



IARIA Congress 2026

The 2026 IARIA Annual Congress on Frontiers in Science, Technology, Services,
and Applications

ISBN: 978-1-68558-399-6

July 5 - 9, 2026

Nice, France

IARIA Congress 2026 Editors

Luigi Lavazza, Università dell'Insubria - Varese, Italy

Timothy T. Pham, Jet Propulsion Laboratory - California Institute of Technology,
USA

Lasse Berntzen, University of South-Eastern Norway, Norway

IARIA Congress 2026

Forward

The 2026 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications (IARIA Congress 2026), held between July 5-th, 2026 and July 9-th, 2026, in Nice, France, continued a series of international events keeping pace with the achievements and challenges our society is facing in science, technologies, services, and applications.

The annual event was a multidomain assembly of scientists, specialists, and decision makers from all economical, educational, and governmental entities, on Social Systems, Software, Data Science Analytics, Communications, Technology, and Networked Services. Apart from classical topics, the congress targeted frontier achievements on Knowledge Science, Data Science, Artificial Intelligence/Machine Learning (AI/ML)-based systems, Self-managing systems, Human-centric technologies, Advanced robotics, Virtual Worlds, Mobility, Sensing, Energy, Electric Vehicles, Green Energy, etc.

The IARIA Congress had a special scientific format where outstanding former IARIA scientists delivered dedicated speeches (Keynote speeches, Tutorial lectures) along with peer-reviewed contributions on the themes of achievements and challenges in science, technologies, services, and applications.

We take here the opportunity to warmly thank all the members of the IARIA Congress 2026 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank the authors who dedicated time and effort to contribute to IARIA Congress 2026. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the IARIA Congress 2026 organizing committee for their help in handling the logistics of this event.

We hope that IARIA Congress 2026 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in our society. We also hope that Nice, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

IARIA Congress 2026 Chairs

IARIA Congress 2026 Steering Committee

Carlos Becker Westphall, Federal University of Santa Catarina, Brazil

Luigi Lavazza, Università dell'Insubria – Varese, Italy

Timothy T. Pham, Jet Propulsion Laboratory - California Institute of Technology, USA

Lasse Berntzen, University of South-Eastern Norway, Norway

Arcady Zhukov, University of Basque Country (UPV/EHU), San Sebastian / Ikerbasque, Basque

Foundation for Science, Bilbao, Spain

Yasushi Kambayashi, Sanyo-Onoda City University, Japan

Gerhard Hube, Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

Renwei (Richard) Li, Southeast University, China

IARIA Congress 2026 Publicity Chairs

Sandra Viciano Tudela, Universitat Politècnica de Valencia, Spain

Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

José Miguel Jiménez, Universitat Politècnica de Valencia, Spain

IARIA Congress 2026 Committee

IARIA Congress 2026 Steering Committee

Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Luigi Lavazza, Università dell'Insubria – Varese, Italy
Timothy T. Pham, Jet Propulsion Laboratory - California Institute of Technology, USA
Lasse Berntzen, University of South-Eastern Norway, Norway
Arcady Zhukov, University of Basque Country (UPV/EHU), San Sebastian / Ikerbasque, Basque
Foundation for Science, Bilbao, Spain
Yasushi Kambayashi, Sanyo-Onoda City University, Japan
Gerhard Hube, Technical University of Applied Sciences Würzburg-Schweinfurt, Germany
Renwei (Richard) Li, Southeast University, China

IARIA Congress 2026 Publicity Chairs

Sandra Viciano Tudela, Universitat Politècnica de Valencia, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain
José Miguel Jiménez, Universitat Politècnica de Valencia, Spain

IARIA Congress 2026 Technical Program Committee

Tamer Abdou, Ryerson University, Canada
Nikunj Agarwal, Amazon, Inc., USA
Nitin Agarwal, COSMOS Research Center | University of Arkansas at Little Rock, USA
Petra Ahrweiler, Institute of Sociology | Johannes Gutenberg University Mainz, Germany
Abiola Akinnubi, Infinity Ward (Activision Publishing) Woodland Hill / COSMOS Research Center Little
Rock, Arkansas, USA
Sedat Akleylek, Ondokuz Mayıs University, Samsun, Turkey
Murat Akpınar, ASELSAN A.Ş., Ankara, Turkey
Raid Rafi Omar Al-Nima, Northern Technical University, Iraq
Alaa Alghazo, Hashemite University, Jordan
Hesham Ali, University of Nebraska Omaha , USA
Ali T. Alouani, Tennessee Technological University, USA
Mohammad Alsulami, University of Connecticut, USA
Mohd Ashraf Bin Ahmad, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia
Slimane Bah, Mohammadia Engineering School - University Mohammed V in Rabat, Morocco
Lasse Berntzen, University of South-Eastern Norway, Norway
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Megha Bhushan, University of Seville, Spain
John Blake, University of Aizu, Japan
Oleksandr Blazhko, National University «Odessa Polytechnic», Ukraine
Natalia Bogach, Peter the Great St. Petersburg Polytechnic University, Russia
Abdelmadjid Bouabdallah, University of Technology of Compiègne, France
Christian Bourret, Université Gustave Eiffel (Paris Est Marne-la-Vallée), France

Frederico Branco, Universidade de Trás-os-Montes e Alto Douro, Portugal
Uwe Breitenbücher, Reutlingen University, Germany
Antonio Brogi, University of Pisa, Italy
Erik Buchmann, Leipzig University / Data Science Center ScaDS.AI, Germany
Isaac Caicedo-Castro, University of Córdoba, Colombia
Ozgu Can, Ege University, Turkiye
Laurence Capus, Université Laval, Québec, Canada
Dirceu Cavendish, Kyushu Institute of Technology, Japan
Julio Cesar Duarte, Military Institute of Engineering, Brazil
Steve Chan, Decision Engineering Analysis Laboratory, USA
Haihua Chen, University of North Texas, USA
André Constantino da Silva, Federal Institute of São Paulo - IFSP, Brazil
Debasree Das, Indian Institute of Technology, Kharagpur, India
Jay Dave, BITS Pilani, Hyderabad Campus, India
Luca Davoli, University of Parma, Italy
Patrizio Dazzi, University of Pisa, Italy
Toon De Pessemier, Imec - WAVES - Ghent University, Belgium
Peter Edge, Ara Institute of Canterbury, New Zealand
Sam Erbatl, Duisburg-Essen University / Deutsche Telekom, Germany
Adrian Florea, "Lucian Blaga" University of Sibiu, Romania
Stefano Forti, University of Pisa, Italy
Matteo Francia, University of Bologna, Italy
Edelberto Franco Silva, Federal University of Juiz de Fora, Brazil
Ronnie Frates Rohrich, Federal University of Technology - Paraná, Brazil
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Imam Barkat Ghiloubi, University of Biskra, Algeria
Maxmillan Giyane, Midlands State University, Zimbabwe
Ramesh Gorantla, Arizona State University, USA
Gulsah Gorucu, University of the Aegean, Greece
Denis Gracanin, Virginia Tech, USA
Gregor Grambow, Aalen University, Germany
Wahida Handouzi, Tlemcen University, Algeria
Bohdan Havano, Lviv Polytechnic National University, Ukraine
Hussein Hazimeh, Lebanese University & Al Maaref University, Lebanon
Olaf Herden, Baden-Wuerttemberg Cooperative State University, Germany
Hans-Joachim Hof, Technische Hochschule Ingolstadt, Germany
Wladyslaw Homenda, Warsaw University of Technology, Poland
Shi-Jinn Horng, Asia University / National Taiwan University of Science and Technology, Taiwan
Gerhard Hube Technical University of Applied Sciences Würzburg-Schweinfurt, Germany
Hocine Imine, Université Gustave Eiffel, France
Orest Ivakhiv, L'viv Polytechnic National University, Ukraine
Fehmi Jaafar, Quebec University at Chicoutimi / Concordia University / Laval University, Canada
Marc Jansen, University of Applied Sciences Ruhr West, Germany
Imad Jawhar, Al Maaref University, Beirut, Lebanon
Felipe Jimenez Alonso, Technical University of Madrid, Spain
Luisa Jorge, CeDRI-IPB & INESC Coimbra, Portugal
Mohammed Jouhari, Mohammed VI Polytechnic University, Morocco
Yasushi Kambayashi, Sanyo-Onoda City University, Japan

Pushpendu Kar, The University of Nottingham, Ningbo, China
Mohamed Kara-Mohamed, Liverpool John Moores University, UK
Mojtaba A. Khanesar, University of Nottingham, UK
Elhaouari Kobzili, National Polytechnic School, Algeria
Koteswararao (Kote) Kondepudi, India Institute of Technology Dharwad (IITDh), India
Dmitry Korzun, Petrozavodsk State University (PetrSU), Russia
Constantine Kotropoulos, Aristotle University of Thessaloniki, Greece
Nane Kratzke, Lübeck University of Applied Sciences, Germany
Dragana Krstic, University of Nis, Serbia
Prarit Lamba, Intuit, USA
Bruno Lamiscarre, NeoMetSys, France
Filipe Lautert, UTFPR, Brazil
Luigi Lavazza, Università dell'Insubria, Varese, Italy
Wenjuan Li, Hong Kong Polytechnic University, China
Lan Lin, Ball State University, USA
Xing Liu, Kwantlen Polytechnic University, Canada
Prabhat Mahanti, University of New Brunswick, Canada
Abderrahim Maizate, RITM Laboratory / ESTC, Hassan II University of Casablanca (UH2C), Morocco
Rakesh Matam, Indian Institute of Information Technology Guwahati, India
Weizhi Meng, Lancaster University, UK
Fernando Moreira, Universidade Portucalense, Portugal
Shintaro Mori, Fukuoka University, Japan
Ioannis Moscholios, University of Peloponnese, Greece
Ejike Nwokoro, HealthNet Homecare, UK
Shashi Raj Pandey, Aalborg University, Denmark
Robert M. Patton, Oak Ridge National Laboratory, USA
Timothy Pham, Jet Propulsion Laboratory, USA
Krzysztof Pietroszek, American University, Washington, USA
Ivan Pires, Universidade de Trás-os-Montes e Alto Douro, Portugal
Chinthaka Premachandra, Shibaura Institute of Technology, Japan
Evgeny Pyshkin, University of Aizu, Japan
Ahmad Qawasmeh, The Hashemite University, Jordan
Md Muzakkir Quamar, King Fahd university of Petroleum and Minerals (KFUPM) / Interdisciplinary centre for smart mobility and logistics (IRC-SML), KSA
Catarina I. Reis, ciTechCare | School of Technology and Management | Polytechnic of Leiria, Portugal
Huamin Ren, Kristiania University College, Norway
Christophe Roche, University Savoie Mont-Blanc, France
Gunter Saake, Otto-von-Guericke-University Magdeburg, Germany
Zsolt Saffer, Institute of Statistics and Mathematical Methods in Economics - Vienna University of Technology, Austria
Sergei Sawitzki, FH Wedel (University of Applied Sciences), Germany
André Schneider de Oliveira, UTFPR/CPGEI/DAELN, Brazil
Hans-Werner Sehring, NORDAKADEMIE - University of Applied Sciences, Elmshorn, Germany
Ivana Semanjski, Universiteit Gent, Belgium
Davide Senatori, Università degli Studi di Genova, Italy
Shouqian Shi, Google, USA
Hemraj Singh, NIT Warangal, India
Miloudia Slaoui, Mohammed V University (UM5) in Rabat, Morocco

Yifei Song, Virginia Tech, Blacksburg, USA
Michael Spranger, Hochschule Mittweida | University of Applied Sciences, Germany
Grażyna Suchacka, Institute of Informatics / University of Opole, Poland
Weiwei Zhu Stone, University of Maryland Eastern Shore, USA
Ye Tang, Amazon, USA
Christos Troussas, University of West Attica, Greece
G. Vadivu, SRM Institute of Science & Technology, Kattankulathur, India
Jos van Rooyen, Huis voor software kwaliteit, The Netherlands
Eric MSP Veith, OFFIS e.V. - Institut für Informatik, Oldenburg, Germany
Tim vor der Brück, Fernfachhochschule Schweiz (FFHS), Switzerland
Zhengxun Wu, Independent researcher, USA
Linda Yang, University of Portsmouth, UK
Shuo Yang, University of Hong Kong, Hong Kong
Jiaqi Yin, Northwestern Polytechnical University, China
Maram Bani Younes, Philadelphia University, Jordan
Elena Zaitseva, University of Zilina, Slovakia
Jorge Zavaleta, CNPq, Rio de Janeiro, Brazil
Yuan Zhang, Nanjing University, P.R. China
Xingyu Zhou, Dow Inc., USA
Arkady Zhukov, University of Basque Country - UPV/EHU | IKERBASQUE - Basque Foundation for Science, Spain

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission to reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Few-Shot Classification Using the OpenAI CLIP Model <i>Hsiao-Ju Lin, Shi-Jinn Horng, Pin-Siang Huang, and En-Ping Lu</i>	1
Permutation Entropy and Statistical Complexity Analysis of MODIS Satellite NDVI Data for Discerning <i>Xylella fastidiosa</i> Infection in Olive Groves in Southern Italy <i>Luciano Telesca and Rosa Lasaponara</i>	7
Proposition of a Brain-Computer Interface based on Pictograms for Communication Purpose <i>Alvaro Fernandez-Rodriguez, Simon Audren, Lou Guiochet, Veronique Lespinet-Najib, Jean-Marc Andre, Francisco Velasco-Alvarez, and Ricardo Ron-Angevin</i>	14
Green Energy Transformation: Policy Reversals, Geopolitical Shocks, and the Digital Infrastructure Imperative <i>Mahmood Ahmad and Arifa Younas</i>	21
Driving Digital Transformation in Precision Engineering: A Tooling Management Case Study <i>Shivam Sharma, Mohamed Kara-Mohamed, Andrew Naylor, Phillip Taylor, and David Eadie</i>	27
Cascaded TSDF for Multi-Camera Volumetric Capture with Perceptual Mesh Reduction <i>Abdul Rehman and Francois Bouille</i>	34
Simplification Modeling System by Projecting Paradigmatic to Syntagmatic Simplification for Explaining Artistic Texts <i>Ryotaro Kamimura</i>	41
Model and Analysis of 5G Network Registration Procedure in Public Safety Scenarios <i>Christian Madsen, Michael Berger, and Henrik Wessing</i>	48
Evolutionary Transfer Optimization for Diversified Investment Using Predictive Distribution for Asset Returns <i>Kiyoharu Tagawa and Yukiko Orito</i>	54
LiDAR-Based Simultaneous Localization, Mapping, and Multiobject Tracking: Combining Semantic Segmentation and Interacting Multiple Model- Extended Object Tracking <i>Akio Okumura, Masafumi Hashimoto, and Kazuhiko Takahashi</i>	60
Majority Voting-Based Update of Point-Cloud Map Using LiDAR Information from Multiple Vehicles <i>Saki Fujisawa, Masafumi Hashimoto, and Kazuhiko Takahashi</i>	67
Evaluating LLM-based Pseudonymization for Cross-Platform Identifiers <i>Yeong Su Lee, Hendrik Bothe, and Michaela Geierhos</i>	73
Agreement-based SVM One-Class Classification: an Ensemble Method for Software Defect Prediction	80

Luigi Lavazza, Sandro Morasca, and Gabriele Rotoloni

Process Simulation and AI: A Perfect Match?! 86
Carlo Simon and Merlin Tankred Philipp Hladik

Assessing the Accessibility of Public Healthcare Web Services in Spain 92
Veronica Alvarez-Lopez and Daniel Fernandez-Lanvin

From Relational and Property Graph Data to Large Language Models 96
Malcolm Crowe and Fritz Laux

PyWIB: A Python Library for a Multi-Modal Approach to Web Interaction Behavior Analysis 103
Guillermo Dylan Carvajal-Aza, Alejandro Alvarez-Varela, Javier De Andres, Martin Gonzalez-Rodriguez, Daniel Fernandez-Lanvin, and Mercedes Paino

Engagement-Aware Response Generation for Educational Robots Using Grounded LLM Prompting 109
Nayanika Reddy Rajoli, Vennela Akula, and Daniela Marghitu

Parkinson's Disease Severity Assessment on the Unified Parkinson's Disease Rating Scale Using Text Embeddings Generated from Speech 114
Christina Koumpogianni and Constantine Kotropoulos

Articulating Multidisciplinary Knowledge Bundles for Urban Planning 120
Robert Laurini

Proactive Link Failure Early Warning for Ultra-Long-Range Maritime 5G via Edge Intelligence and Random Forest 126
Kun Yang, Yu Xu, and Li Qin

Few-Shot Classification Using the OpenAI CLIP Model

Hsiao-Ju Lin

Department of Orthopedic
Surgery,
Taoyuan General Hospital,
Ministry of Health and Welfare
Taoyuan, Taiwan
email: ruru1300837@gmail.com

Shi-Jinn Horng

Department of Computer Science
and Information Engineering,
Asia University
Taichung, Taiwan
email: horngsj@yahoo.com.tw

Pin-Siang Huang

Agaruda Research,
Agaruda,
Taipei, Taiwan
email: sean.huang@agaruda.io

En-Ping Lu

Department of Computer Science
and Information Engineering,
National Taiwan University of
Science and Technology,
Taipei, Taiwan
email: t6233111@gmail.com

Abstract—With the rapid development of artificial intelligence, the application of deep learning models has matured, but it requires large datasets to improve accuracy and generalization. To reduce reliance on large datasets, this study focuses on few-shot learning based on OpenAI's Contrastive Language-Image Pre-Training (CLIP) model and adapter technology. Unlike traditional transfer learning, this method does not require fine-tuning the model but instead combines newly learned image features with existing model features through adapters. The results demonstrate that the proposed method achieved excellent accuracy across multiple validation sets, particularly on the Describable Textures Dataset (DTD), improving from 67.53% to 75.51%, thereby reducing the need for training resources.

Keywords—*Few-shot learning; Zero-shot learning; Deep learning; Machine learning.*

I. INTRODUCTION

Artificial intelligence has rapidly advanced in the fields of vision, audio, and natural language. However, training practical models usually requires large amounts of data, leading to challenges with insufficient datasets. Even with powerful models, a lack of comprehensive datasets can limit their effectiveness. Additionally, fine-tuning large models requires substantial computational resources and can lead to overfitting. With the development of large models like CLIP [1], Segment Anything Model (SAM) [2], and Vision Transformer (ViT) [5] models with zero-shot learning capabilities can address many downstream tasks. This paper explores the technique of combining the CLIP model, used as the backbone network, with adapter neural networks for fine-tuning to achieve few-shot classification. This approach reduces data requirements and training resources, avoids overfitting, and enhances cost-effectiveness in enterprise applications.

The remainder of this paper is organized as follows. Section II provides a review of related work. Section III describes the proposed method. Section IV presents the experimental results. Section V concludes the paper.

II. RELATED WORK

A. CLIP

CLIP is a multimodal model developed by OpenAI. This model is pre-trained on a large-scale dataset of image-text pairs using a contrastive learning method [9], enabling it to process both textual and visual data simultaneously. Its main advantage lies in its ability to learn general representations for both images and text, allowing it to excel in a wide variety of tasks. The primary strength of the CLIP model is its powerful zero-shot

learning capability, which enables efficient classification and retrieval even in the absence of labeled data. Additionally, CLIP's multimodal nature allows it to establish connections between images and text, making it highly valuable for application scenarios that require handling multiple data forms simultaneously. Typical applications include image classification, image-text retrieval, and image generation.

B. Adapter

A lightweight neural network structure is typically used to add new learning capabilities to pre-trained models. The original backbone network is frozen, and only the external network is trained, eliminating the need to retrain the entire model. This approach allows for the preservation of the pre-trained model's knowledge while enabling efficient learning with limited data. Adapters usually consist of a small fully connected layer or MultiLayer Perceptron (MLP) placed between specific layers of the pre-trained model. These adapter modules receive inputs from the pre-trained model during the forward propagation process, generate new features after the adapter's computation, and combine these features with the original ones. This way, the adapter can learn the feature representations needed for new tasks without altering the original features of the pre-trained model. Another important application of adapters is reducing the risk of overfitting. Since the number of parameters in adapters is relatively small, they effectively mitigate overfitting issues on small datasets. This technique also significantly reduces the model's training time and resource requirements, making it particularly suitable for resource-constrained applications. Initially applied in Natural Language Processing (NLP) [11] domain, it has also been proven effective in the computer vision field [10], such as the SAM adapter [12].

C. Embedding learning

In few-shot classification tasks, one of the methods is embedding learning [13], where some samples x from the training dataset are passed through a neural network $f(x)$ for feature extraction and converted into a vector. This process can be seen as transforming an image from a high-dimensional space to a low-dimensional space, where the similarity between classes is easier to discern. Additionally, another network $g()$ can be trained to function similarly to $f()$, converting high-dimensional images into low-dimensional vectors. The classification is then performed by comparing the similarity between this vector and the class vector.

D. Transformer

The Transformer [7] is a neural network architecture based on self-attention mechanisms, introduced in 2017 and widely adopted in sequence modeling tasks. In the field of NLP [6], the introduction of the Transformer model has brought significant advancements. Traditionally, Long Short-Term Memory networks (LSTM) [4] and Recurrent Neural Networks (RNN) [3] were widely used for processing sequential data. However, these models have limitations when dealing with long-range dependencies. Additionally, autoregressive models have important applications in text generation, and the Transformer has improved upon these foundations. The self-attention mechanism of the Transformer allows for parallel processing of entire sequences, overcoming the serial computation limitations of RNNs, thereby significantly improving computational efficiency and performance. The application range of Transformers is extensive, including various NLP tasks such as machine translation, text generation [24], semantic analysis [25], text summarization [26], and question-answering systems. Furthermore, in the field of computer vision, following ResNet [8], the Transformer has also demonstrated strong capabilities. The ViT applies the Transformer's self-attention mechanism to tasks, such as image classification [5], object detection [27], and image segmentation [28], achieving remarkable results. These applications showcase the flexibility and superiority of the Transformer model in handling different data types and application scenarios.

E. Tip-Adapter [21]

The traditional CLIP-Adapter [20] integrates new knowledge with CLIP's pre-trained knowledge by fine-tuning a lightweight residual feature [8] adapter connected to the pre-trained CLIP model. This architecture allows the model to combine newly learned features with the existing visual representations of CLIP. Specifically, when an image is input, CLIP's visual encoder first extracts its features, which are then passed through the lightweight adapter to generate updated feature representations. These updated features are fused with the original CLIP features via residual connections, thus effectively combining new knowledge with pre-trained knowledge, significantly improving the performance of few-shot classification. Tip-Adapter is a training-free variant of CLIP-Adapter designed to enhance the performance of the CLIP model in few-shot classification tasks by using image cache. The principle involves constructing an image cache with a small number of images, storing the feature vectors of these images along with their corresponding labels in the cache. During inference, the model does not require additional parameter updates but instead makes classification decisions by retrieving features from the cache. The advantage of this method lies in its ability to quickly adapt to new tasks by leveraging the pre-trained knowledge of the CLIP model and performing lightweight feature adjustments with a small amount of few-shot data.

III. METHOD

A. Basical Setup

This study was conducted using Python 3.11.6 and PyTorch 2.1.2 for model development, and training was performed on an

NVIDIA GeForce GTX 1080 Ti GPU. All experiments were conducted on the Ubuntu 20.04.1 LTS operating system.

B. Data Augmentation

During the adapter training, a series of data augmentation techniques were introduced to preprocess the dataset, including cropping, rotation, adding various types of noise, etc. These methods are common and effective data augmentation techniques that enhance the model's generalization ability and improve its robustness. After experimentation, the method finally chosen in this paper was to increase image contrast, as described in (1):

$$I'(x, y) = \alpha(I(x, y) - \mu) + \mu \quad (1)$$

In this paper, we use the following variables to describe the process of adjusting image contrast. Let $I(x, y)$ represent the brightness value of the original image at position (x, y) , and $I'(x, y)$ represent the brightness value of the image after contrast adjustment at position (x, y) . Here, α is the contrast adjustment coefficient. When $\alpha > 1$, the contrast increases; μ is the mean brightness value, as shown in (2):

$$\mu = \frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N I(x, y) \quad (2)$$

μ is used to maintain the overall brightness of the image, where M and N represent the width and height of the image, respectively. In this paper, α is set to 2, and the enhanced image is shown in Figure 1. Contrast enhancement is a simple and effective data augmentation method. Many studies have demonstrated that it can significantly improve the model's ability to recognize low-contrast images.

In addition, Gaussian noise was applied for data augmentation. This method perturbs images with zero-mean Gaussian noise of standard deviation σ , as formulated in (3).

$$I'(x, y) = I(x, y) + N(0, \sigma^2) \quad (3)$$

Here, $N(0, \sigma^2)$ represents Gaussian noise with a mean of 0 and a variance of σ^2 . In this paper, σ^2 is set to a random value between 25 and 75. By adding Gaussian noise to the images, we can simulate more real-world scenarios, further enhancing the model's generalization ability and robustness.



Figure 1. Comparison of Contrast Enhancement: Left (Original data from Oxfordflower 17 [14]) and Right (Enhanced in this paper).



