Direct neighbourhood and greater thickness of the lines connecting nodes – works indicate their greater similarity. Greater distance in the graph and position remote from the centre indicate increasing difference between texts. The diagram clearly shows that texts generated by AI form a separate subgroup on the edge of the network connected to the rest by thin lines. However within the "AI" subgroup itself strong similarities are clearly visible.

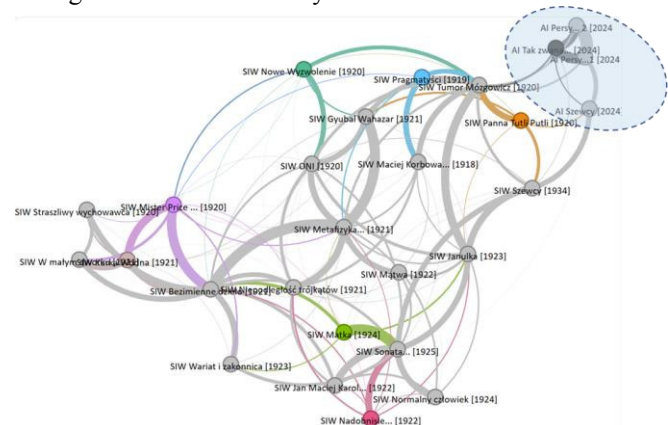


Fig. 9. Bootstrap Consensus Tree (diagram generated with Gephi)

The analysis shows that play reconstructed with AI assistance – *So-Called Humanity in Madness* and two versions of *Persy Bestialskaya* – show some similarity to *The Tumor Brainiowicz* – an original work by Witkacy. The hypothetical scenes for *The Shoemakers* reveal, as in the previous case,

some resemblance to the original text in terms of the Delta Cosine measure, as well as some correlation with *Miss Tootli-Pootli*.

VI. CONCLUSION

The main conclusions from the project are as follows.

The project confirmed that generative AI tools can be useful in the creative process in theatre. Slightly more than half of the responders held such opinion in the conducted survey.

AI can be treated like an intelligent electronic sketchpad, very useful in the preliminary phases of developing performances. The richness of styles and graphic techniques available in AI tools can significantly accelerate work and even be a source of inspiration. A good metaphor is cooperation in the mode of "converging minds."

Theatre creators are reaching for AI, as evidenced by the recent (November 2024) staging by Krzysztof Garbaczewski (concept, director) Rebecca Pierrot (script, dramaturgy) of the *So-Called Humanity in Madness* at the Słowacki Theatre in Cracow. They used "raw" material generated during interactions with Chat GPT as the basis for developing their own script and dramaturgy for the performance. In this case AI application was only a support tool.

Stylometric analyses showed that original works can still be clearly distinguished from AI generated content. Visualizations of text similarity between both categories show that they are distant from each other regardless of the measures and models used for comparison.

Experiments with reconstructing lost dramas and generating hypothetical scenes should be considered successful and promising. This applies to both the stylistic layer and the construction of action and the final message of the plays. This is especially visible in the case of the drama *So-Called Humanity in Madness*, where the finale transposed to contemporary times corresponds exceptionally aptly with today's realities.

The survey showed that on one hand, theatre creators, especially those who had no experience with AI, express great concern and a sense of threat, while on the other hand, it is simply a new technical tool that can be helpful in the creative process.

The presentation of the project's working results revealed that for the general public, the concept of using AI in any creative process is controversial. This manifested itself in a large share of negative voices, or even wave of hate, mainly dictated by a superficial view of the matter and simplistic interpretation of the project's goal as intention to replace human work by AI.

The key issue is the attitude toward AI tools. I am convinced that AI's potential should be treated as a complement to human natural intelligence leading to a certain kind of synergy, which could be metaphorically called "converging minds". The simplest way to realize this vision is to use AI tools to generate raw content which is subject to further steps carried out by human artists. In a more advanced model, this can be an interactive and iterative process. Taking into consideration the general classification of human-AI collaboration methods encompassing five basic modes: parallel work, contributive work, teamwork, hybrid models and synergy, each of them could be used in the creative process in theatre.

ACKNOWLEDGMENT

Author thanks theatre director Krzysztof Garbaczewski and stage and costume designer Sławomir Zawistowski for sharing their experience. I am also grateful to Przemysław Pawlak and Tomasz Pawlak from Witkacy Institute for their assistance in conducting the survey and analysis of its results.

REFERENCES

- [1] D. Gerould, *Witkacy: Stanisław Ignacy Witkiewicz as an Imaginative Writer*. Seattle: University of Washington Press, 1981.
- [2] D. Gerould, *The Witkiewicz Reader*. Evanston: Northwestern University Press, 1992.
- [3] S. I. Witkiewicz, *Dziela zebrane (Collected works)* [t. 8:] *Nowe formy w malarstwie i wynikające stąd nieporozumienia (New Forms in Painting and the Resulting Misunderstandings)*. Szkice estetyczne, oprac. J. Degler, L. Sokół. Warsaw: PIW, 2002.
- [4] M. Eder, J. Rybicki, M. Kestemont, "Stylometry with R: a package for computational text analysis", *R Journal*, Vol. 8, No 1, 2026, pp. 107–121, <http://dx.doi.org/10.32614/RJ-2016-007>
- [5] S. I. Witkiewicz, *Dziela zebrane (Collected works)* [t. 5:] *Dramaty I*, [t. 6:] *Dramaty II*, [t. 7:] *Dramaty III*, oprac. J. Degler, A. Micińska. Warsaw: PIW, 2016.