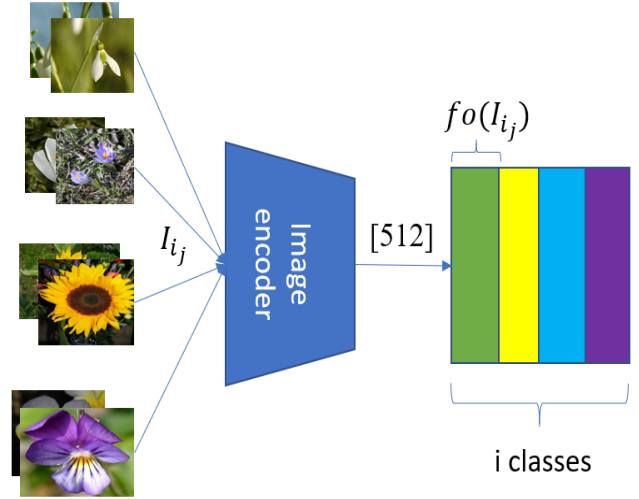
Figure 2. Effect of Gaussian Noise Augmentation, Original Data from Oxfordflower 17 [14].

The addition of Gaussian noise is illustrated in Figure 2. The augmentation effect of Gaussian noise on the image dataset has also been widely studied and proven. We will increase the number of samples in the training dataset to twice the original amount.

C. Image cache

Assume the training dataset has N classes. Let I_{ij} represent a randomly selected image from the i^{th} class in the training dataset, with i serving as the key for that image. The selected image I_{ij} is then input into our CLIP image encoder, which outputs a 512-dimensional vector $fo(I_{ij})$. Using the output vector $fo(I_{ij})$, we construct an image cache according to the classes, as shown in Figure 3. After feature extraction, the training images are stored as key-value pairs in a database. By applying the t-SNE [19] dimensionality reduction technique, we can visualize the distribution of different classes in the space, as illustrated in Figure 4, using 17 classes of images from Oxford Flower 102 to build the image cache. This is the starting point of this paper: using these distributions for classification. In our experimental design, we observe from the TIP-Adapter that the accuracy improves as the number of images increases. Therefore, we directly select 16 images per class to construct this cache. If there are N classes, the cache can be represented as a matrix of size $[16 \times N, 512]$. Cache models have been applied to improve some issues in NLP field, such as Knn-LMs [23], which also demonstrate the power of cache models. This approach allows for the initial rapid categorization of targets, further enhancing the adaptability and flexibility of the model.

There are several advantages to using this method within our model. Since the dataset contains a limited number of samples, a small amount of data can be used to construct this cache. Directly using a complex neural network could easily lead to overfitting. Therefore, we use the vectors generated by the CLIP image encoder, which has strong generalization capabilities. Similar vectors tend to cluster in closer regions within the space. By leveraging this characteristic, we avoid overfitting, reduce the impact of outliers, and decrease the number of learnable parameters, thereby increasing training speed.



[224x224x3]

Figure 3. Image Cache Example Diagram.

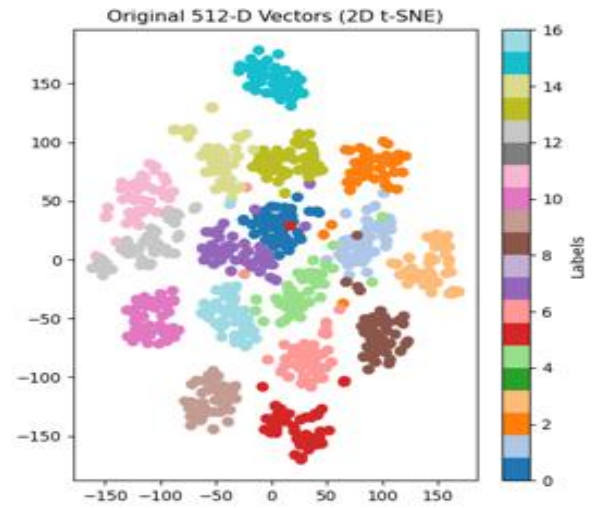


Figure 4. t-SNE Visualization of Oxford17 Data.

D. Image adapter

To make the CLIP model more effective in few-shot classification, this study introduces an adapter architecture as an external framework for the model. Adapters were initially used for fine-tuning in the natural language processing (NLP) domain and have since been widely applied across various deep learning fields. Unlike traditional deep learning neural networks, this method significantly reduces the number of training parameters and training time. It preserves the complexity of the original network while avoiding overfitting, which can occur when the amount of data is too small. The adapter achieves model fine-tuning by inserting lightweight parameter layers into the pre-trained model, without requiring extensive parameter updates across the entire network. In this paper, the adapter consists of a two-layer fully connected MLP with the activation function ReLU. The vectors f_o , obtained by

passing images through the image encoder, are used as training data to train the MLP. The parameters of the MLP are 512×128 and 128×512 , respectively, which can be represented as shown in (4):

$$f_x = W_2 \text{ReLU}(W_1 f_v + b_1) + b_2. \quad (4)$$

W_1 and W_2 are the weight matrices from the input layer to the hidden layer and from the hidden layer to the output layer, respectively. b_1 and b_2 are the bias vectors. We only train the adapter, freezing all the parameters of the image encoder, which reduces our training parameters to just those of the fully connected layers. During the training phase, our architecture is illustrated in Figure 5, and the following sections will introduce the process in sequence.

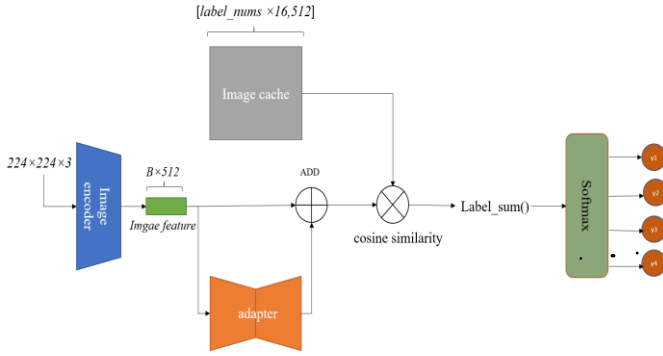


Figure 5. The proposed system architecture.

During the training phase, through the image cache, we construct the cache from other data in the training dataset, which is divided into two paths. One path passes through the MLP adapter for training, while the other remains unprocessed. The vector f_x output by the adapter is then added to the vector f_o , which bypassed the adapter, as shown in (5):

$$f_c = f_x + f_o \quad (5)$$

The vector f_c is compared with the image cache database using dot product similarity. The similarities for the same category are summed to obtain the classification score for that category, as shown in (6):

$$i \in \text{label}, y_i = \sum f_c \cdot \text{image_cache}_i \quad (6)$$

The similarity for each category obtained by this method is then passed through a Softmax function to produce the classification result for the image, as shown in (7):

$$\hat{y}_i = \frac{e^{y_i}}{\sum_j e^{y_j}} \quad (7)$$

Compared to traditional single-point key-value lookups, our approach can better avoid the occurrence of extreme values within the image cache, thereby reducing classification errors. Our loss function is the cross-entropy loss function, as shown in (8):

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^K y_i^{(n)} \log(\hat{y}_i^{(n)}) \quad (8)$$

where N represents the total number of samples, K represents the number of classes, $y_i^{(n)}$ denotes the true label of the n^{th} sample, and $\hat{y}_i^{(n)}$ represents the predicted probability of the n^{th} sample after applying the Softmax function.

IV. EXPERIMENTS

A. Experiment result

To validate the effectiveness of our method, we conducted a series of experiments. First, we trained different adapters on multiple datasets, including Flowers102 [29], Caltech-101 [16], OxfordPets [18], FGVCAircraft [17], and DTD [15]. We randomly selected a fixed number (16) of images from each class in the dataset to construct the image cache. The remaining data was then augmented and used for training, where each class has at most 40 images. Next, we trained the model. The pre-trained weights for the CLIP image encoder were ViT-B/32, which is the most commonly used weight. During training, we used the Adam optimizer and cross-entropy loss function, with an initial learning rate set at 0.001. To better adapt to changes during the training process, we reduced the learning rate to 80% of its original value every 10 epochs, for a total of 30 epochs. The batch size was set to 32. We kept the output vector dimension of the CLIP image encoder fixed, and our model achieved good accuracy on various few-shot test sets, as shown in Table I, outperforming other benchmark models in this field. These results demonstrate the excellent classification ability of our method and prove its effectiveness. In the following sections, we will provide a more detailed analysis of our experimental results.

To ensure stable performance of the model, we conducted tests and validations on several datasets and compared the results with previously proposed models, as shown. We imported the ViT-B/32 weights into the CLIP model and tested the model's performance according to the experimental setup described in Section IV.A. We also compared it with CLIP-Adapter and Tip-Adapter, both of which are benchmark models that use CLIP for classification tasks.

TABLE I. ACCURACY OF CLASSIFICATION MODELS ON DIFFERENT DATASETS

Method/ Dataset	CLIP	CLIP adapter[20]	Tip adapter[21]	Ours
Oxfordflower 102 [14]	66.35%	93.90%	94.80%	95.14%
OxfordPet [18]	85.98%	87.84%	89.70%	92.84%
DTD [15]	42.27%	65.96%	67.53%	75.51%
Caltech-101 [16]	86.01%	92.49%	92.86%	94.53%
FGVCA [17]	17.89%	32.10%	35.55%	39.85%

B. Efficacy

Since CLIP uses the similarity between text and images, we selected the category with the highest similarity as the predicted result. However, similar textual phrases can also affect

accuracy. For example, in the Oxford Flower 102 dataset, which contains 102 types of flowers, phrases like "a type of," "a type of flower on the ground," or "the picture shows a type of flower" could lead to variations in accuracy. Therefore, we standardized the input format, comparing the similarity using the English phrase "This is a photo of [specific flower]." For instance, for a sunflower, the input would be "This is a photo of sunflower."

According to our experimental data, all datasets showed a significant improvement compared to the original CLIP. Our model also showed a notable improvement on the DTD texture dataset, with accuracy increasing from 67.53% to 75.51%, outperforming Tip-Adapter. The other datasets also showed slight improvements. These results demonstrate the effectiveness and robustness of our model across these datasets.

C. Training parameters

In the field of few-shot classification, the ability to apply the model in scenarios such as data engineering has undeniable value, as it allows classification with a small amount of data. In data engineering, obtaining the first-hand manually labeled data often requires a significant amount of time. If we can reduce training time, we can save considerable time in the initial stages. As shown in Table II, only the corresponding adapters need to be trained, so there is less concern about the equipment being unable to handle the training. Additionally, with only 30 epochs per dataset, the training time can be compressed into just a few hours, significantly reducing the overall development process.

D. Ablation experiment

In this section, we conduct an ablation study to gain a deeper understanding of the contributions of various parts of the model, including the generalization ability of the original CLIP and the impact of the image cache combined with adapters. As shown in the results in Table III, there is a significant difference in accuracy across different datasets when using or not using the adapter strategy. Adding the adapter clearly demonstrates that this approach can indeed improve the model's accuracy. When only using CLIP, we still need to calculate the similarity with text vectors, but it is sometimes challenging to fully capture the features of an image with a textual description, particularly in datasets like DTD textures, where the accuracy is not high. Therefore, we also designed a method to compare each input image with every vector in the image cache, using the label of the most similar vector as the output. The results show that this paper effectively leverages CLIP's image encoder, making it more accurate in classification.

TABLE II. TRAINING PARAMETERS OF COMMON FEW-SHOT MODELS

Method	Parameters
ResNet-12	9.5630 M
EfficientNet-B0 [22]	5.3305M
Ours	0.13171M

TABLE III. ACCURACY ON DIFFERENT TEST SETS UNDER VARIOUS CONDITIONS

	one point	Image cache sum	Image cache sum+adapter
DTD [15]	51.00%	56.87%	75.51%
FGVCA [17]	25.55%	27.52%	47.88%
Caltech-101 [16]	87.25%	84.80%	94.53%
OxfordPet [18]	62.52 %	68.87 %	92.84%
Oxfordflower102 [29]	64.55%	87.32%	95.14%

V. CONCLUSION

This study explored the few-shot learning capabilities of large models for classification tasks by designing a model based on CLIP's image encoder with adapters. Experimental results across multiple datasets demonstrate the effectiveness of the proposed method. Owing to the CLIP backbone, our model exhibits strong generalization ability. Although frozen parameters are still required during inference, the model maintains relatively low parameter and computational costs, making it feasible for edge deployment. Despite promising results, further improvements remain possible.

While our results are promising, several improvements remain possible:

1. Dataset Diversification: Future studies could use more obscure datasets to better validate generalization, as CLIP may have encountered commonly used datasets.
2. Enhanced Data Augmentation: Beyond contrast and Gaussian noise, more effective augmentation methods could be explored.
3. Architecture Enhancement: Since only CLIP's image encoder was used, future work could leverage its text encoder or explore alternative architectures to improve performance and generalization.

ACKNOWLEDGEMENT

This work was supported in part by the National Science and Technology Council under contract number 114-2221-E-468 -014 -.

REFERENCES

- [1] A. Radford et al., "Learning transferable visual models from natural language supervision," *International conference on machine learning*, 2021, pp. 8748–8763.
- [2] A. Kirillov et al., "Segment anything," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [3] L. R. Medsker et al., "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64–67, p. 2, 2001.
- [4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [6] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

- [7] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 2006, vol. 2, pp. 1735–1742.
- [10] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.
- [11] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," *International conference on machine learning*, 2019, pp. 2790–2799.
- [12] J. Wu et al., "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.
- [13] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2593–2601.
- [14] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 2006, vol. 2, pp. 1447–1454.
- [15] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.
- [16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *2004 conference on computer vision and pattern recognition workshop*, 2004, pp. 178–178.
- [17] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [18] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," *2012 IEEE conference on computer vision and pattern recognition*, 2012, pp. 3498–3505.
- [19] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [20] P. Gao et al., "CLIP-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [21] R. Zhang et al., "Tip-adapter: Training-free CLIP-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.
- [22] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *International conference on machine learning*, 2019, pp. 6105–6114.
- [23] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019.
- [24] T. Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [26] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," *International conference on machine learning*, 2020, pp. 11328–11339.
- [27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *European conference on computer vision*, 2020, pp. 213–229.
- [28] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.
- [29] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008, pp. 722–729.

Permutation Entropy and Statistical Complexity Analysis of MODIS Satellite NDVI

Data for Discerning *Xylella fastidiosa* Infection in Olive Groves in Southern Italy

Luciano Telesca and Rosa Lasaponara
 Institute of Methodologies for Environmental Analysis
 National Research Council
 Tito, Italy
 email: luciano.telesca@cnr.it; rosa.lasaponara@cnr.it

Abstract—This study utilizes the Complexity–Entropy Causality Plane (CECP) to investigate the temporal dynamics of Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) time series, focusing on olive groves in Southern Italy. The primary objective is to identify the presence of *Xylella fastidiosa*, an aggressive phytopathogen responsible for the devastating Olive Quick Decline Syndrome (OQDS). Our results demonstrate that the application of the CECP framework facilitates a highly effective discrimination between infected and healthy olive orchards. Our analysis reveals that infected datasets are characterized by a lower permutation entropy (H) and larger statistical complexity (C) than uninfected ones. By employing an embedding dimension of $d_x = 4$ a wider separation between infected and uninfected datasets in the CECP is obtained, evidenced also by an increased Mahalanobis distance and superior Receiver Operating Characteristic (ROC) classification performance. These findings validate the CECP as an optimal diagnostic metric for identifying subtle irregularities within MODIS NDVI data due to *Xylella fastidiosa*-induced degradation of vegetation.

Index Terms—MODIS; NDVI; CECP; *Xylella fastidiosa*

I. INTRODUCTION

Invasive pests and non-native plant bacteria are increasingly recognized as major threats worldwide, capable of causing severe plant diseases with far-reaching effects on both natural ecosystems and agricultural production. A prominent example is *Xylella fastidiosa*, considered one of the most hazardous plant pathogens globally. This bacterium is responsible for serious diseases that particularly affect crops such as grapevines, citrus, and olive trees. Within the European Union, focusing specifically on olives, it is estimated that *Xylella fastidiosa* could lead to annual production losses of up to 5.5 billion euros, jeopardizing approximately 70% of the EU's mature olive tree output. Consequently, early detection of infections caused by this pathogen is essential to implement effective strategies for minimizing its economic and ecological impact [2].

In 2013, *Xylella fastidiosa* was first detected in southeastern Italy [3]. Since then, the pathogen has spread rapidly to several other European countries, leading to a serious phytosanitary crisis [4].

Traditionally, visual inspection has been the primary method for detecting *Xylella fastidiosa*, due to its simplicity and

low cost. However, the accuracy of this approach is highly dependent on the subjective assessment of disease severity.

In recent years, Remote Sensing (RS) techniques have gained increasing attention for monitoring vegetation dynamics, including changes caused by plant pathogens [5]. Several RS-based studies in phytopathology have focused on methods leveraging multi-temporal and multi-spectral satellite imagery to track land cover changes. Statistical approaches, such as principal component analysis [6] and curve-fitting techniques [7], have proven effective in identifying vegetation alterations across large areas.

The potential of MODIS satellite data for monitoring pest and parasite outbreaks at both landscape and field scales has been explored in [8], [9], where MODIS vegetation indices proved effective in tracking the degradation of pine tree vegetation caused by the parasite *Toumeyella parvicornis*. Additionally, analyses using multifractal detrended fluctuation analysis and the Fisher-Shannon informational method applied to MODIS evapotranspiration time series have demonstrated that these satellite-derived data provide a reliable means of assessing the pathogenic effects of *Xylella fastidiosa* on olive trees [10]–[12].

The Normalized Difference Vegetation Index (NDVI) is a well-established indicator of vegetation vigor, photosynthetic activity, and canopy structure, and is particularly suitable for monitoring the effects of *Xylella fastidiosa* infection in olive groves. The pathogen disrupts xylem water transport, inducing water stress, leaf desiccation, and progressive canopy loss, which result in reduced near-infrared reflectance and increased red reflectance, effectively captured by NDVI. Satellite-based NDVI time series, especially those derived from the MODIS sensor, provide consistent, long-term observations with high temporal resolution, enabling large-scale monitoring of olive-growing areas. In infected groves, NDVI temporal profiles typically exhibit reduced seasonal amplitude, altered phenological patterns, and increased temporal irregularity compared to healthy areas. When combined with nonlinear time-series analysis and statistical complexity measures, MODIS NDVI can reveal early dynamic changes associated with infection, making NDVI a robust, scalable, and cost-effective tool for the surveillance and monitoring of *Xylella fastidiosa* impacts.

In this study, we analyze time series of MODIS NDVI

pixels covering both infected and healthy olive groves in Southern Italy using the Complexity–Entropy Causality Plane (CECP) that enables a topological characterization of time series through the distribution of ordinal patterns, providing a robust alternative to traditional information-theoretic and fluctuation-based approaches previously applied to this type of data.

The Complexity–Entropy Causality Plane (CECP) was first introduced by Rosso *et al.* in 2007 [13] and has since become a widely used diagnostic tool. The CECP represents a two-dimensional space combining normalized permutation entropy (H) with a statistical complexity measure (C). While H measures the degree of randomness or unpredictability in a time series, C captures non-trivial correlations and structural patterns. By projecting time series onto the $H \times C$ plane, it is possible to differentiate between various stochastic and deterministic processes, allowing the identification of transitions between chaotic and noisy regimes [16]. The CECP has been successfully applied in environmental science, including the analysis of air pollution dynamics [17], river discharge variability under anthropogenic influence [19], and rainfall time series [18].

In Section II, the applied methods of time series analysis will be described and in Section III the investigated dataset will be presented. Section IV will show the obtained results that will be discussed in Section V.

II. METHODS

A. The permutation entropy and statistical complexity

Permutation entropy is computed by determining the Shannon entropy of a probability distribution derived from ordinal (or permutation) patterns. These patterns are extracted by applying a sliding window across a time series to generate symbolic representations. This specific distribution, often termed the *ordinal distribution*, is estimated following the symbolization framework proposed by Bandt and Pompe [1].

To formalize this methodology, consider a discrete time series $\{x_t\}_{t=1,\dots,N_x}$. The sequence is decomposed into $n_x = N_x - (d_x - 1)\tau_x$ overlapping vectors. Each vector consists of $d_x > 1$ elements sampled at a temporal distance of $\tau_x \geq 1$. For a given set of parameters, the p -th partition is defined as:

$$w_p = (x_p, x_{p+\tau_x}, x_{p+2\tau_x}, \dots, x_{p+(d_x-1)\tau_x}), \quad (1)$$

where $p = 1, \dots, n_x$ represents the partition index. The Bandt–Pompe approach relies on two fundamental parameters: the *embedding dimension* d_x and the *embedding delay* τ_x . While the seminal work by Bandt and Pompe [1] focused on consecutive observations ($\tau_x = 1$), the framework was subsequently generalized to include arbitrary delays. Because the selection of d_x and τ_x is critical to the analysis, extensive research has been dedicated to identifying optimal parameter configurations for different signal types [13] [16].

Following the partitioning of the signal, each vector w_p is mapped to a specific permutation $\pi_p = (r_0, r_1, \dots, r_{d_x-1})$. This permutation represents the unique ordering of the set $\{0, 1, \dots, d_x - 1\}$ required to arrange the components of w_p

in non-decreasing order. Formally, the ordinal pattern π_p is established by the relation:

$$x_{p+r_0\tau_x} \leq x_{p+r_1\tau_x} \leq \dots \leq x_{p+r_{d_x-1}\tau_x}. \quad (2)$$

To ensure a unique symbolic representation in the presence of ties (equal values), the initial chronological sequence is maintained. Specifically, if $x_{p+r_k\tau_x} = x_{p+r_h\tau_x}$, the condition $r_k < r_h$ is imposed for any $k < h$ [1].

To demonstrate this mapping, let us examine a sample sequence $x_t = (6, 4, 3, 3, 8, 10)$ with parameters $d_x = 4$ and $\tau_x = 1$. The initial window is defined as $w_1 = (6, 4, 3, 3)$. Sorting these values gives the sequence $3 \leq 3 < 4 < 6$, which links back to the original positions $x_{1+2}, x_{1+3}, x_{1+1}, x_{1+0}$. Consequently, the resulting ordinal pattern for this segment is $\pi_1 = (2, 3, 1, 0)$.

Although Bandt and Pompe originally mentioned that ties could be resolved by introducing a negligible amount of observational noise to the dataset, this approach is rarely adopted in empirical studies. While high-resolution data typically minimize the frequency of identical values, the choice of tie-breaking strategy becomes a critical consideration when processing low-resolution or heavily quantized signals [16].

After converting all partitions into their respective symbolic representations, we obtain a symbolic sequence $\{\pi_p\}_{p=1,\dots,n_x}$. The corresponding ordinal probability distribution, $P = \{\rho_i(\pi_i)\}_{i=1,\dots,n_\pi}$, is determined by calculating the relative frequency of each unique permutation within the sequence. Formally, this is expressed as:

$$\rho_i(\pi_i) = \frac{\text{count}(\pi_p = \pi_i)}{n_x}, \quad (3)$$

where π_i represents one of the $n_\pi = d_x!$ possible permutations of order d_x .

Using the distribution P , we calculate its Shannon entropy, defining the *permutation entropy* as:

$$SH(P) = - \sum_{i=1}^{n_\pi} \rho_i(\pi_i) \log_2 \rho_i(\pi_i). \quad (4)$$

This metric serves as a proxy for the structural stochasticity of the time series. Specifically, $SH \approx \log_2 n_\pi$ implies a nearly random process, whereas $SH \approx 0$ reflects highly predictable or deterministic behavior.

It is well-established that the estimator in (4) exhibits a negative bias, particularly for finite datasets [20], [22]. To mitigate this, we employ the Miller–Madow correction [21], which refines the estimate as follows:

$$S(P) = SH(P) + \frac{N_0 - 1}{2N}, \quad (5)$$

where N_0 denotes the count of permutations with non-zero probability and N represents the total number of observed patterns.

Finally, to facilitate comparisons across different embedding dimensions, we utilize the *normalized permutation entropy*:

$$H(P) = \frac{S(P)}{\log_2 n_\pi}, \quad (6)$$

which maps the entropy values onto the unit interval $[0, 1]$.

The embedding dimension d_x defines the symbolic resolution of the system, with the total number of potential ordinal patterns given by $n_\pi = d_x!$. To achieve a statistically consistent estimation of the ordinal probability distribution, Bandt and Pompe [1] suggest selecting d_x such that $d_x! \ll N_x$. Simultaneously, the embedding delay τ_x determines the relevant time scale for the dynamical analysis. In our case $\tau_x = 1$.

Rosso *et al.* [13] introduced the CECP as a diagnostic tool, designed to differentiate between different processes.

The CECP integrates the normalized permutation entropy H from (6) with the statistical complexity measure C to establish a bivariate diagnostic space, where C is analyzed as a function of H . Indicating as $P = \{\rho_i(\pi_i)\}_{i=1, \dots, n_\pi}$ the observed ordinal distribution and as $U = \{1/n_\pi\}_{i=1, \dots, n_\pi}$ the uniform distribution,

$$C(P) = \frac{D(P, U) H(P)}{D_{\max}}, \quad (7)$$

where the Jensen–Shannon divergence $D(P, U)$ is calculated as:

$$D(P, U) = S\left(\frac{P+U}{2}\right) - \frac{1}{2}S(P) - \frac{1}{2}S(U). \quad (8)$$

The term $D_{\max}$ serves as a normalization constant, ensuring that the complexity measure is bounded. It represents the maximum possible divergence between P and U , which occurs when the system is concentrated in a single state (i.e., P follows a Kronecker delta distribution, $\delta_{1,i}$). This constant is defined as:

$$D_{\max} = -\frac{1}{2} \left[(n_\pi + 1) \log_2(n_\pi + 1) - 2 \log_2(2n_\pi) + \log_2 n_\pi \right]. \quad (9)$$

Unlike entropy, C vanishes at both poles of the dynamical spectrum: complete order (periodic or constant signals) and total disorder (white noise). By quantifying the interplay between information storage and structural dissimilarity, C provides essential diagnostic details that complement the randomness quantified by H , making the CECP a powerful tool for detecting subtle regime shifts in complex systems.

B. The ROC analysis

Receiver Operating Characteristic (ROC) analysis serves as a fundamental framework for assessing the diagnostic efficacy of binary classifiers [24]. In this context, instances are categorized into two discrete classes, typically labeled as positive or negative. The predictive performance is determined by the resulting confusion matrix, which consists of four possible outcomes: True Positive (TP), where a positive instance is accurately identified; False Negative (FN), where a positive instance is erroneously classified as negative; True Negative (TN), signifying the correct identification of a negative instance; and False Positive (FP), where a negative instance is incorrectly assigned a positive label [24].

To quantify classification accuracy, we define the *True Positive rate* (TPr) and the *False Positive rate* (FPr) as follows:

$$\text{TPr} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (10)$$

$$\text{FPr} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (11)$$

The ROC curve characterizes the trade-off between the TPr (sensitivity) and the FPr (1 – specificity) across a range of decision thresholds. To construct the curve for a specific parameter, the data from classes C_1 and C_2 are consolidated and ranked. A sliding threshold T is then varied across the data range. Assuming the distribution mean of C_1 exceeds that of C_2 , a TP is recorded when a value from C_1 is $\geq T$, while a FP occurs when a value from C_2 is $\geq T$. The optimal classification threshold is identified as the point on the ROC curve with the minimum Euclidean distance to the ideal coordinates (0, 1), representing the highest achievable discriminative power.

C. The Mahalanobis distance

The statistical divergence between two datasets can be quantified using the Mahalanobis distance D_M , defined as:

$$D_M(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \sqrt{(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^T \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)}, \quad (12)$$

where the indices i and j represent the respective groups under comparison, and $\bar{\mathbf{x}}$ denotes the vector of mean values for the parameters of interest. The term $\mathbf{S}_p$ refers to the pooled covariance matrix. Consequently, D_M provides a robust multivariate measure of the dissimilarity between groups, accounting for the internal geometry of the data distribution.

III. DATA AND STUDY AREA

The data exhibit a spatial resolution of 250m and an 16-day sampling rate. They are readily accessible online [27] and are also present in the Google Earth Engine (GEE) cloud database.

The NDVI is obtained from the visible (red) and near-infrared (NIR) by using the following formula

$$\text{NDVI} = \frac{\text{NIR} - \text{red}}{\text{NIR} + \text{red}}. \quad (13)$$

The differential reflectance in these bands provides a means of monitoring density and vigor of green vegetation growth using the spectral reflectivity of solar radiation. Green leaves commonly have larger reflectances in the near-infrared than in the visible range. As the leaves undergo water stress, become diseased or die, they become more yellow and reflect significantly less in the near-infrared range. Therefore the NDVI is regarded as a reliable indicator for land cover variations [14] since its temporal evolution is strongly linked to changes in the state of surface.

In this study, we analyzed three datasets of MODIS NDVI pixels derived from olive groves affected by *Xylella fastidiosa* corresponding to sites where the infection was first reported in 2015 (X2015), in 2016 (X2016), and in 2017 (X2017). As a reference, we also examined datasets of pixels from two healthy olive groves located in the Foggia and Matera areas, which exhibit climatic and topographic conditions almost similar to those of the infected sites. Before performing the CECP analysis of the data, we firstly removed from the datasets outliers. Outliers in the NDVI time series were detected using a moving median filter. For a given series, a sliding window of size 5 was used

to compute the local median, and points that deviated from this local median by more than three times the local median were removed. This approach is robust to short-term fluctuations and noise, such as cloud contamination or sensor errors, because it evaluates each point relative to its local temporal context rather than to the global statistics of the series. The NDVI time series span the period from 2010 to 2022, with each pixel represented by 288 samples. Given the length of the available NDVI time series ($N = 288$ samples per pixel), the analysis was restricted to embedding dimensions $d_x \in \{3, 4\}$, following the conservative criterion proposed by Amigó *et al.* [15], which requires $5 \cdot d_x! \leq N$ to ensure a statistically reliable estimation of the ordinal probability distribution. This condition is satisfied for $d_x = 3$ ($5 \cdot 3! = 30 \leq 288$) and $d_x = 4$ ($5 \cdot 4! = 120 \leq 288$), but violated for $d_x = 5$ ($5 \cdot 5! = 600 > 288$). We processed only data with a gap percentage below 25%. Figure 1 shows the time variation of the spatial means of NDVI data from each of the five investigated sites.

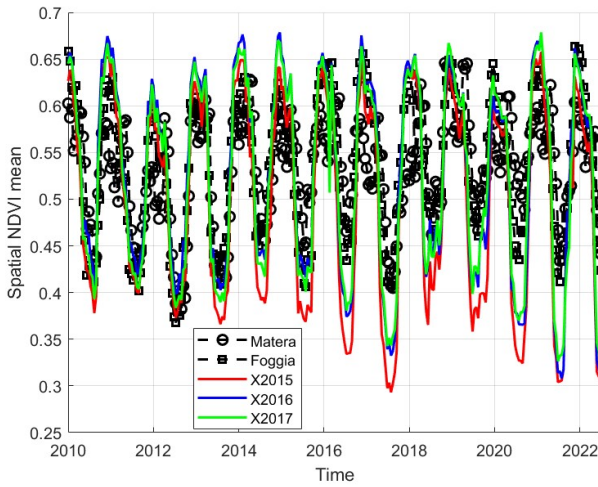


Figure 1. Temporal evolution of the spatial means of NDVI for the investigated sites.

IV. RESULTS

Figure 2 reports the box-plot distributions of H and C for $d_x = 3$. The analysis reveals clear differences in the dynamical behavior of the uninfected datasets (Matera and Foggia) compared to the infected datasets (X2015, X2016, and X2017). In particular, the uninfected datasets are characterized by a larger median entropy with respect to the X2015–X2017 sites, indicating a lower degree of intrinsic organization in the underlying dynamics. In contrast, the infected datasets exhibit a larger decrease in H , consistent with diminished stochasticity in the signal evolution.

An opposite tendency is observed for the complexity measure. The uninfected datasets show lower median values of C . Conversely, the X2015–X2017 datasets display larger complexity levels. The simultaneous decrease in entropy and increase in complexity observed in the latter datasets suggests a transi-

tion from disordered, noise-dominated toward more structured states.

The pairwise analysis conducted in the CECF (Figure 3) further quantifies the separation between the datasets. The CECF plots reveal Matera and Foggia datasets consistently occupying the low-complexity/high-entropy region of the plane; while the X-series datasets cluster significantly toward the high-complexity/low-entropy region. The Mahalanobis distance for Matera vs X-data range from $D_M = 1.53$ to $D_M = 0.95$; while a bit smaller Mahalanobis distance is for Foggia vs X-data, ranging from $D_M = 1.13$ to $D_M = 0.56$. These values confirm a quantifiable divergence in the statistical quantifiers of the uninfected and the infected sites. The ROC curves (Figure 4) yield significant Area Under the Curve (AUC) values from 0.86 to 0.74 for Matera vs X-data and from 0.79 to 0.66 for Foggia vs X-data, providing a measure of diagnostic accuracy. The discrimination between uninfected and infected (X2015 and X2016) datasets is good (but larger for Matera). However, for the more recent X2017 dataset, the AUC indicates a less efficient discrimination between the two types of datasets.

Similar results (not shown) are obtained for $d_x = 4$. Tables I to IV report the values of D_M , thresholds, AUC , TPr, and FPr for H and C for $d_x = 3$ and $d_x = 4$. From a synoptical evaluation of the reported results, $d_x = 4$ provides a bit more refined characterization of the signal dynamics, offering higher sensitivity in discriminating between uninfected and infected series compared to $d_x = 3$.

Table I. SUMMARY OF PERMUTATION ENTROPY (H) AND COMPLEXITY (C) METRICS FOR MATERA ($d_x = 3$).

Metric	Matera vs X2015	Matera vs X2016	Matera vs X2017
D_M	1.530	1.265	0.953
Permutation Entropy (H)			
Threshold $_H$	0.969	0.971	0.973
AUC $_H$	0.859	0.816	0.745
FPR $_H$	0.259	0.315	0.419
TPR $_H$	0.856	0.829	0.797
Complexity (C)			
Threshold $_C$	0.031	0.031	0.031
AUC $_C$	0.858	0.814	0.743
FPR $_C$	0.181	0.178	0.187
TPR $_C$	0.778	0.692	0.568

Table II. SUMMARY OF PERMUTATION ENTROPY (H) AND COMPLEXITY (C) METRICS FOR FOGGIA ($d_x = 3$).

Metric	Foggia vs X2015	Foggia vs X2016	Foggia vs X2017
D_M	1.127	0.836	0.561
Permutation Entropy (H)			
Threshold $_H$	0.972	0.972	0.980
AUC $_H$	0.791	0.734	0.665
FPR $_H$	0.218	0.299	0.290
TPR $_H$	0.677	0.681	0.549
Complexity (C)			
Threshold $_C$	0.029	0.030	0.024
AUC $_C$	0.789	0.731	0.662
FPR $_C$	0.349	0.338	0.450
TPR $_C$	0.809	0.718	0.707

V. CONCLUSION AND FUTURE WORK

The results obtained in this study provide a robust dynamical framework for characterizing the impact of *Xylella fastidiosa* on

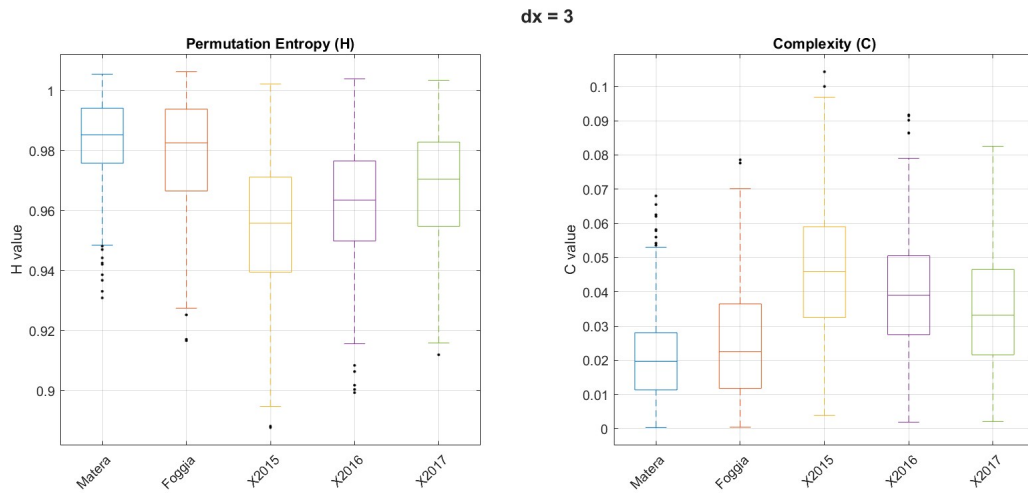


Figure 2. Box-plots of H and C for the investigated datasets and embedding dimension $d_x = 3$.

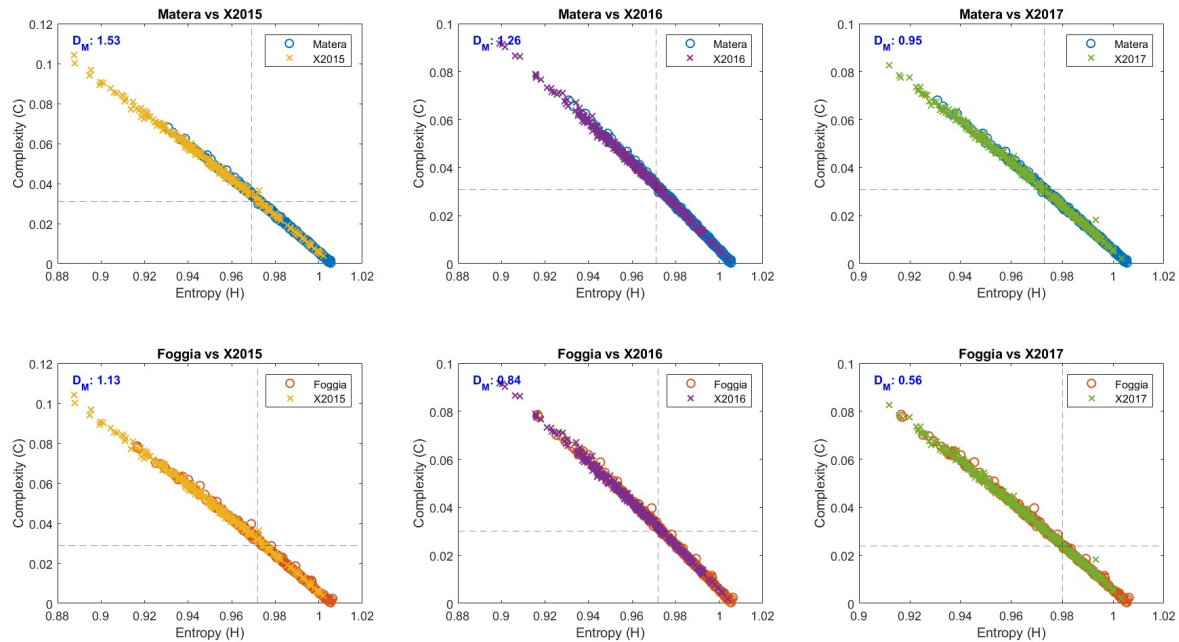


Figure 3. CECPs for $d_x = 3$. The dotted vertical and horizontal lines mark the optimal thresholds for H and C .

Table III. SUMMARY OF PERMUTATION ENTROPY (H) AND COMPLEXITY (C) METRICS FOR MATERA ($dx = 4$).

Metric	Matera vs X2015	Matera vs X2016	Matera vs X2017
D_M	1.666	1.382	1.111
Permutation Entropy (H)			
Threshold $_H$	0.936	0.939	0.943
AUC $_H$	0.878	0.834	0.777
FPR $_H$	0.175	0.242	0.297
TPR $_H$	0.806	0.785	0.747
Complexity (C)			
Threshold $_C$	0.090	0.089	0.090
AUC $_C$	0.872	0.827	0.772
FPR $_C$	0.208	0.216	0.208
TPR $_C$	0.818	0.754	0.647

Table IV. SUMMARY OF PERMUTATION ENTROPY (H) AND COMPLEXITY (C) METRICS FOR FOGGIA ($dx = 4$).

Metric	Foggia vs X2015	Foggia vs X2016	Foggia vs X2017
D_M	1.135	0.837	0.617
Permutation Entropy (H)			
Threshold $_H$	0.928	0.931	0.931
AUC $_H$	0.789	0.729	0.673
FPR $_H$	0.243	0.325	0.421
TPR $_H$	0.718	0.694	0.694
Complexity (C)			
Threshold $_C$	0.102	0.086	0.096
AUC $_C$	0.780	0.720	0.665
FPR $_C$	0.267	0.423	0.315
TPR $_C$	0.725	0.780	0.588

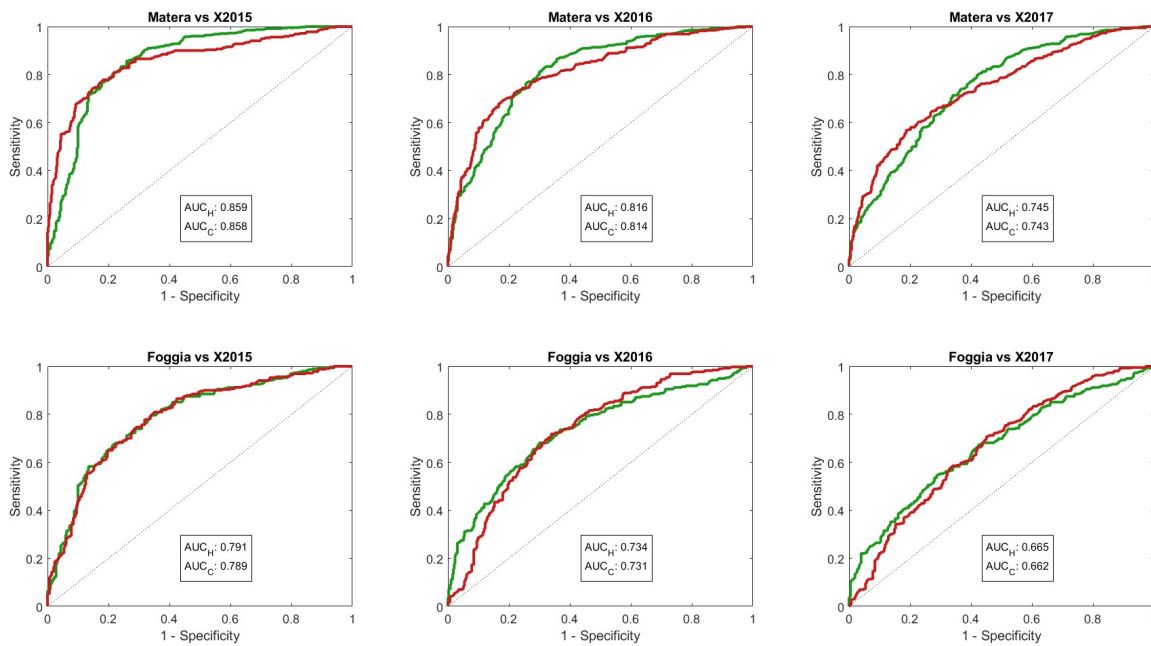


Figure 4. ROC curves (bottom panels) for $d_X = 3$. The green and red ROC curves refer to H and C respectively.

olive groves.

The transition from uninfected sites (Matera and Foggia) to the infected datasets (X2015–X2017) signifies a profound shift in the system’s dynamical regime: the high values of H and low levels of C observed in Matera and Foggia indicate a system in a flexible, resilient equilibrium; the signal fluctuates freely, reflecting a “natural noise” where a healthy canopy absorbs environmental variations without losing its dynamical “plasticity”. In contrast, the simultaneous decrease in H and increase in C in the X-datasets suggests that the infection imposes a specific “structure” onto the signal; the resulting desiccation is not a random process but a structured, ordered degenerative pathway. This is also observed in Figure 1, where the NDVI time series of the infected pixels reveals a progressive structured decay—most notably following the 2015 outbreak—with the seasonal troughs becoming progressively deeper. In fact, in the infected sites, the thinning of the olive tree canopy causes the 250 m pixel to be increasingly influenced by the underlying herbaceous vegetation. Unlike the stable, evergreen nature of healthy olive trees, grass follows a significantly more extreme seasonal cycle, characterized by rapid growth followed by quick senescence. The dominance of these rigid seasonal cycles makes the signal dynamics in the infected sites more “predictable” compared to the stochastic stability of a healthy evergreen canopy. Consequently, the infected sites exhibit a decrease in permutation entropy, as the biological signal is replaced by a more deterministic, seasonally-driven pattern. Furthermore, the transition from a stable perennial state to a structured decline introduces non-trivial temporal correlations that enhance the C .

The values of the AUC and the Mahalanobis Distances confirm that non-linear analysis is a powerful diagnostic tool. Matera serves as the “gold standard” for olive grove purity, making the contrast with infected sites highly evident. The D_M from 1.53 to 0.95 quantifies this clear divergence. A smaller D_M between 1.13 and 0.56 and slightly lower AUC values suggest that Foggia, while healthy, may possess a dynamics closer to the infected Apulian sites. This could be attributed to shared pedoclimatic conditions or similar land management practices; nevertheless, the statistical distinction remains significant.

In summary, the approach based on the CECF does more than merely identify the presence of *Xylella fastidiosa*; it quantifies the pathogen’s ability to dismantle the plant’s biological regulation mechanisms. This diagnostic framework effectively characterizes the transition where the inherent flexibility of a healthy life state is replaced by the rigid, structured dynamics of a pathological decline in olive trees.

REFERENCES

- [1] C. Bandt and B. Pompe, “Permutation entropy: A natural complexity measure for time series,” *Physical Review Letters*, vol. 88, p. 174102, 2002.
- [2] European Commission Joint Research Centre, “Monitoring the impacts of *Xylella fastidiosa*,” https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/monitoring-impacts-xylella-fastidiosa-2019-11-12_en, [retrieved: May, 2025].
- [3] M. Saponari, D. Boscia, F. Nigro, and G. P. Martelli, “Identification of DNA sequences related to *Xylella fastidiosa* in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (Southern Italy),” *Journal of Plant Pathology*, vol. 95, pp. 659–668, 2013.
- [4] M. Jeger, J. Bragard, D. Caffier, T. Candresse, E. Chatzivassiliou, K. Dehnen-Schmutz, and G. Lovell, “Scientific opinion on the updated pest categorisation of *Xylella fastidiosa*,” *EFSA Journal*, vol. 16, p. 5357, 2018.

- [5] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, pp. 265–274, 2009.
- [6] J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, pp. 2992–3004, 2007.
- [7] G. Chen and D. Needell, "Compressed sensing and dictionary learning," *Finite Frame Theory: A Complete Introduction to Overcompleteness*, vol. 73, pp. 201–241, 2016.
- [8] L. Telesca, M. Lovallo, A. Pisarenko, and R. Lasaponara, "Exploring long-term anomalies in the vegetation cover of peri-urban parks using the Fisher–Shannon method," *Entropy*, vol. 24, p. 1784, 2022.
- [9] L. Telesca, M. Lovallo, A. Pisarenko, and R. Lasaponara, "Urban and peri-urban vegetation monitoring using satellite MODIS NDVI time series, singular spectrum analysis, and Fisher–Shannon statistical method," *Sustainability*, vol. 15, p. 11039, 2023.
- [10] L. Telesca, N. Abate, F. Faridani, M. Lovallo, and R. Lasaponara, "Revealing traits of phytopathogenic status induced by *Xylella fastidiosa* in olive trees by analyzing multifractal and informational patterns of MODIS satellite evapotranspiration data," *Physica A*, vol. 629, p. 129163, 2023.
- [11] L. Telesca, N. Abate, F. Faridani, M. Lovallo, and R. Lasaponara, "Discerning *Xylella fastidiosa*-infected olive orchards in the time series of MODIS Terra satellite evapotranspiration data by using the Fisher–Shannon analysis and the multifractal detrended fluctuation analysis," *Fractal and Fractional*, vol. 7, p. 466, 2023.
- [12] L. Telesca, N. Abate, M. Lovallo, and R. Lasaponara, "Investigating the impact of *Xylella fastidiosa* on olive trees by the analysis of MODIS Terra satellite evapotranspiration time series by using the Fisher information measure and the Shannon entropy: A case study in Southern Italy," *Remote Sensing*, vol. 16, p. 1242, 2024.
- [13] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, "Distinguishing noise from chaos," *Physical Review Letters*, vol. 99, p. 154102, 2007.
- [14] K. E. Huemrich, T. A. Black, P. G. Jarvis, J. H. McCaughey, and E. G. Hall, "High temporal resolution NDVI phenology from micrometeorological radiation sensor," *Journal of Geophysical Research*, vol. 104, p. 27935, 1999.
- [15] J. M. Amigó, M. B. Kennel, and L. Kocarev, "The permutation entropy rate equals the metric entropy rate for ergodic information sources and ergodic dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 228, no. 1, pp. 77–85, 2007.
- [16] L. Zunino, M. C. Soriano, and O. A. Rosso, "Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach," *Physical Review E*, vol. 86, p. 046210, 2012.
- [17] F. Ma, Q. Fan, and G. Ling, "Complexity-entropy causality plane analysis of air pollution series," *Fluctuation and Noise Letters*, vol. 21, p. 2250011, 2021.
- [18] A. S. A. D. Silva, I. D. D. C. Barreto, M. Cunha-Filho, R. S. C. Menezes, B. Stosic, and T. Stosic, "Multiscale complexity analysis of rainfall in Northeast Brazil," *Water*, vol. 13, p. 3213, 2021.
- [19] M. Suriano, L. F. Caram, and O. A. Rosso, "Daily streamflow of Argentine rivers analysis using information theory quantifiers," *Entropy*, vol. 26, p. 56, 2024.
- [20] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.
- [21] G. A. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology: Problems and Methods*, H. Quastler, Ed. Glencoe, IL: Free Press, pp. 95–100, 1955.
- [22] J. De Gregorio, D. Sánchez, and R. Toral, "Entropy estimators for Markovian sequences: A comparative analysis," *Entropy*, vol. 26, p. 79, 2024.
- [23] R. López-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, pp. 321–326, 1995.
- [24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
- [25] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series," *Physical Review E*, vol. 80, p. 046103, 2009.
- [26] M. Saponari, A. Giampetruzzi, G. Loconsole, D. Boscia, and P. Saldarelli, "*Xylella fastidiosa* in olive in Apulia: Where we stand," *Phytopathology*, vol. 109, pp. 175–186, 2019.
- [27] NASA Land Processes Distributed Active Archive Center, "LP DAAC – Land Products," <https://lpdaac.usgs.gov>, [retrieved: May, 2025].
- [28] L. Laipelt, R. H. B. Kayser, A. S. Fleischmann, A. Ruhoff, W. Bastiaanssen, T. A. Erickson, and F. Melton, "Long-term monitoring of evapotranspiration using the SEBAL algorithm and Google Earth Engine cloud computing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 81–96, 2021.
- [29] MathWorks File Exchange, "Fast natural visibility graph (NVG) for MATLAB," <https://www.mathworks.com/matlabcentral/fileexchange/70432>, [retrieved: May, 2025].
- [30] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles," *Physical Review E*, vol. 85, p. 016117, 2012.
- [31] B. A. Gonçalves, L. Carpi, O. A. Rosso, and M. G. Ravetti, "Time series characterization via horizontal visibility graph and information theory," *Physica A*, vol. 464, pp. 93–102, 2016.
- [32] K. Schneider, W. van der Werf, M. Cendoya, M. Mourits, J. A. Navas-Cortés, A. Vicent, and A. Oude Lansink, "Impact of *Xylella fastidiosa* subspecies *pauca* in European olives," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 9250–9259, 2020.