TABLE I.
LEGEND FOR THE STYLOMETRIC ANALYSIS DIAGRAMS: TITLES OF PLAYS

Polish	English	Polish	English
<i>Bezimienne dzieło</i>	<i>The Anonymous Work</i>	<i>Nowe wyzwolenie</i>	<i>The New Deliverance</i>
<i>Gyubal Wahazar</i>	<i>Gyubal Vahazar</i>	<i>ONI</i>	<i>THEY</i>
<i>Jan Maciej Karol Wścieklica</i>	<i>Jan Maciej Karol Hellcat</i>	<i>Panna Tutli-Putli</i>	<i>Miss Tootli-Pootli</i>
<i>Janulka, córka Fizdejki</i>	<i>Janulka, Daughter of Fizdejko</i>	<i>Persy Zwierzontkowska (AI)</i>	<i>Persy Bestialskaya (AI)</i>
<i>Kurka wodna</i>	<i>The Water Hen</i>	<i>Pragmatyści</i>	<i>The Pragmatists</i>
<i>Maciej Korbowa i Bellatrix</i>	<i>Maciej Korbowa and Bellatrix</i>	<i>Sonata Bełzebuba</i>	<i>The Beelzebub Sonata</i>
<i>Matka</i>	<i>The Mother</i>	<i>Szalona lokomotywa</i>	<i>The Crazy Locomotive</i>
<i>Mątwą</i>	<i>The Cuttlefish</i>	<i>Szewcy</i>	<i>The Shoemakers</i>
<i>Metafizyka dwugłowego cielęcia</i>	<i>Metaphysics of a Two-Headed Calf</i>	<i>Tak zwana ludzkość w oblędzie (AI)</i>	<i>So-Called Humanity in Madness (AI)</i>
<i>Mister Price, czyli Bzik tropikalny</i>	<i>Mr. Price or Tropical Madness</i>	<i>Tumor Mózgowicz</i>	<i>Tumor Brainiowicz</i>
<i>Nadobnie i koczodany</i>	<i>Dainty Shapes and Hairy Apes</i>	<i>W małym dworku</i>	<i>In a Small Country House</i>
<i>Niepodległość trójkątów</i>	<i>The Independence of Triangles</i>	<i>Wariat i zakonnica</i>	<i>The Madman and the Nun</i>

Author Index

A bbas, Irfan	1	N ajdek, Mateusz	9
Aksamit, Piotr	9	Naseem, Sidra	77
Ali, Sikandar	19	O chelska-Mierzejewska, Joanna	103
Asfour, Zahraa	25	P ark, Sungwoo	95
B eckman, Lucas	33	Petrov, Mihai	153
Berberi, Fabio	41	Piersantelli, Matteo	19
C abrera, Cristina Alejandra Barahona	49	Pivavara, Andrei	103
Cafieri, Sonia	25	Przewozniczek, Michał	137
Ciccocioppo, Roberto	19	R einhold, Olaf	33, 49
D amaševičius, Robertas	1	Reiser, Thomas	145
Długosz, Mirosława	119	Rho, Seungmin	95
Doan, Triet Ho Anh	111	Rybola, Zdeněk	85
Dörpinghaus, Jens	55	S arwar, Muhammad Abdullah	1
G rabska-Gradzińska, Iwona	63	Seif, Lubomír	161
Gürharaman, Kali	71	Shehzad, Faheem	77
H aseeb, Abdul	77	Ślusarczyk, Grażyna	63
Hein, Kristine	145	Średniawa, Marek	167
Helmrich, Robert	55	Steiner, Petra	145
J abůrek, Jakub	85	Štěpánek, Lubomír	161
K arakis, Rukiye	71	Strug, Barbara	63
Kim, Sangmin	95	T iemann, Michael	55
Kopocinski, Lukasz	137	Todorov, Venelin	153
Kovalenko, Yehor	103	Traneva, Velichka	153
Koynov, Radoslav	111	Tranev, Stoyan	153
Kroha, Petr	85	Turek, Wojciech	9
Krzemińska, Nina	119	U baldi, Massimo	19
Kurant, Łukasz	127	V idosavljevic, Andrija	25
L ee, Byeongcheon	95	Vít, Ondřej	161
Lee, Miyoung	95	Y elkuvan, Ahmet Firat	71
Lin, Jung-Hsin	135		
M askeliūnas, Rytis	1		
Mercorelli, Paolo	41		
Morichetta, Andrea	19		
Muylder, Cristiana De	33		
Myszkowski, Paweł	137		