Proposition of a Brain-Computer Interface based on Pictograms for Communication Purpose

Álvaro Fernández-Rodríguez 

Department of Basic Psychology
University of Málaga
Málaga, Spain
e-mail: alvarofr@uma.es

Simon Audren

Lou Guiochet

Véronique Lespinet-Najib 

Jean-Marc André 

Laboratoire IMS (CNRS UMR 5218)
Cognitive Team, Bordeaux INP-ENSC
Talence, France
e-mail: saudren@ensc.fr,
lguiochet@ensc.fr,
veronique.lespinet@ensc.fr,
jean-marc.andre@ensc.fr

Francisco Velasco-Álvarez 

Ricardo Ron-Angevin 

Department of Electronic Technology
University of Málaga
Málaga, Spain
e-mail: fvelasco@uma.es,
rron@uma.es

Abstract—Brain-Computer Interface (BCI) technology enables direct communication between the brain and external devices, bypassing traditional neuromuscular pathways. This study presents a pictogram-based BCI designed to address the communication needs of patients with Amyotrophic Lateral Sclerosis (ALS) through a hierarchical menu architecture. A preliminary evaluation involving five able-bodied participants was conducted to assess the system's performance. The results demonstrate the feasibility of the proposed BCI as a functional Augmentative and Alternative Communication (AAC) tool. Although the current framework appears viable, further optimization regarding usability and interaction efficiency is required to ensure its effectiveness in the target clinical population.

Keywords—brain-computer interface (BCI); assistive technology; event-related potential (ERP); augmentative and alternative communication (AAC); pictogram.

I. INTRODUCTION

Brain-Computer Interfaces (BCIs) represent a specialized class of assistive technology that facilitates interaction through the exclusive modulation of neural signals [1]. The primary objective of these systems is to enhance the user's autonomy within their environment [2]. In cases of neurodegenerative conditions such as Amyotrophic Lateral Sclerosis (ALS), conventional assistive devices—such as voice recognition and eye-tracking—often prove inadequate due to the severe impairment of muscular and ocular functions [3]. Consequently, for individuals facing profound motor limitations, BCIs emerge as a critical alternative by circumventing the requirement for neuromuscular pathways.

In these instances, BCIs provide a robust alternative by decoding Electroencephalographic (EEG) activity to re-establish functional interaction [3]. One of the most effective EEG-based methods is the use of Event-Related Potentials (ERPs), which correspond to brain responses elicited by attended stimuli [4].

A BCI speller is a typical brain-computer interface system for communication purposes. This technology can provide users with severe motor disabilities, such as patients with ALS, with an assistive device controlled by brain activity.

Most BCI spellers are based on the P300 ERP. The P300 signal is a positive voltage deflection occurring approximately 300 ms after an infrequent or significant stimulus is perceived [5]. The P300 amplitude typically ranges between 2–5 μV and is symmetrically distributed around central scalp areas, showing greater amplitude in occipital than in frontal regions [6]. Most of these spellers are based on the P300 speller first developed by Farwell and Donchin [5]. In this BCI, a 6×6 matrix of letters arranged in rows and columns is presented to the user. The user focuses attention on the matrix element he or she wishes to select while each row and column is flashed, or intensified, randomly. After a number of flashes, the symbol that the system identifies as the intended target is presented on the screen. This paradigm is known as the Row–Column Paradigm (RCP).

Augmentative and Alternative Communication (AAC) systems empower individuals with severe speech and motor impairments by facilitating interaction through selection-based visual interfaces [7]. Specifically, pictogram-based AAC frameworks have been shown to optimize communication efficiency and enhance overall accessibility [8].

In this study, we introduce a preliminary pictogram-based AAC framework specifically engineered to facilitate communication for individuals with profound motor impairments. The proposed system aims to enhance daily communicative autonomy and, consequently, improve the overall quality of life of this population. By leveraging intuitive visual representations, the architecture seeks to reduce cognitive load while maintaining high functional utility in real-world interaction scenarios. The main contribution of this work is the preliminary design and evaluation of a configurable

pictogram-based ERP-BCI for AAC, intended to support functional communication through the selection of meaningful pictograms rather than conventional letter-by-letter spelling. Specifically, we hypothesized that, in this preliminary evaluation with able-bodied users, participants would achieve adequate online performance and efficient message composition through the selection of pictograms.

This paper is organized as follows: Section II describes the experimental setup and presents details about the proposed AAC system. The results and discussion are presented in Sections III and IV, respectively, followed by the conclusion and future work in Section V.

II. METHODS

A. Participants

Five healthy university students (aged 18.20 ± 0.45 years; 3 females, 2 males; referred to as SU01–SU05) participated in this study. None of them had previous experience using a BCI system. The study was approved by the Ethics Committee of the University of Malaga and met the ethical standards of the Declaration of Helsinki. According to self-reports, all participants had no history of neurological or psychiatric illness and had normal or corrected-to-normal vision. Written informed consent was obtained from all participants prior to data collection.

B. Data Acquisition and Signal Processing

The EEG was recorded at a sample rate of 250 Hz using the electrode positions: Fz, Cz, Pz, Oz, P3, P4, PO7 and PO8, according to the 10/20 international system. All channels were referenced to the left mastoid and grounded to position AFz. Signals were amplified by an acti-CHamp amplifier (Brain Products GmbH, Munich, Germany).

EEG data acquisition and signal processing were managed via the UMA BCI Speller software [9], an open-source tool developed by the UMA-BCI group [10]. Built upon the established BCI2000 platform [11], this software provides a streamlined interface that simplifies complex system configurations. Signal classification was executed using a step-wise linear discriminant analysis (SWLDA) classifier, adhering to the methodology implemented in the BCI2000 P300Classifier module. This approach facilitates the customization of classifier parameters for each participant, thereby ensuring optimized performance during real-time (online) operation.

C. Proposed system

The proposed AAC system is specifically designed for individuals with severe motor impairments, including those unable to utilize muscle-based communication channels such as speech. Configured via the UMA-BCI Speller tool [9], the system allows for precise adaptation to individual user requirements. While the current implementation is tailored

for Spanish-speaking users, the architecture is language-agnostic and can be reconfigured for any language.

The interface layout, menu options, and pictogram selection were defined and validated by clinical experts from the ALS Center at Hôpital Pellegrin de Bordeaux, leveraging their extensive experience with this patient demographic.

The pictograms used in the proposed system were obtained from the ARASAAC database (Centro Aragonés para la Comunicación Aumentativa y Alternativa) [12] and were selected for their relevance in AAC contexts. The selection of these pictograms ensures consistency with established clinical and educational practices for non-verbal communication systems.

Communication is facilitated through the sequential selection of pictograms organized within a hierarchical menu architecture (Figure 1). Each submenu integrates control functions to optimize user flow, including navigation to previous or primary menus, alert triggering, and text buffer management (deletion and validation). Furthermore, a virtual keyboard is included to enable the composition of personalized messages. The complete AAC framework comprises 18 distinct menus, designed to promote navigational fluidity and intuitive command selection while minimizing cognitive load. The main menu includes five main categories: needs (“NECESIDADES”), wants (“QUIERO”), questions (“PREGUNTA”), feelings and emotional states (“SENTIMIENTOS”), and physical states (“ESTADO FÍSICO”). In addition, the main menu includes two direct response options: Yes (“SÍ”) and No (“NO”). These menus are shown in Figure 1.

The “+” symbol in the commands indicates that selecting the option leads to a submenu with additional commands. Following the RCP, stimuli were presented by intensifying each row and column of a 6×6 matrix. The intensification, also known as a flash, consisted of an overlay of a high-contrast, red-tinted image of a well-known face (Billie Eilish) against a white background. This stimulus selection was based on the work of [13], who demonstrated that familiar human faces with high visual contrast elicit more robust ERPs, thereby enhancing performance in attention-based BCI systems. While the number of functional pictograms varied across submenus, the 6×6 grid dimensions remained constant. Consequently, all rows and columns were flashed during the RCP cycle, including those containing empty cells. To maintain system stability, selections corresponding to empty cells triggered no state changes in the interface.

The stimulus presentation timing was defined by an intensification duration of 192 ms and an inter-stimulus interval (ISI) of 96 ms, resulting in a stimulus-onset asynchrony (SOA) of 288 ms. The number of sequences per selection was dynamically adjusted based on the calibration phase and individual user performance, as further detailed in Section II-D. To facilitate user orientation, a 4-second pre-selection pause was implemented, followed by a post-selection delay of 1472 ms to allow for visual feedback and

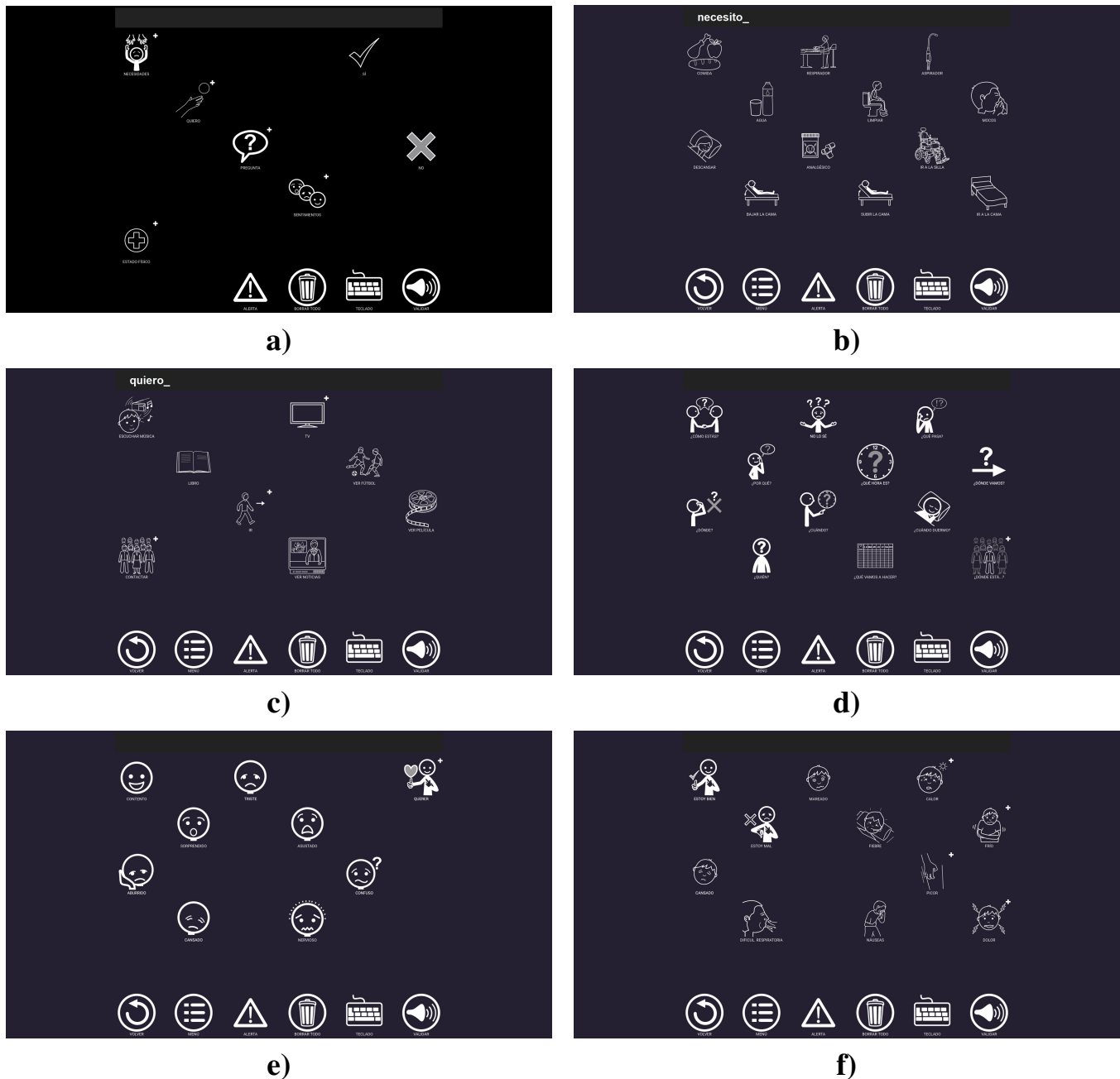


Figure 1. Overview of the pictogram-based augmentative and alternative communication (AAC) interface operated through the event-related potential (ERP)-based brain-computer interface (BCI). The figure illustrates a) the main menu and several representative submenus used for communication and interaction, including b) needs, c) wants, d) questions, e) emotional states, and f) physical states.

gaze shifting before the subsequent trial.

In summary, the platform enables the synthesis of verbal messages through the sequential selection of pictograms. Upon final validation, the system transmits the constructed message to caregivers or family members via simultaneous Text-To-Speech (TTS) output and visual display. For example, to communicate the request “I want to watch the news,” the user navigates the hierarchy by selecting the following sequence: “want+”, “watch news”, and “validate”

(Figure 2). This multimodal feedback ensures clear and effective communication in real-world environments.

D. Procedure

Each experimental session began with a detailed explanation of the protocol, followed by EEG cap placement and impedance verification. Once signal quality was confirmed, participants received standardized instructions on how to operate the system. Specifically, they were asked to attend to

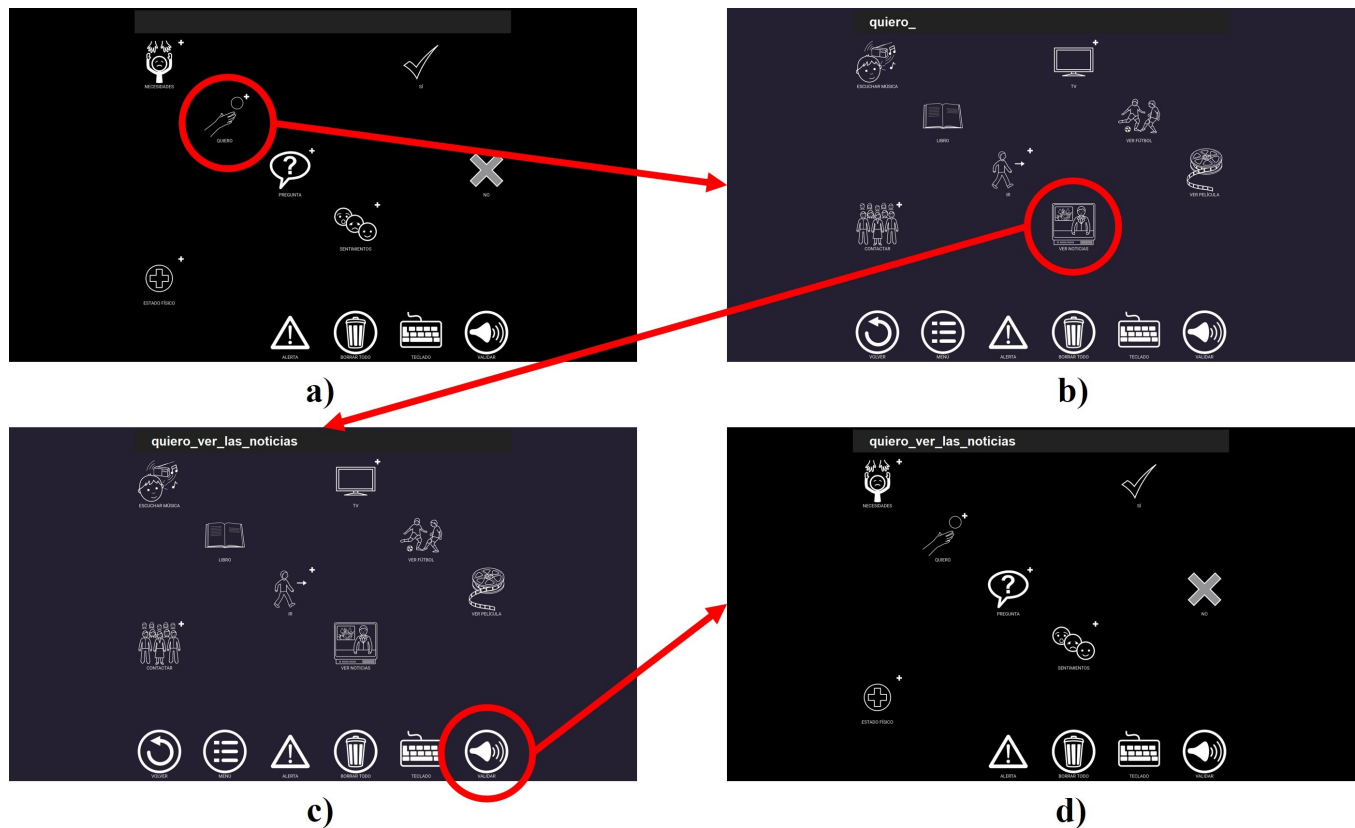


Figure 2. Examples of command selection in the proposed augmentative and alternative communication (AAC) interface for expressing the intention to watch the news: a) the user starts from the main menu and selects the *want* (“*QUIERO*”) submenu; b) the specific desire to watch the news is selected (“*VER NOTICIAS*”); c) the selection is validated using the *validate* command (“*VALIDAR*”) so that the command is issued; and d) the interface automatically returns to the main menu.

the desired target command within the matrix and silently count the number of times it was intensified, in order to enhance attentional engagement and promote a robust P300 response.

The interface was based on a classical 6×6 RCP. In this configuration, one stimulus sequence consisted of the sequential intensification of all 12 matrix elements (6 rows and 6 columns), resulting in 12 flashes per sequence. Although, from a theoretical standpoint, a single sequence is sufficient to determine the target coordinates, reliable P300 detection typically requires multiple repetitions to increase the Signal-to-Noise Ratio (SNR) and improve classification accuracy. Accordingly, the number of stimulus sequences was treated as an adjustable parameter and optimized depending on the experimental phase. The procedure comprised two consecutive stages: (i) an offline calibration phase and (ii) an online control phase.

1) Calibration phase: Calibration was performed using a dedicated calibration menu consisting of a 6×6 alphanumeric matrix (letters and digits) superimposed with pictograms (Figure 3). This layout was intentionally designed to resemble the final AAC interface in order to maintain visual consistency and reduce contextual changes between

calibration and online use.

Participants were instructed to spell three predefined Spanish four-letter words: *GUAY*, *TREN*, and *FILO*. The number of stimulus sequences was fixed at five during calibration. Consequently, each target item was intensified 10 times (5 row flashes and 5 column flashes). The total time required to complete a single selection was 22.752 s. This duration was determined by an SOA of 288 ms applied to 12 flashes across five sequences (17.28 s), plus a 4 s pre-selection interval and a 1.472 s post-selection interval.

After data acquisition, a subject-specific P300 classifier was computed using Stepwise Linear Discriminant Analysis (SWLDA). Model performance was estimated by means of a three-fold cross-validation procedure. Based on the resulting accuracy values, the number of stimulus sequences to be used during the online phase was individualized for each participant. Sequence optimization followed a performance-maximization criterion. First, classification accuracy was evaluated using a maximum of five sequences. If 100% accuracy was achieved at this level, a retrospective analysis was conducted to determine the minimum number of sequences required to maintain perfect performance. If ceiling accuracy was not reached within five sequences, the

configuration yielding the highest accuracy at this maximum length was retained for the online session. In all cases, the minimum number of sequences was set to three. This adaptive procedure sought to balance classification reliability and communication speed.



Figure 3. Interface used during the calibration task. The highlighted fifth column illustrates the stimulation mechanism of the row-column paradigm (RCP), in which entire rows and columns of the matrix are sequentially intensified. The face stimulus shown in the figure has been pixelated for copyright reasons.

2) *Online phase:* Prior to initiating BCI-based control, participants were allowed to freely explore the AAC interface using a conventional computer mouse. This familiarization period ensured that users understood the system structure, available commands, and menu navigation logic without the cognitive demands imposed by BCI operation.

Once participants reported being comfortable with the interface, they were instructed to prepare three independent command sets. Each set consisted of six consecutive command selections (18 commands in total) to be completed exclusively through BCI control and without interruption. Participants were free to determine the semantic content of each set, provided that it could be generated using the available AAC commands. If a selection error occurred, participants were required to correct it and continue the sequence until the intended six-command set was successfully completed. Between command sets, the system was paused to allow rest periods and minimize fatigue. This design aimed to emulate realistic AAC use while maintaining experimental control over task structure and workload.

III. RESULTS

A. Calibration phase

The classification accuracy as a function of the number of stimulus sequences is shown in Table I. Overall, calibration performance was high. Four out of five participants (SU01, SU02, SU04, and SU05) achieved 100% accuracy. For these participants, ceiling performance was reached with three stimulus sequences or fewer, indicating robust P300 responses that enabled reliable target detection with a reduced number of repetitions. Notably, SU04 and SU05

achieved 100% accuracy from the first sequence. In contrast, participant SU03 did not reach 100% accuracy during calibration. Although performance varied across sequences, the maximum accuracy obtained was 91.67% using five sequences. According to the predefined selection criterion for the online phase, the mean number of stimulus sequences chosen per participant was 3.40 ± 0.89 . For all participants except SU03, three sequences were sufficient to ensure near-perfect or perfect performance. Participant SU03 required five sequences and did not achieve ceiling accuracy.

B. Online phase

Table II summarizes the online performance during the BCI-AAC control phase. Overall accuracy remained high during real-time operation, with a mean accuracy of 95.84%. Two participants (SU01 and SU05) achieved 100% online accuracy, with no false positives (FP) or false negatives (FN). Participants SU03 and SU04 showed minor performance reductions compared to calibration, with one false negative each, while SU02 presented one false positive and one false negative, resulting in 90% accuracy. Importantly, despite the reduced number of sequences for most users (3.40 ± 0.89), classification accuracy remained above 90% in all cases, supporting the robustness of the ERP-BCI system in a realistic AAC control scenario.

The mean task completion time to execute the 18 commands was 319.51 s. As expected, participant SU03, who required five sequences, showed the longest completion time (409.54 s), illustrating the well-known trade-off between speed and accuracy in P300-based BCIs. Participants operating with three sequences completed the task substantially faster (mean ≈ 297 s), demonstrating that reducing the number of repetitions can significantly improve communication rate without severely compromising accuracy.

IV. DISCUSSION

The present results suggest that the proposed ERP-based BCI system enables reliable control of an AAC interface, achieving high online accuracy ($95.84 \pm 4.24\%$) with low inter-user variability. In addition, the proposed system offers an extensive and flexible set of commands, which can be easily configured using the UMA BCI Speller software [9].

In conventional letter-by-letter AAC systems, conveying a complete message (e.g., expressing a need) typically requires multiple selections, making communication slower than when using nested pictograms. Recent advances in text prediction and generative Artificial Intelligence (AI) have already been incorporated into speller-based systems to enhance communication speed and efficiency [14]. In contrast, although some recent work has begun to explore the integration of predictive or AI-based language support into AAC systems beyond traditional spellers [8], its application to pictogram-based ERP-BCI interfaces remains limited.

TABLE I. CALIBRATION ACCURACY (%) AS A FUNCTION OF THE NUMBER OF STIMULUS SEQUENCES FOR EACH PARTICIPANT, INCLUDING THE MEAN $\pm$ STANDARD DEVIATION (SD).

Participant	Sequences				
	1	2	3	4	5
SU01	66.67	100.00	100.00	100.00	100.00
SU02	33.33	75.00	100.00	100.00	100.00
SU03	58.33	91.67	83.33	75.00	91.67
SU04	100.00	100.00	100.00	100.00	100.00
SU05	100.00	100.00	100.00	100.00	100.00
Mean $\pm$ SD	71.66 $\pm$ 28.63	91.33 $\pm$ 12.10	96.67 $\pm$ 7.46	95.00 $\pm$ 11.18	98.33 $\pm$ 3.73

TABLE II. ONLINE PERFORMANCE RESULTS FOR EACH PARTICIPANT, INCLUDING THE MEAN $\pm$ STANDARD DEVIATION (SD), DURING THE ONLINE PHASE. TRUE POSITIVES (TP) CORRESPOND TO CORRECTLY SELECTED TARGET COMMANDS; FALSE POSITIVES (FP) CORRESPOND TO SELECTIONS OF NON-TARGET COMMANDS; FALSE NEGATIVES (FN) CORRESPOND TO TRIALS IN WHICH AN EMPTY CELL (I.E., A CELL WITHOUT AN ASSOCIATED COMMAND) WAS SELECTED, RESULTING IN NO EXECUTED COMMAND DESPITE AN INTENDED TARGET.

User	Sequences	TP	FP	FN	Accuracy (%)	Time (s)
SU01	3	18	0	0	100.00	285.12
SU02	3	18	1	1	90.00	316.80
SU03	5	17	0	1	94.44	409.54
SU04	3	18	0	1	94.74	300.96
SU05	3	18	0	0	100.00	285.12
Mean $\pm$ SD	3.40 $\pm$ 0.89	17.80 $\pm$ 0.45	0.20 $\pm$ 0.45	0.60 $\pm$ 0.55	95.84 $\pm$ 4.24	319.51 $\pm$ 52.01

Extending generative and predictive strategies to pictogram-driven AAC systems could substantially improve message construction efficiency while preserving the advantages of symbolic communication.

Nevertheless, the present findings should be considered preliminary, as the sample included only five participants without severe motor impairments. Therefore, the results cannot be directly generalized to users with severe motor disabilities, who may face additional challenges related to fatigue, attention, visual control, and prolonged use. Future work should involve testing the system with this population, in collaboration with clinical professionals, relatives, and caregivers, to evaluate its usability, communication efficiency, and required adaptations.

V. CONCLUSIONS

This study demonstrates the feasibility of controlling a flexible and customizable pictogram-based AAC interface through an ERP-based BCI with high reliability. The proposed system, defined and validated by clinical experts from the ALS Center at Hôpital Pellegrin de Bordeaux, provides a wide and configurable set of commands, allowing users to construct meaningful messages beyond simple letter-by-letter spelling. These commands can be modified and adapted to patient needs.

The high online performance observed across participants supports the robustness of the approach and its suitability for practical communication scenarios. By combining symbolic

communication with BCI control, the system represents a promising step toward more functional and user-centered assistive communication solutions.

Future work should focus on validating the system with end users presenting severe motor impairments and assessing its long-term usability in real-world environments.

ACKNOWLEDGMENTS

This work was partially supported by the project PID2024-160781OB-I00 (ICCADIS), funded by MCIN (Ministerio de Ciencia e Innovación)/AEI (Agencia Estatal de Investigación) /10.13039/501100011033/FEDER, UE (Fondo Europeo de Desarrollo Regional). This work was also partially supported by the University of Malaga (UMA). The authors would like to thank all participants for their cooperation.

REFERENCES

- [1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002, ISSN: 13882457. DOI: 10.1016/S1388-2457(02)00057-3.
- [2] E. Karikari and K. A. Koshechkin, "Review on brain-computer interface technologies in healthcare," *Biophysical Reviews*, vol. 15, no. 5, pp. 1351–1358, 2023, ISSN: 18672469. DOI: 10.1007/s12551-023-01138-6.
- [3] E. Aust et al., "Impairment of oculomotor functions in patients with early to advanced amyotrophic lateral sclerosis," *Journal of Neurology*, vol. 271, no. 1, pp. 325–339, 2024, ISSN: 14321459. DOI: 10.1007/s00415-023-11957-y.

- [4] B. Z. Allison, A. Kubler, and J. Jin, “30+ years of P300 brain–computer interfaces,” *Psychophysiology*, vol. 57, no. 7, pp. 1–18, 2020. DOI: 10.1111/psyp.13569.
- [5] L. A. Farwell and E. Donchin, “Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials,” *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988, ISSN: 00134694. DOI: 10.1016/0013-4694(88)90149-6.
- [6] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “Toward enhanced P300 speller performance,” *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 15–21, 2008, ISSN: 01650270. DOI: 10.1016/j.jneumeth.2007.07.017.
- [7] Y. Elshar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor, “Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability,” *Sensors (Switzerland)*, vol. 19, no. 8, 2019, ISSN: 14248220. DOI: 10.3390/s19081911.
- [8] R. Hervás, S. Bautista, G. Méndez, P. Galván, and P. Gervás, “Predictive composition of pictogram messages for users with autism,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 5649–5664, 2020, ISSN: 18685145. DOI: 10.1007/s12652-020-01925-z.
- [9] F. Velasco-Álvarez et al., “UMA-BCI Speller: an Easily Configurable P300 Speller Tool for End Users,” *Computer Methods and Programs in Biomedicine*, vol. 172, pp. 127–138, 2019, ISSN: 0169-2607. DOI: 10.1016/J.CMPB.2019.02.015.
- [10] UMA-BCI Group, *Uma-bci group website*, Accessed: May 2026. [Online]. Available: <https://umabci.uma.es>.
- [11] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, “BCI2000: A general-purpose brain-computer interface (BCI) system,” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004, ISSN: 00189294. DOI: 10.1109/TBME.2004.827072. arXiv: 1741-2560/2/4/008 [10.1088].
- [12] Centro Aragonés para la Comunicación Aumentativa y Alternativa, *Arasaac*, [retrieved: April, 2026]. [Online]. Available: <https://arasaac.org>.
- [13] X. Zhang, J. Jin, S. Li, X. Wang, and A. Cichocki, “Evaluation of color modulation in visual P300-speller using new stimulus patterns,” *Cognitive Neurodynamics*, vol. 0123456789, no. Birbaumer 1999, 2021, ISSN: 18714099. DOI: 10.1007/s11571-021-09669-y.
- [14] J. Hong, W. Wang, and L. Najafizadeh, “ChatBCI, a P300 speller BCI with context-driven word prediction leveraging large language models, from concept to evaluation,” *Scientific Reports*, vol. 16, no. 1, p. 6379, 2026, ISSN: 2045-2322. DOI: 10.1038/s41598-025-25660-7.

Green Energy Transformation: Policy Reversals, Geopolitical Shocks, and the Digital Infrastructure Imperative

Mahmood Ahmad
Water, Energy and Food Policy Research Wing
RIZQ Trust
Lahore, Pakistan
e-mail: mahmood4404@gmail.com

Arifa Younas
Water, Energy and Food Policy Research Wing
RIZQ Trust
Lahore, Pakistan
e-mail: arifayounas490@gmail.com

Abstract— The global energy transition faces a major governance challenge: balancing rapid renewable energy expansion with policy reversals, geopolitical disruptions, and the growing energy demands of Artificial Intelligence (AI) data centers. These pressures risk diverting clean energy gains toward digital infrastructure while leaving energy-insecure nations vulnerable to fossil fuel dependence and unequal digital development. The urgency of this challenge is underscored by record-breaking global temperatures from 2023 to 2025, highlighting the escalating costs of delayed climate action. While previous studies have examined renewable deployment, policy changes, and data center sustainability separately, limited research has explored their combined impacts. This paper addresses that gap by analyzing four interconnected developments: the 2025 global renewable energy surge, which added 793 GW of new capacity; the rollback of U.S. clean-energy policies under the Trump administration; the 2026 Strait of Hormuz closure, which demonstrated that energy security and decarbonization can reinforce one another; and the increasing resource demands of AI-driven data center expansion. Based on these findings, the paper proposes a phased and enforceable policy framework that includes technology efficiency mandates, incentives for strategic data center siting, and automatic safeguards for developing countries. The framework is assessed through comparative case studies from China, India, the Nordic region, Bhutan, and Pakistan. The analysis concludes that energy security and decarbonization are fundamentally aligned objectives. Evidence from the Hormuz crisis showed that renewable-rich economies were more resilient to energy shocks, with Spain experiencing significantly lower electricity price increases than fossil fuel-dependent Singapore. The study also observed substantial gains in clean-energy equities following the crisis, reinforcing the strategic value of accelerating the renewable transition.

Keywords - Green energy transition; AI infrastructure; Machine learning model sharing.

I. INTRODUCTION

Recent years have underscored the urgency of accelerating the global transition toward low-carbon energy systems. According to the Copernicus Climate Change Service, the years 2023, 2024, and 2025 were among the warmest on record, with global average temperatures exceeding the 1.5°C benchmark established under the Paris

Agreement [5]. Climate observations further indicate that the Arctic is warming at a rate approximately three to four times higher than the global average, reflecting a persistent and systemic climate trend rather than a temporary fluctuation [5]. Scientific assessments have demonstrated that a warmer atmosphere retains greater quantities of moisture, increasing both the frequency and intensity of extreme weather events. The Intergovernmental Panel on Climate Change (IPCC) has repeatedly warned that continued warming beyond the 1.5°C threshold may increase the likelihood of irreversible environmental and socio-economic consequences [13].

At the same time, the global energy transition is experiencing both unprecedented progress and significant uncertainty. Rapid deployment of renewable energy technologies has strengthened confidence in the technical feasibility of decarbonization. However, political shifts, geopolitical disruptions, and emerging demands from digital infrastructure are creating new pressures that could reshape the trajectory of the transition. Four interconnected developments are particularly influential.

First, renewable energy deployment reached record levels during 2025, reflecting accelerating investment in solar, wind, and associated clean-energy technologies. Second, changes in United States energy policy introduced uncertainty regarding the long-term stability of climate and clean-energy commitments. Third, geopolitical tensions surrounding the Strait of Hormuz highlighted the continuing vulnerability of global energy markets to supply disruptions and price volatility. Fourth, the rapid expansion of Artificial Intelligence (AI) and large-scale data centers has generated growing concerns regarding electricity consumption, water use, and infrastructure sustainability.

Although these developments are frequently discussed within separate policy and academic domains, their interactions have received comparatively limited attention. Existing studies typically focus on renewable energy expansion, energy security, climate policy, or digital infrastructure as independent subjects. This separation obscures the extent to which these forces collectively influence decarbonization strategies, investment decisions, and long-term energy resilience. The present study

addresses this gap by examining these dynamics within a unified analytical framework.

The primary objective of this research is to evaluate how renewable-energy expansion, policy reversals, geopolitical risks, and AI-driven infrastructure growth interact to shape the future of the global energy transition. Building upon comparative evidence from multiple national and regional case studies, the paper develops a policy framework intended to balance climate objectives, energy security, economic competitiveness, and sustainable digital development.

Several limitations should be acknowledged. The analysis is based primarily on publicly available datasets, institutional reports, and secondary literature rather than original field observations. Furthermore, ongoing developments in global energy markets, international politics, and AI technologies may influence future outcomes beyond the scope of the present assessment. While the selected case studies provide valuable insights, they are not intended to represent the full diversity of regional experiences.

The remainder of this paper is organized as follows. Section II examines the four principal dynamics currently shaping the global energy transition, including renewable-energy deployment, policy change, geopolitical risk, and digital infrastructure expansion. Section III evaluates the economic rationale supporting accelerated energy transformation. Section IV presents evidence from circular-economy initiatives and international case studies. Section V proposes a phased policy framework for sustainable digital infrastructure development. Finally, Section VI summarizes the principal findings and outlines policy recommendations for governments, regulators, and industry stakeholders.

II. FOUR DYNAMICS SHAPING THE ENERGY TRANSITION

A. The 2025 Renewable Surge

The global renewable energy sector is projected to experience an unprecedented expansion in 2025, marking a critical milestone in the global transition toward clean energy systems. Approximately 793 GW of new renewable energy capacity is expected to be added worldwide during the year, representing more than 90% of all new power generation capacity. This trend reflects the accelerating dominance of renewable technologies within the global energy mix [7].

Solar energy remains the primary driver of this growth. According to Ember Energy's *Global Electricity Review 2025*, around 380 GW of solar capacity was installed in the first half of 2025 alone, representing a 64% increase compared to the same period in the previous year [7]. This rapid acceleration highlights the increasing global adoption of solar technologies driven by cost reductions and policy support.

The scale and pace of this expansion demonstrate not only strong international commitment to decarbonization but also the structural transformation of global electricity systems. Rather than merely supplementing fossil fuels, renewable energy is increasingly replacing conventional generation sources, signaling a fundamental shift in energy system architecture.

In addition, the International Energy Agency projects that between 2025 and 2030, approximately 4,600 GW of additional renewable capacity will be deployed globally [11]. This projection builds on current momentum and underscores the long-term trajectory toward renewable dominance in global electricity generation.

Table 1 summarizes key indicators of global renewable capacity for 2024 and 2025, illustrating the scale of this transition and its implications for future energy system development.

TABLE I. GLOBAL RENEWABLE CAPACITY: KEY METRICS (2024–2025)

Metric			
	2024 (Actual)	2025 (Projected)	Source
Global New Renewable Capacity Added	585 GW	793 GW	(House of Commons Library 2026) [10]
Share of All New Power Generation	92.5%	>90%	(House of Commons Library 2026) [10]
China Solar Installed Base	~750 GW	>1,000 GW	(House of Commons Library 2026) [10]
India Annual Wind+Solar Build	~25 GW	40–45 GW	(House of Commons Library 2026) [10]

China surpassed 1,000 GW of installed solar capacity in 2025, following an additional 210 GW added within six months, exceeding the total installed capacity of the United States as of the end of 2024 [7]. India also recorded strong growth, installing 31.25 GW of solar and wind capacity between April and November 2025, with projections indicating a total annual addition of 40–45 GW, supported by competitive auctions and declining technology costs [7]. Despite this rapid expansion, a key concern remains whether renewable energy systems can consistently deliver investment returns that are competitive with fossil fuel-based generation in the long term.

B. U.S. Policy Reversal and the Geothermal Exception

The 2025, United States energy policy overall significantly altered the trajectory of the clean energy transition by rolling back several Inflation Reduction Act (IRA) provisions. Key measures included the removal of wind and solar tax incentives, suspension of federal

renewable leasing programs, and the revocation of Electric Vehicle (EV) mandates. Collectively, these actions reduced projected clean energy growth by approximately half by 2035 [9]. In addition, the repeal of the Environmental Protection Agency's greenhouse gas finding created regulatory uncertainty, contributing to the delay or cancellation of over \$14 billion in clean energy investments. Despite this policy contraction, Enhanced Geothermal Systems (EGS) emerged as a notable exception. EGS utilizes established oil and gas drilling technologies, including hydraulic fracturing and horizontal drilling, to create engineered geothermal reservoirs in diverse geographic settings. The U.S. Department of Energy estimates that geothermal energy could eventually generate up to three times the output of the current U.S. nuclear fleet by 2050 [26]. Major technology firms, including Google and Meta, have increasingly invested in geothermal ventures to secure stable, 24/7 carbon-free electricity for Artificial Intelligence (AI) data centers, which require continuous and reliable power supply that intermittent renewables alone cannot consistently guarantee [29]. This emerging "triple advantage"—clean generation, baseload reliability, and subsurface energy storage—positions geothermal energy as a complementary pillar within future low-carbon energy systems.

C. The Hormuz Geopolitical Scenario

The Strait of Hormuz facilitates the transit of approximately 20–21 million barrels of oil per day, representing nearly one-fifth of global oil consumption and serving as a critical supply route for Asian economies. In 2026, escalating geopolitical tensions culminated in a major disruption of this corridor, significantly reducing global energy flows [10]. As a result, oil supply through the strait declined sharply from over 20 million barrels per day in February to approximately 3.8 million barrels per day by early April, marking one of the most severe disruptions in modern energy markets.

The economic consequences were immediate and substantial. Global crude oil prices surged toward \$150 per barrel, while refined fuel prices in key trading hubs such as Singapore exceeded \$290 per barrel. According to the International Energy Agency, the crisis could increase global fuel expenditures by \$1–2 trillion in 2026, reflecting higher costs rather than increased consumption [12]. In addition, disruptions extended to Liquefied Natural Gas (LNG) and fertilizer markets, affecting approximately 20% of LNG trade and one-third of global fertilizer flows, with disproportionate impacts on import-dependent Asian economies [12].

However, the crisis also highlighted structural differences in energy system resilience. Countries with higher renewable penetration demonstrated greater price stability. For example, Spain, with approximately 57% renewable electricity generation, experienced significantly

lower post-disruption electricity price increases compared to fossil fuel-dependent systems such as Singapore. Financial markets also responded to these structural signals, with clean energy Exchange-Traded Funds (ETFs) recording gains between 54% and 118% following the disruption [33]. These outcomes underscore a key systemic insight: renewable-based systems reduce exposure to fossil fuel volatility and improve macroeconomic resilience during supply shocks.

Higher fossil fuel prices may also accelerate electrification trends, including electric vehicle adoption and grid storage expansion. Historically, major energy shocks, such as the 1970s oil crisis and the 2021 Texas grid failure, have reinforced structural shifts toward diversified and electrified energy systems.

D. The Digital Infrastructure Paradox

The rapid expansion of Artificial Intelligence (AI) infrastructure presents a growing tension between digital growth and sustainability objectives. The United States alone hosts more than 5,400 data centers, accounting for nearly half of global capacity [23]. Goldman Sachs projects that AI-driven workloads will increase data center electricity demand by approximately 165% by 2030 (Goldman Sachs 2025), while McKinsey estimates sustained annual capacity growth of approximately 33% through the same period [18]. Data center energy demand is expected to double between 2024 and 2026. Northern Virginia, the world's largest data center hub, has already reached vacancy rates below 1%, reflecting extreme infrastructure concentration [3]. A multi-state analysis indicates that data center operations imposed approximately \$4.3 billion in system-wide electricity costs in 2024, much of which is redistributed to non-digital consumers through rate structures [6].

Water consumption presents an additional constraint. Large-scale facilities can consume up to 5 million gallons of water per day, comparable to the needs of medium-sized urban populations [28]. In Texas, projected water demand from data centers may reach 399 billion gallons annually by 2030, equivalent to approximately 6.6% of statewide supply [25]. In Loudoun County, Virginia, freshwater consumption has increased by 250% since 2019 due to rapid data center expansion [17].

These trends highlight a structural paradox: while AI systems drive productivity gains, their physical infrastructure imposes increasing pressure on electricity, water, and land systems. Table 2 summarizes these risks and corresponding policy responses discussed in Section VI.

TABLE II. AI DATA CENTER SUSTAINABILITY RISKS AND POLICY RESPONSES

Risk Category			
	Current Impact	Projected (2030)	Policy Response
Electricity Demand	Data center power to double by 2026 [5]	+165% vs 2024	Dedicated generation

		baseline [13]	>50 MW (Sec. VI-B)
Water Consumption	Up to 5M gal/day per large facility [28]	Texas alone: 399B gal/yr [25]	Zero-water cooling mandate (Sec. VI-A)
Grid Cost Burden	\$4.3B in 7-state ratepayer costs (2024) [11]	Accelerating without reform	Citizens-first power allocation
Developing Nation Exposure	No binding safeguards exist	Digital colonialism risk	Automatic protections (Sec. VI-D)
Carbon Footprint	Grid-tied growth in fossil-fuel regions	Structural lock-in risk	PUE ≤ 1.2 ; co-located renewables

III. ECONOMIC CASE FOR THE ENERGY TRANSITION

The economic rationale for the energy transition is supported by converging arguments related to cost reduction, risk mitigation, and long-term competitiveness. These drivers have strengthened significantly since 2020, with the primary barrier being institutional inertia rather than economic feasibility.

Renewable and nuclear energy sources are increasingly cost-competitive with new fossil fuel generation. Since 2010, the Levelized Cost Of Electricity (LCOE) for utility-scale solar photovoltaic and onshore wind has declined by more than 85% and 70% respectively, making them among the lowest-cost options for new electricity generation in many markets [15]. However, these comparisons refer strictly to generation costs and should not be confused with retail electricity prices, which also include transmission, distribution, and policy-related costs.

Operational efficiency improvements further strengthen the economic case. Zero-water cooling technologies significantly reduce long-term operational expenditures, while deployment in cold-climate regions can reduce cooling energy demand by 30–40% [11]. In addition, waste heat recovery from data centers can generate supplementary revenue streams, as demonstrated by Eco Datacenter in Sweden [11]. Energy independence also improves price stability by reducing exposure to fossil fuel volatility. Furthermore, green compute infrastructure increasingly benefits from Environmental, Social, and Governance (ESG) investment premiums, attracting institutional capital inflows. A critical but underexplored dimension of efficiency lies in machine learning model sharing. Requiring accessibility of large AI models across organizations—similar to open-source software frameworks—could reduce redundant training processes, lower cumulative energy consumption, and accelerate technology diffusion in developing economies. In parallel, introducing mandatory AI model efficiency standards, analogous to vehicle fuel economy regulations, would create structured incentives for computational efficiency and energy-aware AI development.

IV. CIRCULAR ECONOMY EVIDENCE

Empirical evidence from multiple countries demonstrates that sustainable energy transitions are both economically viable and operationally scalable. Table III presents selected case studies illustrating circular economy applications in energy systems.

TABLE III. CIRCULAR ECONOMY CASE STUDIES IN SUSTAINABLE ENERGY TRANSITION

Location			
	Model	Key Innovation	Outcome
Falun, Sweden (Eco Datacenter)	Climate-positive data center	Server waste heat (30–32°C) piped to district heating & absorption chillers	Net carbon-negative operation; waste heat monetized as revenue
Bhutan	Sovereign renewable model	100% hydropower; state-owned Bitcoin mining (>\$780M revenue)	~33% of GDP; public salaries funded; full domestic value capture
Lancaster, California	Streamlined solar permitting	Permitting time reduced from 6 months to 45 minutes	Unemployment : 17% (2009) to 6% (2023)
Wunsiedel, Germany	Rural sector coupling	Surplus wind/solar integrated with forestry waste processing	Functioning circular rural economy
Norway (Oslo/Trondheim)	Zero-emission urban infrastructure	Zero-emission construction sites; Powerhouse building generates 2× its consumed energy	Net energy surplus supplied to local microgrid

Denmark's North Sea energy island project further illustrates large-scale renewable integration. Upon completion, it is expected to supply electricity to approximately 10 million households, representing a major advancement in offshore energy infrastructure [8].

Collectively, these examples demonstrate that the transition to sustainable energy systems does not necessarily increase costs. Instead, it restructures energy production economics by improving efficiency, enabling revenue diversification, and strengthening system resilience.

V. POLICY FRAMEWORK FOR SUSTAINABLE DIGITAL INFRASTRUCTURE

The economic and operational dynamics outlined above highlight the need for a phased regulatory framework for sustainable digital infrastructure. This framework integrates empirical insights from China, India, the Nordic region, Bhutan, and Pakistan, as discussed in earlier sections. The proposed framework is summarized in Table IV.

TABLE IV. PHASED POLICY FRAMEWORK FOR SUSTAINABLE DATA CENTER INFRASTRUCTURE

PHASE			
	<i>Time frame</i>	<i>Mandate</i>	<i>Expected Outcome</i>
A- Technology	2025–2027	Zero-water cooling, PUE ≤ 1.2 , heat recovery	15% GHG/energy reduction; eliminate water stress
B – Energy Independence	2025–2030	Dedicated generation for facilities >50 MW	Eliminate grid competition with residents
C– Geographic Rationalization	2027–2035	Tax credits for $T < 10^{\circ}\text{C}$ regions; carbon/water pricing	30–40% cooling energy reduction
D– Developing Nation Protections	Immediate	Auto-safeguards: zero-water, on-site gen >20 MW, revenue sharing	Prevent digital colonialism; ensure local benefit
E – Model Efficiency Standards	2026 onward	AI model efficiency analogous to vehicle fuel economy	Reduce computation demand growth; lower energy baseline

A. Technology Mandates (2025–2027)

Zero-water cooling should be mandatory for data centers operating in water-stressed regions (below 1,700 m³ per capita annually). Microsoft’s chip-level cooling systems demonstrate up to 15% reductions in energy and emissions while eliminating water consumption [19].

Power Usage Effectiveness (PUE) standards should be capped at 1.2 for new facilities, with legacy systems required to reach 1.3 by 2030. Leading hyperscale facilities operated by Google and Microsoft already achieve PUE values near 1.1 [19].

Heat recovery systems must be integrated into all new facilities, enabling reuse in district heating, agriculture, or industrial processes, as demonstrated in Nordic deployments [11].

B. Energy Independence Framework (2025–2030)

Facilities exceeding 50 MW should be required to secure dedicated generation capacity rather than relying on residential grids. Options include Small Modular Reactors (SMRs), co-located renewable systems, or hybrid renewable-storage configurations [11]. This approach

reduces social conflict and prevents competition between industrial and residential electricity demand.

C. Geographic Optimization (2027–2035)

Incentives should prioritize deployment in regions with average annual temperatures below 10°C. Cold-climate siting reduces cooling demand by 30–40% while improving system efficiency [11]. Priority regions include Iceland, Norway, Finland, Canada, and northern Europe.

D. Developing Country Protections

Countries with limited water availability or energy deficits should be subject to automatic safeguards, including mandatory zero-water cooling, on-site energy generation requirements, public water-use monitoring, and revenue-sharing mechanisms. Development financing institutions should enforce compliance to prevent asymmetric resource extraction and ensure equitable development outcomes [21].

E. AI Model Efficiency and Sharing

Mandatory efficiency standards for AI systems should be introduced, similar to Corporate Average Fuel Economy (CAFE) standards in transportation. In addition, large-scale AI models should be subject to structured sharing frameworks to reduce redundant training, lower aggregate energy consumption, and improve access for developing economies [18].

V. CONCLUSION

In 2025, global renewable capacity additions reached approximately 793 GW, representing over 90% of all new electricity generation [7]. While this marks significant progress compared to 2024 levels, current deployment remains insufficient to meet long-term climate stabilization pathways.

Policy uncertainty in major economies continues to constrain investment, while AI-driven infrastructure expansion introduces new structural demand pressures [9]. However, evidence from recent geopolitical disruptions indicates that renewable-intensive systems exhibit greater resilience to energy price shocks and supply chain instability [10].

The central finding of this study is that energy security and decarbonization are not competing objectives but structurally aligned systems outcomes. The limiting factor is not technological capability but governance design and policy coordination.

ACKNOWLEDGMENT

The authors acknowledge the RIZQ Water, Energy and Food Policy Research Wing (WEF PRW) at Lahore, Pakistan of the support.

REFERENCES

- [1] BloombergNEF, "Energy Transition Investment Trends 2025." New York, NY, USA: Bloomberg Finance L.P., 2025.

- [2] Bhutan Ministry of Finance, "Digital Economy and Sovereign Wealth Strategy." Thimphu, Bhutan: Royal Government of Bhutan, 2024.
- [3] CBRE Group, Inc., "Global Data Center Trends 2025," 2025. [Online]. Available: <https://www.cbre.com/insights/reports/global-data-center-trends-2025>. Accessed: May 2026.
- [4] City of Lancaster, "Solar Permitting Reform: Economic Impact Assessment." Lancaster, CA, USA: Office of the Mayor, 2023.
- [5] Copernicus Climate Change Service, "Global Temperature Monitoring and Attribution." Reading, U.K.: European Centre for Medium-Range Weather Forecasts (ECMWF), 2025. [Online]. Available: <https://climate.copernicus.eu>. Accessed: May 2026.
- [6] Energy Innovation Policy & Technology, "Data Center Cost Allocation Analysis: Seven-State Study." San Francisco, CA, USA: Energy Innovation, 2024.
- [7] Ember, "Global Electricity Review 2025." London, U.K.: Ember Climate, 2025. [Online]. Available: <https://ember-climate.org>. Accessed: May 2026.
- [8] European Commission, "European Green Deal: Progress Report 2025." Luxembourg: Publications Office of the European Union, 2025.
- [9] Goldman Sachs Research, "AI to Drive 165% Increase in Data Center Power Demand by 2030." New York, NY, USA, 2025. [Online]. Available: <https://www.goldmansachs.com/intelligence/pages/ai-is-poised-to-drive-165-increase-in-data-center-power-demand.html>. Accessed: May 2026.
- [10] House of Commons Library, "Israel/US–Iran Conflict 2026: Reopening the Strait of Hormuz Research Briefing," Apr. 24, 2026. [Online]. Available: <https://commonslibrary.parliament.uk/research-briefings/cbp-10636/>. Accessed: May 2026.
- [11] International Energy Agency (IEA), "Renewable Energy and Data Centers: Market Analysis and Outlook to 2030." Paris, France, 2025. [Online]. Available: <https://www.iea.org/reports>. Accessed: May 2026.
- [12] International Energy Agency (IEA), "Oil Market Report – April 2026," 2026. [Online]. Available: <https://www.iea.org/reports/oil-market-report-april-2026>. Accessed: May 2026.
- [13] Intergovernmental Panel on Climate Change (IPCC), "Climate Change 2023: Synthesis Report." Geneva, Switzerland: IPCC, 2023.
- [14] Intergovernmental Panel on Climate Change (IPCC), "Climate Change 2022: Mitigation of Climate Change (AR6 WGIII)." Cambridge, U.K.: Cambridge Univ. Press, 2022.
- [15] International Renewable Energy Agency (IRENA), "Renewable Power Generation Costs in 2024." Abu Dhabi, UAE, 2025. [Online]. Available: <https://www.irena.org/publications>. Accessed: May 2026.
- [16] City of Lancaster, "Solar Permitting Reform: Economic Impact Assessment." Lancaster, CA, USA, 2023.
- [17] Loudoun County Water Authority, "Annual Water Usage Report: Data Center Impact Analysis," 2024. [Online]. Available: <https://www.loudounwater.org>. Accessed: May 2026.
- [18] McKinsey & Company, "AI-Ready Data Centre Infrastructure: Growth Projections 2023–2030." New York, NY, USA: McKinsey Global Institute, 2024.
- [19] Microsoft Corp., "2024 Environmental Sustainability Report: Zero-Water Data Center Design." Redmond, WA, USA, 2024. [Online]. Available: <https://www.microsoft.com/sustainability>. Accessed: May 2026.
- [20] Pakistan Ministry of Energy (Power Division), "National Power System Update: Circular Debt and Capacity Payments." Islamabad, Pakistan, 2024.
- [21] Rocky Mountain Institute (RMI), "The Economics of Clean Energy Transition in Emerging Markets." Basalt, CO, USA, 2024.
- [22] Royal Government of Bhutan, "Digital Economy Development Strategy." Thimphu, Bhutan: Gross National Happiness Commission, 2024.
- [23] Statista, "Data Centers Worldwide by Country, 2025," 2025. [Online]. Available: <https://www.statista.com>. Accessed: May 2026.
- [24] Texas Water Development Board (TWDB), "Projected Water Demand: Data Center Sector Analysis." Austin, TX, USA, 2025.
- [25] U.S. Department of Energy, "Data Center Energy Consumption Projections 2024–2028." Washington, DC, USA, 2024.
- [26] U.S. Department of Energy, "Enhanced Geothermal Systems: Technology and Market Projections." Washington, DC, USA, 2024.
- [27] U.S. Energy Information Administration, "Strait of Hormuz: World Oil Transit Chokepoints," 2024. [Online]. Available: https://www.eia.gov/international/analysis/special-topics/World_Oil_Transit_Chokepoints. Accessed: May 2026.
- [28] U.S. Geological Survey, "Industrial Water Use in the United States: Data Center Facilities." Reston, VA, USA, 2024.
- [29] V. Vaitheeswaran, "Geothermal Energy and the U.S. Clean Power Transition," "The Economist," London, U.K., 2025.
- [30] W. Stadtwerke, "Sector Coupling and Circular Economy in Wunsiedel: Annual Sustainability Report," 2024.
- [31] M. Ahmad, H. Isaad, and F. A. Akhtar, "Market dynamics of Pakistan's energy transition," "IntechOpen," 2025. doi: 10.5772/intechopen.1009235.
- [32] M. Ahmad, Ed., "Water pricing, demand management, and allocative efficiency," in "Water Policy in Pakistan." Cham, Switzerland: Springer, 2023, ch. 11. doi: 10.1007/978-3-031-36131-9_11.
- [33] Morningstar, "Hormuz Crisis Shows Clean Energy Is Key to Reducing the Economic Cost of Fossil Fuel Dependence," "PR Newswire," May 14, 2026. [Online]. Available: <https://www.morningstar.com/news/pr-newswire/20260514ln59938/hormuz-crisis-shows-clean-energy-is-key-to-reducing-the-economic-cost-of-fossil-fuel-dependence>. Accessed: May 2026.

Driving Digital Transformation in Precision Engineering: A Tooling Management Case Study

Shivam Sharma¹, Mohamed Kara-Mohamed¹, Andrew Naylor¹, Phillip Taylor², and David Eadie²

¹School of Engineering and Built Environment, Liverpool John Moores University
Liverpool, UK

email: s.n.sharma@ljmu.ac.uk

²Bryken Ltd, Randles Road, Knowsley Industrial Park South
Knowsley, Liverpool, UK

email: pt@bryken.com; de@bryken.com

Abstract—The precision engineering sector faces increasing pressure to enhance efficiency, traceability, and cost control, which are all challenges compounded by legacy systems and manual processes. This case study examines the digital transformation initiative of a UK-based Computer Numerical Control (CNC) precision engineering company, aimed at modernising tooling management. By leveraging low-code platforms, Azure cloud services, and real-time analytics through Power BI (Business Intelligence), fragmented paper-based systems are successfully replaced by integrated digital systems. The new system includes QR-coded tools, automated workflows, and centralised dashboards that enable traceability. This in turn enables actionable insights and continuous improvement. Some of the key outcomes of this transformation include improved data accuracy, reduced downtime, better cost tracking, and enhanced operator engagement. The initiative not only delivers immediate operational gains but also lays a foundation for advanced capabilities such as predictive maintenance and Internet of Things (IoT) integration. This transformation demonstrates how Small to Medium-sized Enterprises (SMEs) can adopt scalable, cost-effective technologies to drive innovation and competitiveness in a rapidly evolving manufacturing landscape. Furthermore, this case study highlights a scalable and generalisable approach that can inspire other companies and sectors in engineering to accelerate their own digital transformation journeys.

Keywords—digital transformation; tooling management; Industry 4.0; CNC precision engineering; Power Platform.

I. INTRODUCTION

The manufacturing industry is undergoing a significant transformation driven by the need for higher efficiency, stringent quality assurance, and optimised operational costs. Precision engineering and CNC machining operations are challenged by complex tooling inventories, the need for traceability, and inefficiencies arising from manual tracking processes.

This paper addresses the following research questions: (1) How can low-code and cloud-based platforms digitise tooling management in a CNC precision engineering SME? (2) What quantitative improvements in accuracy and downtime result from this transformation? (3) How reproducible and generalisable is the proposed solution?

The emergence of Industry 4.0 technologies presents new opportunities to address these concerns [2]. Highly cited studies underscore the importance of data-centred approaches, such as Enterprise Resource Planning (ERP), IoT-enabled tool tracking, and real-time visualisation platforms [1]. Traditional paper-based workflows are increasingly recognised as

impediments to productivity, contributing to errors, delays, and reduced responsiveness [3]. In contrast, cloud-based platforms, process automation, and integrated analytics redefine best practices in the sector.

This paper presents a case study from Bryken Ltd (a precision engineering company, Knowsley, UK). The case study details the company's digital transformation initiative focused on tooling management, covering streamlining of data capture, integration of disparate systems, and improvement of operational visibility, ultimately positioning Bryken for long-term resilience, innovation, and competitiveness.

The remainder of this paper is organised as follows. Section II presents the case study, describing the company context, the legacy tooling process and improvement criteria, the methodology and technologies adopted, the key performance indicators used for evaluation, and the resulting operational and financial impact. Finally, Section III concludes the work and outlines directions for future development.

II. THE CASE STUDY

A. Company Overview

Bryken Ltd specialises in the production of high-tolerance components for a range of industries including oil and gas, automotive, defence, and healthcare. The company operates a suite of advanced CNC machines and manages an extensive inventory of high-performance cutting tools sourced from multiple suppliers. Historically, Bryken relied on a combination of paper forms, spreadsheets, and manual operator input to track tool vending, usage, and return. This decentralised approach results in inefficiencies, frequent data discrepancies, and limited visibility into the real-time status of tooling across the shop floor.

The digital transformation initiative is launched to modernise tooling processes by implementing an integrated digital system capable of real-time data acquisition, automated reporting, and end-to-end visibility. The initiative seeks to eliminate manual paperwork, reduce tool-related downtime, improve inventory accountability, and enhance the accuracy of cost allocation per manufacturing order. This project is not just a technological upgrade, but a strategic move to future-proof Bryken's operational model.

The project is also aligned with Bryken's vision of moving toward data-driven manufacturing, where prior adoption of data-driven methodologies has been shown to benefit just-in-time production and continuous improvement metrics [4].

B. Current Tooling Process and Improvement Criteria

To effectively address the tooling challenges, we began by conducting a comprehensive audit of the existing workflows, systems, and information processes. In the context of CNC precision engineering, the tooling process refers to the management of cutting tools used in machining operations, including their vending, allocation to machines, monitoring of usage, and return. Under the legacy approach, the process began with operators using paper-based first-off sheets to record tool usage for each new job. This manual documentation served as the initial point of reference for tool deployment. Next, operators would manually input job numbers associated with each tool transaction, linking the paper records to specific machining operations. Information from these sheets and operator entries was then fed into fragmented supplier reports, requiring supervisors to reconcile the data manually. Because there was no centralised or automated link between these steps, discrepancies could easily arise, from mismatches in operator entries to gaps between shop floor records and supplier data, making it difficult to maintain real-time accuracy and accountability throughout the tooling workflow. This resulted in frequent errors in job identification, lack of reliable cost allocation, and poor visibility of tool availability. Such inefficiencies affected the entire engineering process. For

example, production downtime increased when tools were unavailable or unaccounted for, supervisors and procurement staff spent significant time reconciling data manually, and management struggled to forecast costs accurately. Collectively, these limitations reduced productivity, inflated operational costs, and undermined the company's competitiveness.

Figure 1 summarises the manual processes Bryken Ltd relied upon for tooling management before the implementation of the digital transformation initiative. As the figure shows, the workflow for tool usage and inspection followed a fully manual process. Step 1 involved recording first-off inspection results on paper sheets, with no digital linkage to job tracking systems. Step 2 required operators to manually enter job numbers, which often resulted in transcription errors and inconsistencies. Step 3 saw supervisors reviewing supplier tool usage reports by hand, delaying feedback and limiting visibility of performance data. Step 4 highlights the lack of automated data collection or cost tracking, making it difficult to monitor tool inventory and allocate costs accurately. Altogether, the figure illustrates the fragmented and inefficient nature of the legacy system, providing context for the need to transition to a fully integrated, digital workflow that overcomes these specific operational limitations.



Figure 1. Flow diagram of the pre-project tooling process.

The enhancement process starts by establishing the functional and technical requirements based on input from production operators, engineering leads, and procurement personnel. The collected feedback states the need to centralise and standardise tool vending data from multiple external suppliers, integrate machine-level tool usage with operator tracking, and replace multiple paper-based forms with a single unified digital interface. The required system should also support real-time monitoring of tooling costs per manufacturing order to offer granular traceability for root cause analysis, and it should facilitate customisable reporting for production and quality assurance teams. Key design criteria include user-friendliness, scalability, data integrity, and interoperability with existing Enterprise Resource Planning (ERP) and database systems. Importantly, the system should cater to users with varying levels of digital literacy and should integrate

seamlessly with ongoing shop floor activities without introducing disruptions.

C. Methodology and Technology Used

The solution is implemented using a modular and scalable architecture based on Microsoft Power Platform and Microsoft Azure cloud computing services. As illustrated in Figure 2, the system follows a four-layer architecture: (1) Data Capture — QR codes on tools and first-off sheets at shop floor level; (2) Integration — Microsoft Power Automate for email parsing and workflow automation; (3) Storage — Microsoft SharePoint as the centralised data repository; (4) Analytics — Microsoft Power BI dashboards with custom Data Analysis Expressions (DAX) measures for real-time Key Performance Indicator (KPI) reporting. Several integrated systems and procedures have been implemented, as indicated below.

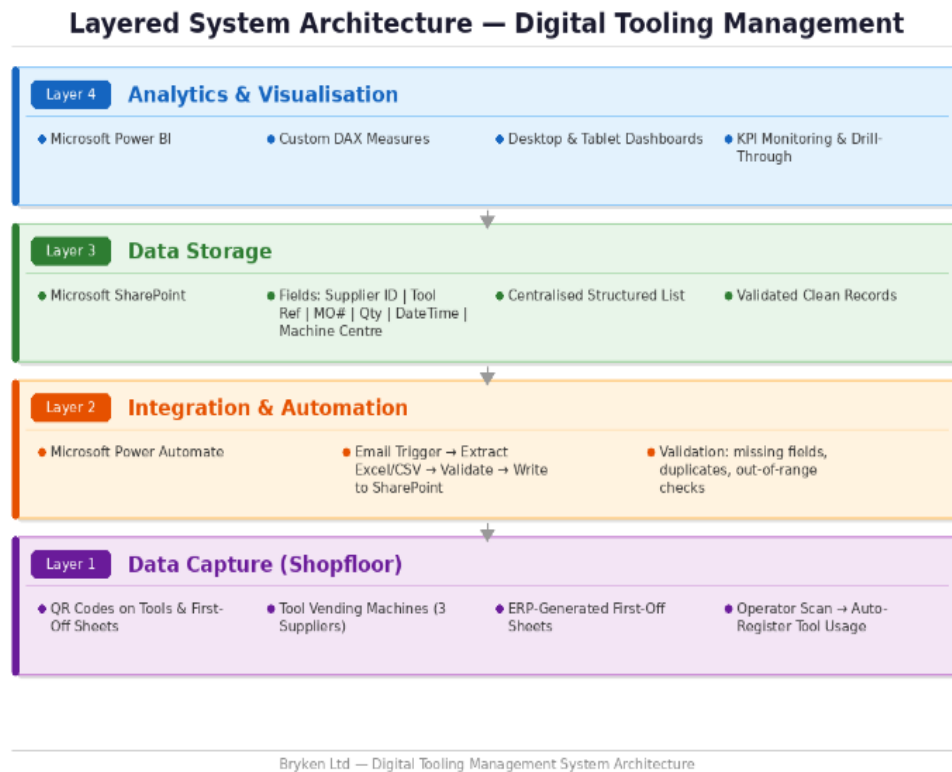


Figure 2. Layered system architecture of the digital tooling management solution.

C.1. Tooling QR Codes

First, QR codes are introduced to uniquely tag individual tools, enabling operators to scan and register tool usage with precision. Then, each first-off sheet incorporates a QR code which links each manufacturing order with its details containing tooling requirements, as shown in the example in Figure 3. The first-off sheet is a process quality document generated by the ERP system that accompanies every new manufacturing order. It contains inspection instructions, gauging requirements, and operator guidance. The sheet stays at the CNC machine that is used for the order and is kept handy until the manufacturing order is completed. Whenever tools are needed, operators scan the QR code of the sheet and the unique QR code of the required tool at the tool vending machine, which registers the tools used against that manufacturing order. In the legacy process, operators manually recorded job numbers and tool usage on the sheet, which often led to errors and inconsistencies. In the new system, the embedded QR codes allow operators and management to capture this information automatically by scanning, thereby eliminating manual entry, reducing errors, and standardising the recording of tool consumption across machines and shifts.



BRYKEN LIMITED		PROCESS QUALITY FORM	
D: 40	P: 0	ISSUE: 03	AUTHORITY: KEN RONAN
1000 ID: P05790055M		ISSUE: 7	MANUFACTURING ORDER ID: MO027503

Basket weight allowance of 15Kg means maximum number of parts per basket should not exceed 200

GAUGING REQUIRED AT MACHINE

VERNIER - MICROMETER
PLUG GAUGES - 5.35/5.65 + 3.00/3.05 + 6.20/6.50 + 7.75/8.05 25.40/25.45 (P1300) M14 X 1.5 SCREW RINGS (GO SR042 + NO-GO) PLAIN RING GAUGES PR 037 + PR 038 SURFTEST (USE CRANKED PROBE TO CHECK FINISH 0.8 FINISH ON 14.8 DEEP FACE) NOTE - 3.0/3.05 DIA IS CRITICAL CHECK SURFACE FINISH IS CORRECT THE FULL LENGTH OF THE HOLE !!! CHECK CONCENTRICITY: QA ALERT 17/08 & 23/05 TO BE ISSUED LIGHT SOURCE/LAMP MUST BE AT THE MACHINE TO IMPROVE VISUAL INSPECTION !!!!

CARE POINTS

i) INSPECT TO DRAWING HANDLE WITH CARE. PLACE IN BASKETS IN ROWS. CUSTOMER REJECTION INTERNAL AND EXTERNAL BURRS ON 8 X 5.5 CROSS HOLES 11/07/2017 CUSTOMER COMPLAINT - SWARF INDENTS ON LARGE O/D CUSTOMER RETURN - GROOVE PRESENT ON CRITICAL DIAMETER - 11/12/2023 REF QA 23/05 See over for historic non-conformance scrapage incident details so as to be aware...

OPERATIONAL FIRST OFF

DATE	OP	MACHINE	INSPECTION OPERATION	REMARKS	SIGN/INIT
------	----	---------	-------------------------	---------	-----------

Figure 3. An example of the new first-off sheet that incorporates a QR code for tool tracking.

C.2. Tooling Inventory Record System

A crucial aspect of the operational process involves accurately tracking tool usage on the shop floor. Bryken currently works with three tool suppliers, each managing their own vending machines installed on site. These vending machines are owned by the tool suppliers and include an inbuilt computer system that records all tool vending activity. Whenever an operator vends a tool, the transaction is automatically logged in the supplier's system.

To facilitate internal monitoring and analysis, the tool suppliers automatically generate daily tool usage reports and send them to Bryken via email in Excel or CSV format. This timely and well-structured reporting ensures a clear record of inventory movements, which is essential for process improvement and operational efficiency. In the previous process, Bryken's supervisors manually opened each supplier email, reviewed the attached spreadsheets, and checked the tool usage information. This approach was time consuming, prone

to human errors, and sometimes resulted in inconsistent data, making the information less reliable.

In the new system, the entire process has been automated. Power Automate detects incoming supplier emails, extracts the data from the attachments, applies validation checks, and uploads the verified data directly to SharePoint. The data schema captured per transaction includes Supplier ID, Tool Reference Code, Manufacturing Order Number, Quantity Vended, Transaction Date/Time, and Machine Centre ID. The Power Automate flow triggers on email receipt, extracts the Excel/CSV attachment, validates each row against defined field rules, and writes clean records to a structured SharePoint list. This automation removes manual work, ensures data accuracy, and provides structured information that is immediately available for analysis and reporting. Figure 4 illustrates the transformation from a manual, error-prone process of handling tooling reports to a streamlined and automated workflow using Power Automate.

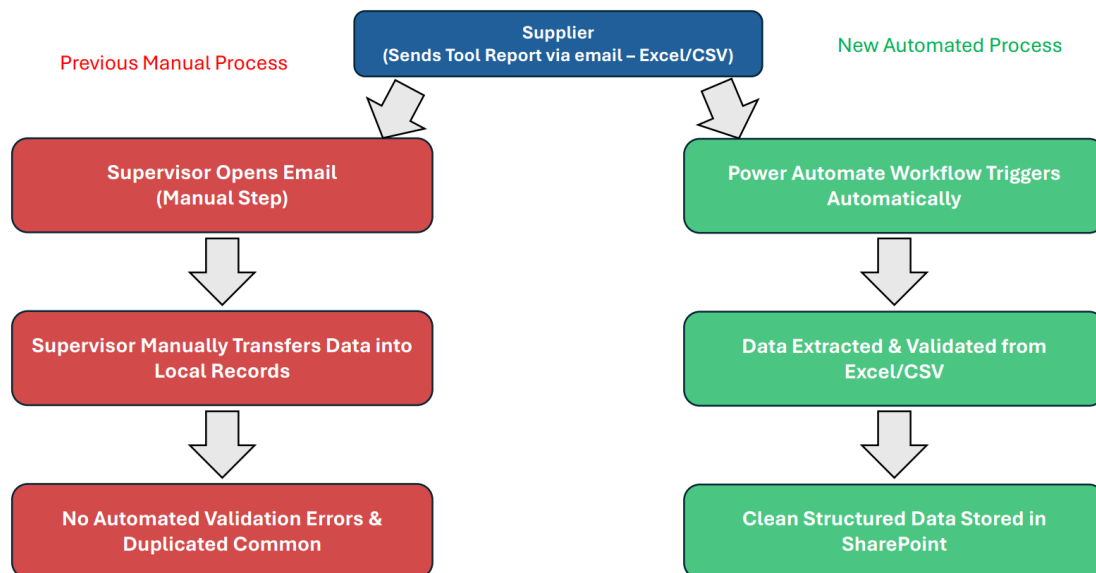


Figure 4. Tool usage report workflow (before vs. after).

C.3. Tooling Dashboard

Utilising information collected from the two procedures described above; a Tooling Dashboard is also introduced. Figure 5 illustrates an example of the new Power BI Tooling Dashboard, which provides an overview of tool usage and spending. The dashboard displays total tool expenditure, total manufacturing orders, quantities produced, product IDs, and the top five products with the highest tool costs. It also shows trends in tool usage over the past months. Dynamic filtering and

interactive drill-through capabilities allow supervisors, production planners, and quality engineers to investigate anomalies and analyse information by machine, product, or date. Advanced DAX measures were used to create tailored key performance indicators, ensuring that the dashboards are optimised for both desktop and shop floor tablet use, enabling quick, data-driven decision-making. The design draws on best practices from early adopters of intelligent manufacturing shop floors [5].

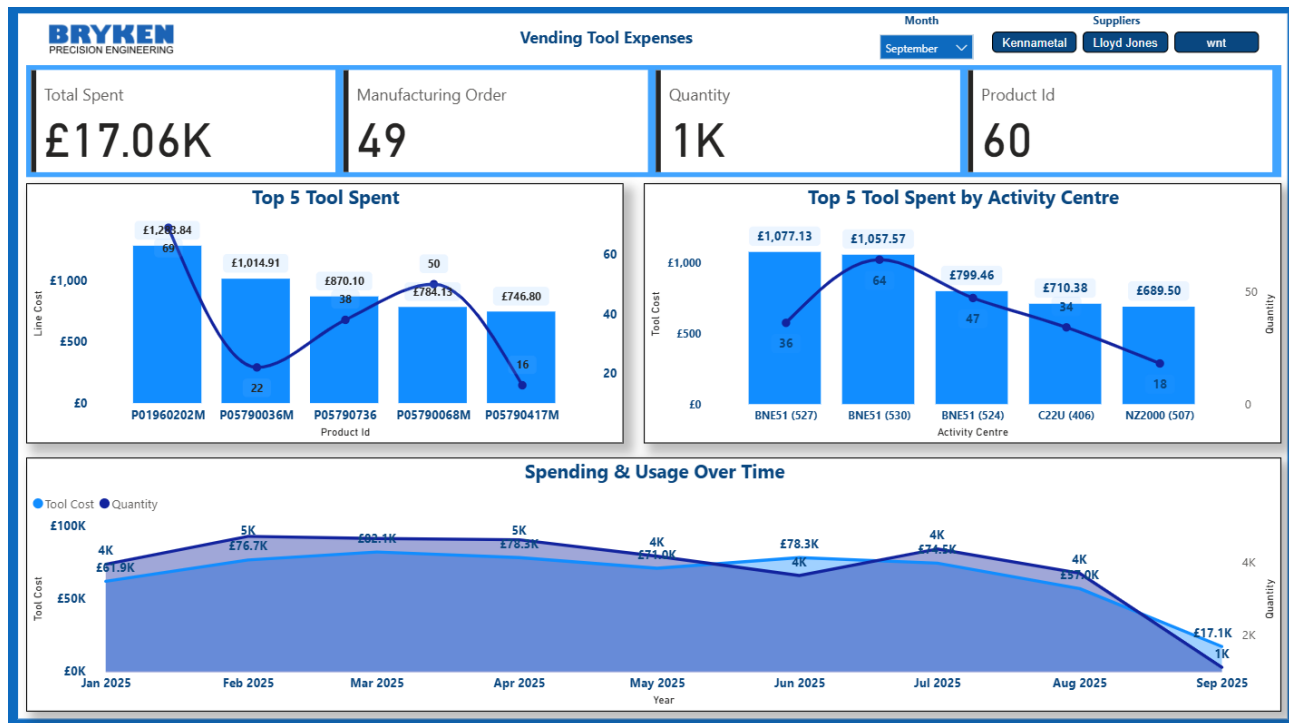


Figure 5. Overview of the Tooling Dashboard.

Figure 6 provides another screen from the Tooling Dashboard. The page provides a detailed breakdown of individual manufacturing orders, showing the tooling cost per product in both actual value and as a percentage of the selling

price. This level of detail enables planners and management to understand product-level profitability, identify high-cost items, and make informed decisions to optimise tooling usage and production efficiency.

Manufacturing Order	Product Id	Completion Date	Quantity	Quantity Made	Quantity Scrapped	Tooling Price Per Part	Selling Price	Tool Percentage	Actual Tooling Cost
MO025987	P01530014M	12 November 2025	1,294.00	1,294.00	0.00	£0.05	£11.20	0.47%	£68.50
MO028700	P05790754	12 November 2025	702.00	702.00	0.00	£0.04	£1.44	2.57%	£25.95
MO028659	P05790784	12 November 2025	533.00	533.00	0.00	£0.59	£7.37	7.95%	£312.47
MO028799	P05791306M	12 November 2025	41.00	32.00	9.00	£2.82	£500.00	0.56%	£90.18
MO028858	P05980142M	12 November 2025	590.00	567.00	23.00	£0.46	£2.87	15.87%	£258.24
MO028418	P05160024	11 November 2025	885.00	881.00	4.00	£0.33	£9.92	3.37%	£294.53
MO028718	P05160026	11 November 2025	2,700.00	2,535.00	0.00	£0.07	£3.02	2.22%	£169.72
MO028287	P05790055M	11 November 2025	7,556.00	7,468.00	88.00	£0.64	£4.89	13.16%	£4,805.29
MO028373	P05790081M	11 November 2025	5,000.00	4,096.00	0.00	£0.09	£1.33	6.83%	£372.12
MO028672	P05790651	11 November 2025	370.00	370.00	0.00	£0.36	£11.67	3.11%	£134.23
MO028706	P05790680M	11 November 2025	400.00	396.00	2.00	£1.11	£23.35	4.76%	£439.80
MO028701	P05790691	11 November 2025	2,332.00	2,332.00	0.00	£0.10	£1.26	8.18%	£240.40
MO028457	P05791054M	11 November 2025	350.00	318.00	8.00	£5.20	£47.77	10.88%	£1,653.36
MO028906	P06570052M	11 November 2025	322.00	322.00	0.00	£0.27	£8.06	3.37%	£87.54
MO028886	P06570053	11 November 2025	403.00	403.00	0.00	£0.14	£9.17	1.53%	£56.50
MO028784	P00640075M	10 November 2025	650.00	536.00	0.00	£0.14	£4.90	2.78%	£72.98
MO028565	P03730003M	10 November 2025	584.00	584.00	0.00	£0.44	£8.42	5.24%	£257.57
MO028113	P05790146	10 November 2025	2,500.00	2,329.00	0.00	£0.01	£1.91	0.58%	£25.89
MO028270	P05790552M	10 November 2025	377.00	370.00	7.00	£1.10	£8.94	12.36%	£408.73
MO028432	P05790598M	10 November 2025	204.00	178.00	26.00	£0.20	£14.58	1.38%	£35.93
MO028759	P05790661M	10 November 2025	1,126.00	1,110.00	16.00	£0.58	£2.27	25.47%	£641.86
MO028658	P05790762	10 November 2025	994.00	971.00	23.00	£0.29	£1.80	15.97%	£279.05
MO028578	P05790874M	10 November 2025	507.00	498.00	9.00	£0.22	£3.58	6.24%	£111.19
MO028661	P05791014M	10 November 2025	145.00	142.00	3.00	£2.60	£38.54	6.75%	£369.67
MO028461	P05791216	10 November 2025	255.00	250.00	5.00	£0.30	£8.78	3.38%	£74.25
MO028840	P05790126	07 November 2025	1,576.00	1,576.00	0.00	£0.14	£1.56	9.17%	£225.45

Month: January, February, March, April, May, June, July, August, September, October, November, December

Figure 6. The Tooling Dashboard: breakdown of the manufacturing cost per product.

D. KPI Definition and Measurement Methodology

The following KPIs were defined and measured to evaluate the transformation.

(1) Stock Discrepancy Rate: the percentage of daily tool transactions where the quantity in the supplier's system did not match Bryken's internal records. A baseline was established over three months before implementation by manually cross-

referencing supplier Excel reports with paper first-off sheets, while post-implementation data was extracted from SharePoint automatically. The rate is calculated as:

$$\text{Stock Discrepancy Rate} = (\text{mismatched records} / \text{total records}) \times 100\%.$$

The baseline was approximately 18%; after implementation it fell below 4.5%, representing a reduction of over 75%.

(2) Tool-Related Downtime: the production time lost due to tools being unavailable or incorrectly allocated. Shift supervisors logged downtime events prior to implementation, whereas afterwards these were tracked via dashboard alerts. A reduction of approximately 30% was observed within the first three months.

(3) Tooling Cost as a Percentage of Selling Price: calculated per manufacturing order from Power BI, enabling profitability analysis at the order level. These figures are based on operational observations from a single site over three months and should be treated as indicative rather than statistically validated benchmarks.

E. Results and Impact

The deployment of the new digital tooling system brought significant improvements across Bryken's operations. Tool usage data has become readily accessible, enabling shift supervisors and engineers to identify usage patterns, detect anomalies, and take proactive measures to avoid tool shortages. The elimination of paper-based documentation led to an immediate reduction in administrative overhead and freed operator time for production activities.

Real-time tracking of tool consumption and availability allows job initialisation to be optimised and minimises tool-related delays. Operators have become more engaged in the tooling process, supported by intuitive interfaces requiring minimal training. Procurement teams also benefit from accurate consumption data, improving forecasting, reducing over-ordering, and strengthening supplier negotiations. The quantitative gains in stock-discrepancy and downtime reduction reported in Section II-D translate directly into these operational benefits. It is important to note, however, that the automated system does not eliminate errors entirely: the remaining discrepancies mainly stem from unavoidable operational factors such as occasional human errors in tool withdrawals, misplacement of tools on the shop floor, or delays in data updates from supplier systems. Nonetheless, the system significantly reduces these issues and enhances overall process reliability.

Crucially, the system also enables the identification of potentially unprofitable orders before production begins. One manufacturing order, with a quantity of 150 units, was analysed and revealed that the tooling required would cost more than the revenue generated from the sale. This insight prompted the implementation of a minimum order quantity of 3,000 units for that product, immediately preventing production of the unprofitable batch and avoiding unnecessary material and tooling costs.

These improvements represent not only a major operational gain but also a substantial financial impact. Reducing stock

discrepancies ensures that procurement teams no longer over-order tools, preventing capital from being tied up in unnecessary inventory. Minimising downtime increases machine utilisation, translating into higher throughput and reduced lead times for customers. Accurate, real-time tooling data also helps the company provide better cost estimates and remain competitive.

The analysis of manufacturing orders shows that tooling consumption depends on the machining process involved. More complex operations, such as multi-axis CNC machining, generally require a higher number of tools and greater tool wear, whereas simpler processes such as turning or vertical milling tend to use fewer tools. This variation highlights how the type of machining directly influences tooling requirements and overall production costs, reinforcing the need for accurate, real-time tracking to capture these differences.

The analysis also shows that the initial fixed percentage-based allocation of tooling costs — where, during the quoting process, a constant percentage of manufacturing cost was allocated to tooling based on best-guess practice — can often be inaccurate, either overestimating or underestimating the actual requirements for a job. With real-time data now available, managers can avoid overstating profitability when tooling needs are lower than estimated, while also preventing unexpected cost overruns when requirements are higher. This improved visibility supports more accurate quoting, strengthens procurement planning, and ensures that tooling strategies are aligned with the true demands of each machining process.

Furthermore, the Tooling Dashboard enables management to conduct weekly and monthly reviews of tooling costs at the manufacturing-order level. This allows the team to compare actual tooling expenses against expected benchmarks, identify cost-saving opportunities, and drive continuous-improvement initiatives. Identifying excessive tool usage on repeat jobs, now monitored using the same dashboard infrastructure, has led to process standardisation and tool-life optimisation. The system also creates the foundation for future integrations, such as predictive maintenance and closed-loop feedback based on machine performance and tool-wear analytics.

In terms of cultural impact, the transformation has fostered a stronger data-driven mindset across the organisation. Staff at all levels now value accurate, real-time data and have embraced digital practices more readily. The success of this tooling project has become a catalyst for broader digital initiatives within the company, including paperless job sheets, digital document viewers, and live machine-monitoring systems. The improved transparency has also strengthened cross-functional collaboration between shop floor operators and engineering staff.

III. CONCLUSION AND FUTURE WORK

The digital transformation initiative at Bryken Ltd demonstrates how even SMEs can unlock substantial value through the strategic deployment of low-code and cloud-based technologies. By integrating real-time data, automating routine processes, and enhancing the visibility of critical operations, Bryken has not only improved its tooling efficiency but also elevated its overall operational maturity.

Looking ahead, Bryken plans to scale this tooling system further by incorporating IoT data for real-time tool-wear monitoring, implementing machine-learning models to predict tooling failures, and extending the digital platform to other consumables such as coolant and fixture components. There is also an ambition to link tooling data with quality metrics, enabling end-to-end traceability from raw material to final inspection.

Future work will also address reproducibility by publishing anonymised workflow schemas and SharePoint data models to support replication in similar SME contexts. Additionally, benchmarking against recent Industry 4.0 implementations in the literature [6][7] will strengthen the evidential basis of the reported improvements.

This case study illustrates that digital transformation does not require disproportionately high capital investment or complex proprietary systems. With thoughtful planning, a clear understanding of operational constraints, and strategic use of accessible technologies, manufacturers can position themselves for long-term resilience, agility, and competitiveness in an increasingly data-driven world. It is hoped that this example will encourage other SMEs in the precision engineering sector to embrace similar digital strategies and share their learnings with the wider manufacturing community.

REFERENCES

- [1] National Academy of Science and Engineering, "Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative Industrie 4.0: Final report of the Industrie 4.0 Working Group," 2013.
- [2] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Bus. Inf. Syst. Eng.*, vol. 6, pp. 239–242, 2014.
- [3] S. Wang, J. Wan, D. Li, and C. Zhang, "Implementing smart factory of Industrie 4.0: An outlook," *Int. J. Distrib. Sens. Netw.*, vol. 12, no. 1, 2016.
- [4] L. D. Xu, E. L. Xu, and L. Li, "Industry 4.0: State of the art and future trends," *Int. J. Prod. Res.*, vol. 56, no. 8, pp. 2941–2962, 2018.
- [5] R. Y. Zhong, C. Xu, C. Chen, and G. Q. Huang, "Big data analytics for physical internet-based intelligent manufacturing shop floors," *Int. J. Prod. Res.*, vol. 55, no. 9, pp. 2610–2621, 2017.
- [6] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for Industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 3, pp. 18–23, 2015.
- [7] T. Stock and G. Seliger, "Opportunities of sustainable manufacturing in Industry 4.0," *Procedia CIRP*, vol. 40, pp. 536–541, 2016.

Cascaded TSDF for Multi-Camera Volumetric Capture with Perceptual Mesh Reduction

Abdul Rehman and François Bouille

French Touch Factory

Paris, France

email: {abdul | f.bouille}@frenchtouchfactory.com

Abstract—Volumetric capture requires adaptive resolution: fine detail for faces/hands, coarse for bodies/legs. We present a 5-level cascaded Truncated Signed Distance Function (TSDF) that allocates resolution by perceptual importance, achieving uniform-fine quality (2.4mm head error) with 83% memory reduction (4.8GB vs 28.5GB) and 4× speedup in TSDF fusion (32 Frame Per Second (FPS) vs 8 FPS for fusion only). To enable production deployment, we introduce: (1) 8-neighbor outlier filtering rejecting 73% of artifacts pre-fusion, (2) two-tier reduction (voxel+quadric) achieving 30:1 compression while preserving cascade detail, and (3) normal-validated texture transfer preventing color bleeding. The pipeline processes 9-camera captures with 512ms reconstruction latency (1.95 Hz) for near-real-time preview and 850ms export latency (1.18 Hz), generating Visual Effects (VFX)-compatible Alembic files (50K triangles, 2048×2048 texture, 98% coverage) suitable for offline batch processing. The proposed pipeline enables efficient, production-scale volumetric video capture by combining adaptive reconstruction with robust geometry and texture processing, supporting both near-real-time preview and high-quality offline asset generation.

Keywords—volumetric capture; TSDF fusion; mesh decimation; texture reprojection; near-real-time reconstruction.

I. INTRODUCTION

High-performance volumetric capture of humans is essential for Virtual Reality (VR), gaming, and film production. The core challenge is integrating noisy depth data from multiple sensors into clean, watertight meshes while balancing temporal coherence, geometric detail, and practical memory constraints. KinectFusion [1] shows TSDF based fusion but produces high-polygon meshes unsuitable for direct commercial export. DynamicFusion [2] adds temporal coherence via non-rigid alignment, but requires offline post-processing. Commercial systems [3] produce high-quality results but demand 30+ controlled cameras in fixed studio environments. Deep learning approaches [4][5] achieve impressive completions but sacrifice the sub-millimeter geometric precision required for VFX pipelines.

In this paper, we address three main challenges in production-ready capture. First, we propose an 8-neighbor pre-fusion depth consistency check that eliminates silhouette bleeding before volumetric integration. Second, we introduce a two-tier mesh reduction combining voxel clustering with edge-aware quadric decimation using spatial weighting, achieving 30:1 compression while preserving perceptually important detail. Third, we propose normal-validated texture reprojection from the high-resolution mesh to the decimated mesh, preventing cross-surface color bleeding while maintaining 98% texture coverage.

Our primary contribution is a 5-level cascaded TSDF that allocates voxel resolution by perceptual importance: 2mm for head and hands, 3–4mm for shoulders, 6mm for torso, and 8mm for legs. We validate this through:

- **Cascade ablation:** comparing uniform-coarse (8mm), uniform-fine (2mm), and 3/4/5/6-level configurations. Our 5-level cascade matches uniform-fine quality (2.4mm vs. 2.3mm head error) with 83% memory reduction and 4× TSDF fusion speedup.
- **Production pipeline:** pre-fusion outlier filtering (73% artifact rejection), two-tier reduction (30:1 compression), and normal-validated texture transfer (98% coverage, 5.5 dB Peak Signal-to-Noise Ratio (PSNR) gain).
- **End-to-end deployment:** 500,000+ production frames exported to Alembic with a constant 2 GB RAM footprint.

The rest of the paper is structured as follows. Section II reviews related work. Section III describes the proposed pipeline. Sections IV–VI detail the TSDF fusion, mesh reduction, and texture reprojection methods. Section VII presents the results and Section VIII concludes the paper.

II. RELATED WORK

Volumetric fusion systems, such as KinectFusion and DynamicFusion [1][2], produce high-polygon meshes with millions of triangles requiring offline reduction, while commercial systems [3] demand 30+ cameras in fixed studio setups. Deep learning approaches [4]–[6] improve completeness but sacrifice geometric precision. A direct quantitative comparison against neural baselines is beyond the scope of this work, as our pipeline targets deterministic sub-millimeter precision required by VFX production constraints that current learning-based methods do not guarantee. Classical TSDF methods, such as Curless and Levoy [7], use weighted averaging but cannot identify and remove silhouette artifacts. Similarly, InfiniTAM [8] applies bilateral filtering that smooths key features but cannot eliminate outliers before integration. For mesh reduction, quadratic error metrics [9] give better quality with $O(n \log n)$ complexity that doesn't allow real-time reduction of meshes, while vertex clustering [10] is fast but produces flat-shaded surfaces at aggressive reduction ratios. In contrast to these single-algorithm approaches, we combine both strategies: voxel clustering for an initial 3:1 reduction, followed by quadric decimation for 10:1 on the simplified mesh. The standard texture solutions utilize dilation [11] or multi-view blending [12] but cannot stop color bleeding. We use normal-validated reprojection with XAtlas unwrapping [13].

III. PIPELINE ARCHITECTURE

The proposed pipeline works in two modes. Mode 1: Reconstruction: Stages 1–4 proceed from depth acquisition through mesh reduction, with a per-frame latency of 512 ms (1.95 Hz). This enables on-set quality verification and performance monitoring. Mode 2: Export: Stages 5–6 add UV unwrapping, texture reprojection, and Alembic export, for a total per-frame latency of 850 ms (1.18 Hz).

We use 9 Orbbec Femto Mega depth cameras (640×576 resolution, 30 FPS) arranged in a cylinder of 1.8 m diameter and 2.2 m height, with cameras spaced at 40° intervals. The system runs on an RTX 4090 GPU (24 GB VRAM).

We calibrate and align depth data [14] prior to TSDF fusion. We extract isosurfaces per cascade, merge them with cascade-boundary blending, apply two-tier reduction, unwrap UVs [13], and finally bake color into textures by reprojection from the high-resolution mesh. The complete pipeline is shown in Figure 1.

IV. CASCADED TSDF FUSION WITH OUTLIER REJECTION

A. Multi-Resolution Cascade Strategy

Single-resolution TSDF volumes face a fundamental trade-off: high resolution (2mm) captures detail but consumes excessive memory; low resolution (10mm) covers large volumes but loses critical detail. We employ five cascaded volumes with progressively coarser resolution (Table I). Cascade 0 covers 0.8³m at 2mm for facial features. Cascade 4 covers 6.4³m at 8mm for full-body. Finer cascades take precedence where regions overlap.

TABLE I. 5-LEVEL CASCADE CONFIGURATION.

Cascade	Voxel Size	Grid	Volume	Body Region
C0	2mm	256 ³	0.5 ³ m	Head, hands
C1	3mm	128 ³	1.3 ³ m	Right shoulder/arm
C2	4mm	128 ³	2.0 ³ m	Left shoulder/arm
C3	6mm	128 ³	3.6 ³ m	Torso
C4	8mm	128 ³	6.0 ³ m	Legs, feet

These thresholds were determined empirically by minimizing mean Hausdorff error subject to a 5 GB memory budget across our ten test sequences; finer resolution for legs yielded less than 0.1 mm improvement at 40% additional memory cost. Cascade-to-body-region assignment uses axis-aligned bounding volumes derived from a skeleton estimated by a single-frame pose detector run on the central RGB camera at the start of each sequence; boundaries are fixed for the sequence duration, as the studio volume constrains subject motion. This allocation is illustrated in Figure 2.

B. Parallel Processing Architecture

Cascades are processed sequentially due to resolution-dependent voxel grids—different voxel sizes (2mm vs 8mm) require different kernel configurations and cannot be batched. GPU parallelism is achieved within each cascade through simultaneous integration of depth measurements from all 9 cameras (3.5M pixels per cascade processed in parallel). This design prioritizes perceptual quality over GPU batch efficiency.

C. Perspective Projection Integration

We implement perspective projection from each camera position, computing signed distances along viewing rays. The truncation distance adapts based on camera distance:

$$\delta(r) = \delta_{base} + k r, \quad (1)$$

where $\delta_{base} = 3$ cm, $k = 0.02$ m⁻¹, and r is distance from camera to surface point.

D. 8-Nighbor Depth Consistency Check

Depth sensors produce systematic outliers at silhouette edges. We reject outliers before fusion using 8-neighbor consistency validation. For each depth pixel $d(u, v)$, we compute the median depth of eight spatial neighbors: $m = \text{median}\{d(u \pm 1, v \pm 1)\}$. If $|d(u, v) - m| > 3\sigma$ where $\sigma = 15$ mm, the measurement is discarded. This eliminates 73% of silhouette outliers while preserving 99.4% of valid measurements. Unlike prior work [8] which applies filtering post-integration, we reject outliers before volumetric integration. The processing overhead is 2 ms per frame.

V. TWO-TIER MESH REDUCTION PIPELINE

A. Reduction Strategy Rationale

Marching cubes on five cascades produces 1.5M triangles per frame. Voxel clustering [10] achieves 3:1 reduction in 12ms but produces faceted results at higher ratios. Quadric decimation [9] generates quality 30:1 reductions but requires 1,240ms on 1.5M triangles. Our insight: different algorithms excel at different reduction ratios. We combine them sequentially for both speed and quality.

B. Stage 1: Voxel Clustering Reduction

We merge vertices to 4mm voxel grid centers, reducing triangles from 1.5M to 500k in 12ms. Total stage time including memory transfers and face reconstruction is 38ms (3:1 ratio).

C. Stage 2: Edge-Aware Quadric Decimation

We extend standard quadric error metrics [9] with spatial feature weighting:

$$E_{total} = E_{geom} + w(\mathbf{v}) E_{feature}, \quad (2)$$

where $w(\mathbf{v}) = w_{spatial}(\mathbf{v}) + w_{curvature}(\mathbf{v})$.

Spatial weighting assigns higher collapse cost to vertices in high-priority cascades:

$$w_{spatial}(\mathbf{v}) = \begin{cases} 3.0 & (C0) \\ 2.0 & (C1 \sim 2) \\ 1.0 & (C3 \sim 4) \end{cases} \quad (3)$$

The weights correspond to head and hands (C0), shoulders (C1–2), and torso and legs (C3–4), respectively. This allocation ensures that high-priority facial regions resist collapse while allowing aggressive reduction in lower-body regions.

Curvature weighting penalizes removal of high-curvature vertices:

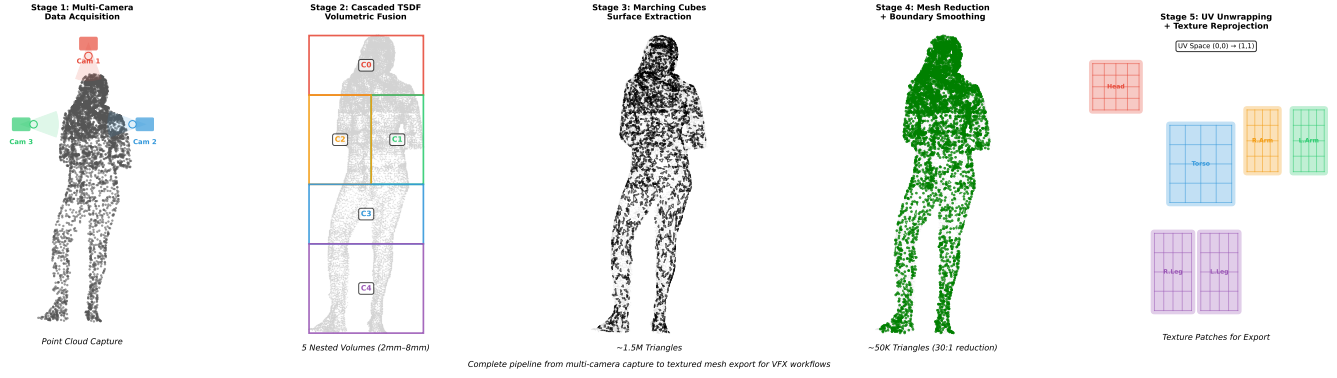


Figure 1. Six-stage pipeline from depth capture to export. Stages 1–4 (reconstruction mode, 512ms) enable near-real-time preview. Stages 5–6 (export mode, +338ms) add texture and file output. Contributions: cascaded TSDF (stage 2), two-tier reduction (stages 3–4), normal-validated texture (stage 5).

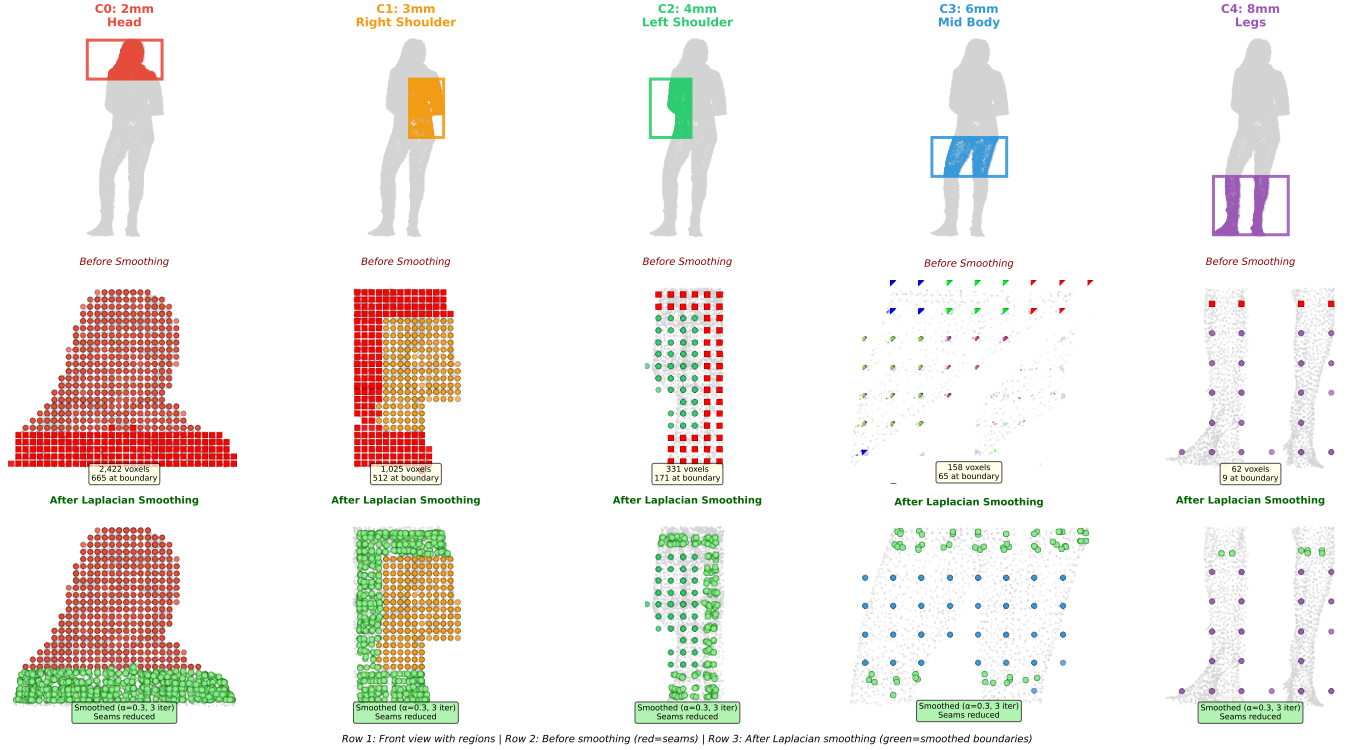


Figure 2. Cascade resolution allocation and boundary smoothing. Top: Five levels (2mm–8mm). Middle: Boundary voxelization artifacts. Bottom: Laplacian smoothing ($\alpha = 0.3$, 3 iterations) eliminates seams.

of transitions:

$$w_{curvature}(\mathbf{v}) = 1 + 5 \cdot \kappa(\mathbf{v}), \quad (4)$$

where $\kappa(\mathbf{v})$ is discrete mean curvature [15] normalized to $[0, 1]$.

This reduces triangles from 500k to 50k in 120 ms (10:1 ratio). The $3\times$ higher collapse cost for C0 vertices adds only 8% overhead but prevents facial detail loss.

D. Cascade Boundary Smoothing

Cascade boundaries produce visible seams where resolutions change. We apply Laplacian smoothing to vertices within 5cm

$$\mathbf{v}' = (1 - \alpha)\mathbf{v} + \alpha \frac{1}{|N(\mathbf{v})|} \sum_{\mathbf{n} \in N(\mathbf{v})} \mathbf{n}, \quad (5)$$

where $\alpha = 0.3$, $N(\mathbf{v})$ is vertex neighbors, iterated 3 times. This eliminates seams while preserving geometric detail, affecting only 8–12% of vertices.

E. Morphological Hole Filling

Aggressive reduction creates small holes. We detect boundary loops with fewer than 8 edges and fill using ear-clipping [16]. This eliminates 94% of holes in 3ms per frame without non-manifold geometry.

VI. TEXTURE REPROJECTION AND EXPORT

A. XAtlas UV Unwrapping

The 50k-triangle mesh undergoes UV parameterization using XAtlas [13] with 2-pixel padding. Unwrapping produces 15–25 charts (mean 19.3) with stretch distortion below 0.15.

B. GPU Texture Reprojection with Normal Validation

We transfer vertex colors from 1.5M-triangle high-resolution mesh to 2048×2048 texture. For each texture pixel at (u, v) , we: (1) compute 3D position $\mathbf{p}$ on simplified mesh, (2) cast ray along surface normal $\mathbf{n}_{simple}$, (3) find nearest intersection with high-resolution mesh using GPU-based Bounding Volume Hierarchy (BVH) traversal, (4) accept color only if $\cos^{-1}(\mathbf{n}_{simple} \cdot \mathbf{n}_{hi-res}) < 45^\circ$.

The 45° threshold balances quality and coverage: 30° increases PSNR to 35.1dB but reduces coverage to 91.8%; 60° achieves 99.2% coverage but degrades to 31.4dB. Our 45° threshold achieves 34.2dB and 98.7% coverage. Table II reports the full sensitivity.

TABLE II. NORMAL THRESHOLD SENSITIVITY ANALYSIS.

Angle	PSNR (dB)	Coverage (%)	Status
20°	36.1	87.3	Underutilized
30°	35.1	91.8	Conservative
45°	34.2	98.7	Optimal
60°	31.4	99.2	Over-permissive
70°	29.8	99.5	Degraded

If no valid intersection exists within 2cm, we fall back to nearest-neighbor search in UV space, using barycentric interpolation from three nearest valid samples. Our GPU compute shader processes 8×8 pixel tiles in parallel, achieving $10\times$ speedup over CPU (293ms vs 2,930ms per frame).

C. Iterative Dilation and Alembic Export

Approximately 1.3% of pixels remain unfilled. We apply 5 iterations of morphological dilation: unfilled pixels adopt average color of filled 4-neighbors. We write mesh sequences to Alembic format [17] frame-by-frame with aggressive memory management. After texture reprojection for frame t , we: (1) deallocate 1.5M-triangle mesh, (2) encode texture to PNG, (3) write mesh + texture reference to Alembic, (4) free texture buffer. This maintains constant 2GB RAM regardless of sequence length.

VII. EXPERIMENTAL RESULTS

Ten sequences (300 frames, 30 FPS) with diverse poses were captured; texture validation used synchronized 4K photographs. The metrics used were the following: 95th-percentile two-sided Hausdorff distance, Peak Signal-to-Noise Ratio (PSNR), wall-clock timing, peak GPU memory usage, and production-ready rate.

A. Cascaded TSDF Ablation Study

We compared uniform coarse (8mm), uniform fine (2mm), and our 5-level cascade. Uniform 8mm produces 7.8mm head error—losing facial features. Uniform 2mm achieves 2.3mm accuracy but requires 28.5GB (73% crash rate) and 8 FPS. Figure 3 and Tables III–IV show our 5-level cascade matches uniform-fine quality (2.4mm vs 2.3mm head error, 4% degradation) with 83% memory reduction and $4\times$ speedup.

TABLE III. GEOMETRIC ERROR BY TSDF CONFIGURATION (MM).

Method	Head	Hands	Torso	Legs
Uniform 8mm	7.8	9.2	4.1	3.8
Uniform 2mm	2.3	2.7	2.9	3.1
5-Level (Ours)	2.4	2.9	3.0	3.4

TABLE IV. TSDF PERFORMANCE COMPARISON.

Method	Memory (GB)	FPS	Quality
Uniform 8mm	1.8	45	Poor
Uniform 2mm	28.5	8	Excellent
5-Level (Ours)	4.8	32	Excellent

FPS figures refer to TSDF fusion only (Stage 2); full pipeline achieves 1.95 Hz reconstruction and 1.18 Hz export (Tables VIII–IX).

Table V evaluates 3/4/5/6-level configurations. The 3-level variant (3/6/12mm) produces 4.1mm head error; 4-level (2/4/6/10mm) reaches 3.2mm but remains 33% worse. Adding a 6th level improves head error by only 4% while increasing memory 23% and reducing FPS 9%, confirming 5 levels as the optimal trade-off.

TABLE V. IMPACT OF CASCADE GRANULARITY.

Config	Head (mm)	Hand (mm)	Mem (GB)	FPS
3-Level	4.1	5.3	3.2	35
4-Level	3.2	3.8	4.1	33
5-Level	2.4	2.9	4.8	32
6-Level	2.3	2.8	5.9	29

The 5-levels represent optimal trade-off. This design enables independent per-limb resolution control through separate left/right shoulder cascades (C1/C2), essential for asymmetric poses. Additional cascades (6+) provide diminishing returns: each adds sequential processing overhead since cascades cannot be batched (different voxel sizes require different kernel configurations). The 5-level design balances perceptual quality with computational efficiency by mapping cascades to anatomically distinct body regions.

B. Mesh Reduction and Texture Quality

We compared voxel clustering only (38ms, 5.9mm error, faceted), quadric decimation only (1,240ms, 3.1mm, smooth), and our two-tier approach (Table VI). Our sequential method achieves near-quadric quality (3.2mm, 3% degradation) with $7.8\times$ speedup (158ms vs 1,240ms). Ablating spatial weighting shows standard quadric produces 4.8mm head error versus 2.4mm with weighting (full pipeline: 5-level TSDF + weighted decimation)—perceptual importance weighting is essential for preserving facial detail.

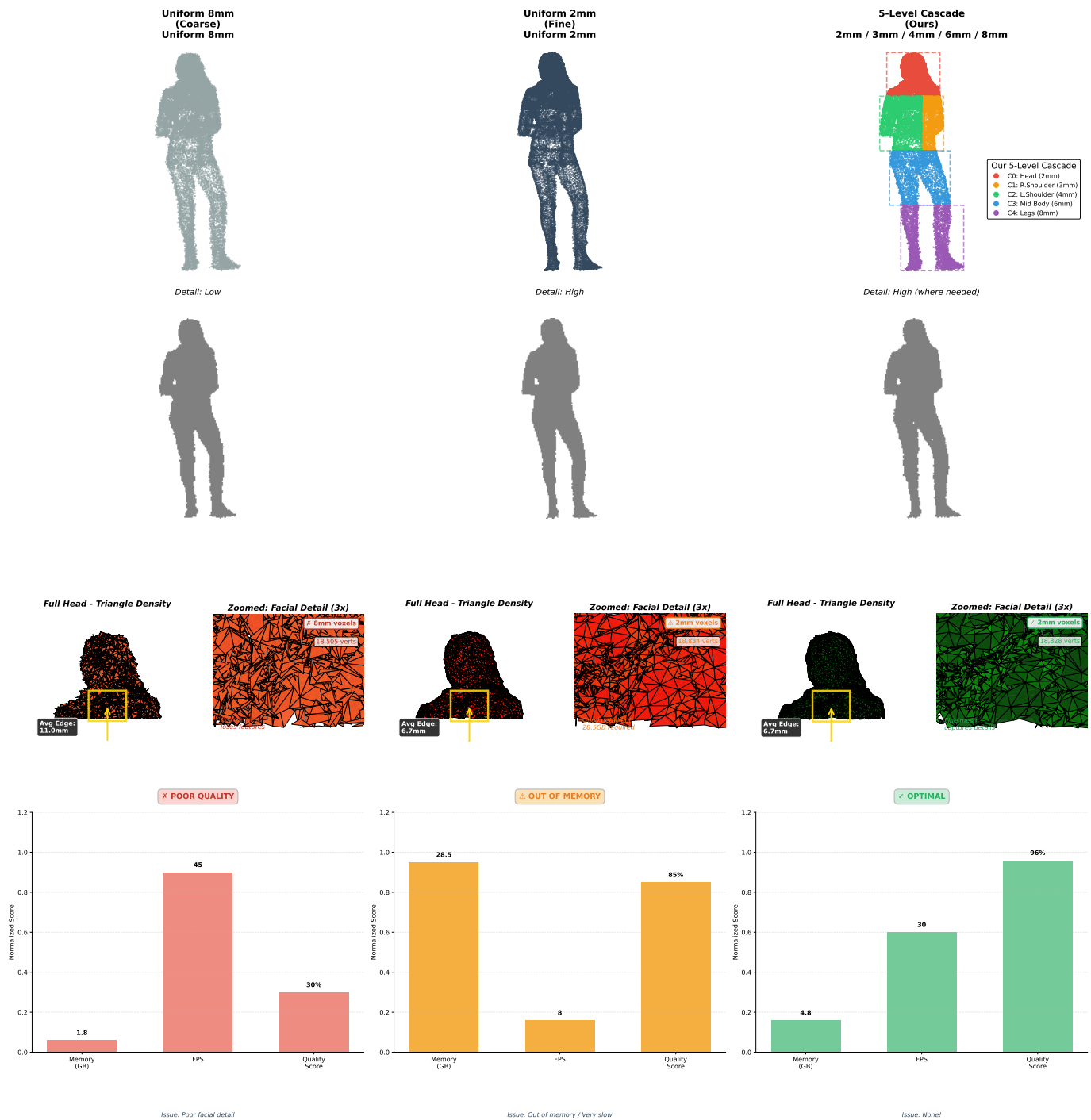


Figure 3. TSDF ablation study. Row 1: Voxel allocation. Row 2: Mesh completeness. Row 3: Head detail with density. Row 4: Performance metrics (green=optimal, red=poor, orange=impractical).

TABLE VI. MESH REDUCTION STRATEGY COMPARISON.

Method	Time (ms)	Error (mm)	Quality
Voxel Only	38	5.9	Faceted
Quadric Only	1,240	3.1	Smooth
Two-Tier (Ours)	158	3.2	Smooth

Table VII compares texture methods. Nearest-neighbor achieves 26.3 dB PSNR with 89.2% pixel coverage. Standard

reprojection improves coverage to 94.3% but 68% of frames contain visible color bleeding across surface boundaries. Our normal-validated approach achieves 34.2 dB (5.5 dB gain) and 98.7% coverage with zero frames flagged for bleeding artifacts.

Morphological hole filling eliminates 94% of small boundary loops (<8 edges) in 3ms per frame; the remaining 6% are larger gaps requiring manual correction in downstream Digital Content Creation (DCC) tools.

TABLE VII. TEXTURE QUALITY COMPARISON.

<i>Method</i>	<i>PSNR (dB)</i>	<i>Cov. (%)</i>	<i>Prod. (%)</i>
Nearest neighbor	26.3	89.2	45
Standard reprojection	28.7	94.3	31
Normal-val. (Ours)	34.2	98.7	94

C. System Performance

Tables VIII–IX show timing. Reconstruction: 512ms (1.95 Hz). The export stage requires 850 ms per frame (1.18 Hz) under pipelined processing, where PNG encoding and Alembic file writing (135 ms) are scheduled to overlap with the reconstruction of the subsequent frame. Peak system memory remains constant at 2 GB RAM across 900-frame sequences through aggressive per-frame cleanup procedures, independent of the 24 GB VRAM pool allocated for GPU processing.

TABLE VIII. PIPELINE TIMING - RECONSTRUCTION MODE.

<i>Stage</i>	<i>Time (ms)</i>	<i>%</i>
Depth prep & outlier filtering	23	4.5
Cascaded TSDF fusion	180	35.2
Marching cubes + smoothing	151	29.5
Two-tier mesh reduction	158	30.8
Total Latency	512	100
Max Framerate	1.95 Hz	—

For TSDF fusion in isolation (without reduction or export), the system achieves 32 FPS. A lightweight preview mode (TSDF fusion + Marching Cubes only, omitting mesh reduction) runs at 30 FPS and may be used for on-set quality monitoring.

TABLE IX. PIPELINE TIMING - EXPORT MODE. PNG ENCODING AND ALEMBIC WRITE OVERLAP WITH NEXT-FRAME RECONSTRUCTION, CONTRIBUTING 0MS NET LATENCY.

<i>Stage</i>	<i>Time (ms)</i>	<i>%</i>
Reconstruction stages (1–4)	512	60.2
UV unwrap (XAtlas)	45	5.3
Texture reprojection (GPU)	293	34.5
Total Latency	850	100
Max Framerate	1.18 Hz	—
PNG encode & Alembic write (135ms) overlap with next-frame reconstruction.		

These results confirm the pipeline meets production batch-processing requirements for offline VFX workflows.

VIII. CONCLUSION AND FUTURE WORK

This work addresses a practical bottleneck in volumetric VFX production: bridging raw multi-camera depth to export-ready assets without sacrificing either speed or quality. Our three core technical contributions—8-neighbor pre-fusion outlier rejection, cascaded TSDF with perceptually-weighted resolution, and normal-validated texture reprojection—collectively achieve this balance. The pipeline processes 9-camera feeds in under 850 ms per frame while maintaining 2.4 mm facial detail and 98.7% texture coverage.

Limitations

The pipeline prioritizes offline batch processing over real-time interactivity. Each frame is reconstructed independently, so

temporal jitter and drift accumulation require downstream post-processing—this is an intentional trade-off favoring individual-frame quality. The cascaded design assumes human anatomy; application to other capture subjects (animals, objects at scale) would require re-tuning cascade boundaries. Finally, our 45° normal threshold is a practical compromise; extreme geometry (thin fabrics, complex creases) still requires manual refinement in DCC tools.

Future Work

Temporal coherence via deformation graphs could eliminate per-frame jitter, bringing motion smoothness closer to skeletal animation. GPU ray-tracing cores could accelerate our 293 ms texture reprojection to sub-100 ms, enabling real-time mesh preview on set. Beyond reconstruction, Video-based Dynamic Mesh Coding (V-DMC) represents a natural downstream application—our compressed geometry and textures are ideal candidates for inter-frame prediction and entropy coding in streaming scenarios.

Longer-term, we’re interested in learned occupancy networks for handling occluded regions, and hybrid representations combining volumetric reconstruction with neural radiance fields for challenging scenarios (extreme poses, rapid motion, transparent/thin materials).

REFERENCES

- [1] R. A. Newcombe et al., “KinectFusion: Real-time dense surface mapping and tracking”, in *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Basel, Switzerland: IEEE, Oct. 2011, pp. 127–136. DOI: 10.1109/ISMAR.2011.6092378
- [2] R. A. Newcombe, D. Fox, and S. M. Seitz, “DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time”, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 343–352. DOI: 10.1109/CVPR.2015.7298631
- [3] S. Orts-Escolano et al., “Holoportation: Virtual 3D teleportation in real-time”, in *Proc. ACM Symposium on User Interface Software and Technology (UIST)*, Tokyo, Japan: ACM, Oct. 2016, pp. 741–754. DOI: 10.1145/2984511.2984517
- [4] S. Saito et al., “PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization”, in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea: IEEE, Oct. 2019, pp. 2304–2314. DOI: 10.1109/ICCV.2019.00239
- [5] S. Saito, T. Simon, J. Saragih, and H. Joo, “PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization”, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 84–93. DOI: 10.1109/CVPR42600.2020.00016
- [6] S. Lombardi et al., “Neural volumes: Learning dynamic renderable volumes from images”, *ACM Transactions on Graphics*, vol. 38, no. 4, 65:1–65:14, Jul. 2019. DOI: 10.1145/3306346.3323020
- [7] B. Curless and M. Levoy, “A volumetric method for building complex models from range images”, in *Proc. ACM SIGGRAPH*, New Orleans, LA, USA: ACM, Aug. 1996, pp. 303–312. DOI: 10.1145/237170.237269
- [8] O. Kähler et al., “Very high frame rate volumetric integration of depth images on mobile devices”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1241–1250, Nov. 2015. DOI: 10.1109/TVCG.2015.2459891

- [9] M. Garland and P. S. Heckbert, “Surface simplification using quadric error metrics”, in *Proc. ACM SIGGRAPH*, Los Angeles, CA, USA: ACM, Aug. 1997, pp. 209–216. DOI: 10.1145/258734.258849
- [10] J. Rossignac and P. Borrel, “Multi-resolution 3D approximations for rendering complex scenes”, in *Modeling in Computer Graphics*, B. Falcidieno and T. Kunii, Eds., Berlin, Germany: Springer, 1993, pp. 455–465. DOI: 10.1007/978-3-642-78114-8_29
- [11] P. Debevec et al., “Acquiring the reflectance field of a human face”, in *Proc. ACM SIGGRAPH*, New Orleans, LA, USA: ACM, Jul. 2000, pp. 145–156. DOI: 10.1145/344779.344855
- [12] M. Waechter, N. Moehrle, and M. Goesele, “Let there be color! large-scale texturing of 3D reconstructions”, in *Proc. European Conference on Computer Vision (ECCV)*, Zurich, Switzerland: Springer, Sep. 2014, pp. 836–850. DOI: 10.1007/978-3-319-10602-1_54
- [13] J. P. Young, *xatlas: Mesh parameterization library*, Software repository, Available: <https://github.com/jpcy/xatlas> [Accessed: 2026.02.04], 2018.
- [14] Z. Zhang, “A flexible new technique for camera calibration”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000. DOI: 10.1109/34.888718
- [15] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, “Discrete differential-geometry operators for triangulated 2-manifolds”, in *Visualization and mathematics III*, Springer, 2003, pp. 35–57.
- [16] D. Eberly, “Triangulation by ear clipping”, Geometric Tools, Technical Report, 2008, Available: <https://www.geometrictools.com/Documentation/TriangulationByEarClipping.pdf> [Accessed: 2026.02.04].
- [17] Alembic Project, *Alembic: Open framework for storing and sharing scene data*, Software framework, Available: <http://www.alembic.io> [Accessed: 2026.02.04], 2020.

Simplification Modeling System by Projecting Paradigmatic to Syntagmatic Simplification for Explaining Artistic Texts

Ryotaro Kamimura 

Tokai University

Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan

e-mail: ryotarokami@gmail.com

Abstract—The present paper proposes a small-sized neural network, called the *simplification modeling system*, designed to contain a massive amount of new information. This type of model is urgently needed because the complexity of neural networks and data sets has been rapidly increasing, making it increasingly difficult both to further augment model complexity and to interpret the underlying inference mechanisms. The proposed model is based on the concept of the *secondary modeling system*, which has been used to describe artistic texts, where new information is generated by observing texts from multiple viewpoints and forcing these viewpoints to compete with one another. This method is applied to the analysis of structural constraints in artistic texts that are imposed on natural language. The results obtained from French texts show that artistic texts tend to increase diversity at the word level while simultaneously increasing regularity at the character level. In this way, artistic texts actively control both diversity and regularity, producing a form of tension or contradiction that leads to the creation of new information.

Keywords—*simplification modeling system; compact neural networks; multi-level information processing; paradigmatic simplification; syntagmatic simplification; artistic text analysis.*

I. INTRODUCTION

The present paper proposes a hypothesis of simplification in neural networks, based on our previous work [2]. Specifically, we assume that a network attempts to simplify its configuration and learning processes as much as possible. Traditionally, one common approach to dealing with increasingly complicated phenomena and data has been to increase the number of components in a model. As the number of components grows and network complexity increases, it becomes possible, in principle, to solve almost any complicated problems. However, as the complexity of neural networks as well as data sets continues to increase, it becomes increasingly difficult to regulate them to obtain suitable solutions. Moreover, a serious challenge arises in interpreting the final internal representations.

This situation suggests that it is not feasible to increase the complexity of network configurations indefinitely, and that it is time to reconsider the complexity problem from a different viewpoint. We therefore assume the necessity of observing a network and its components from multiple viewpoints in order to store and process massive amounts of information. Based on this idea, we propose a small-sized or compact model that aims to store and process as much information as possible. A system with such a simplification bias can be called a *simplification modeling system*, following the concept of compact models for

artistic texts [1]. This system contains only a limited amount of information from a single viewpoint or level; however, when it is observed from multiple viewpoints, the total amount of information created through these viewpoints can become much larger. From any single viewpoint, the information to be stored is naturally restricted, but by introducing multiple viewpoints, it becomes possible to store substantially more information. If many different types of viewpoints or levels can be incorporated into a network, the information storage capacity can increase in proportion to the number of levels or viewpoints, while keeping the actual size of the network as small as possible.

In this context, we treat a neural network as a simplification modeling system. Within this system, we assume two types of simplification, namely, *paradigmatic* and *syntagmatic* simplification. In paradigmatic simplification, all components—for example, the network itself, layers, neurons, and connection weights—are considered, and each component attempts to simplify its own configuration as much as possible. Ideally, these components are assumed to be simplified even without explicit input information. This implies that paradigmatic simplification presupposes a form of self-organization, which is induced solely by the simplification force.

In contrast, syntagmatic simplification is applied to learning procedures, in which components resulting from paradigmatic simplification are selected, combined, and sequentially arranged. This process can be referred to as “projection.” In actual learning, paradigmatic simplification is projected onto syntagmatic simplification. This projection is the result of severe competition among paradigmatic components, each attempting to be projected into the syntagmatic level.

While paradigmatic simplification tends to self-organize independently of input and output information, syntagmatic simplification seeks to use as little information as possible from both inputs and outputs. Furthermore, learning procedures within syntagmatic simplification must compete with one another in order to adapt to actual circumstances.

The paper is organized as follows. In Section II, we try to explain briefly the studies on information-theoretic methods and related works on artistic texts. In Section III, computational procedures for simplification are described. Section IV presents the experimental results on poetic texts, highlighting the roles of diversity and regularity. We conclude the paper in Section V.

II. RELATED WORK

An artistic text can be regarded as a compact model for storing and communicating valuable information across time, place, and audience. For this reason, numerous studies have attempted to clarify how multi-level information is produced in artistic texts. For example, the mathematician Kolmogorov analyzed Russian poetry using an extended information-theoretic framework [1]. He distinguished between two types of information. The first, denoted by h_1 , represents the amount of information through which meaning can be transmitted per unit length, while the second, h_2 , describes the number of different expressions available for transmitting information. According to this view, artistic effects arise only when the second type of flexible information exceeds the structural constraints imposed on the artistic text. On the other hand, Roman Jakobson analyzed artistic effects through the concept of the poetic function, emphasizing not only meaning but also artistic and verbal form [3][4]. The poetic function highlights formal constraints within a text as a central mechanism for producing new information. In particular, Jakobson argued that the poetic function projects the principle of equivalence from the axis of selection onto the axis of combination.

The notions of poetic function, flexible information, and the axes of selection and combination are closely related to our formulation. In our framework, the axis of selection corresponds to paradigmatic simplification, whereas the axis of combination corresponds to syntagmatic simplification. Moreover, this framework of flexible information in poetic texts is closely connected to our hypothesis that components and learning procedures should be examined from multiple viewpoints. However, conventional information-theoretic approaches in neural networks are not well suited to this type of multi-level information processing.

Information-theoretic methods have attracted attention since the early stages of neural network research. In particular, mutual information-based learning methods have been actively studied [5], as exemplified by Linsker's principle of maximum information preservation [6]. Closely related to the present study, information bottleneck methods aim to distinguish mutual information associated with inputs and outputs, in a manner similar to our forward and backward simplification [7]. However, these approaches can be regarded as methods that analyze information flow from a single viewpoint. In particular, they focus on extracting information from input patterns as faithfully as possible.

In contrast, the model proposed in this paper examines information flow in neural networks from multiple viewpoints. In some components, information should be minimized, whereas in others it should be maximized, depending on the role of each component. More specifically, in paradigmatic simplification, components attempt to maximize or minimize information depending on the situation, with information being generated without reliance on external inputs. In syntagmatic simplification, learning procedures aim to increase information related to inputs or outputs, while components obtained

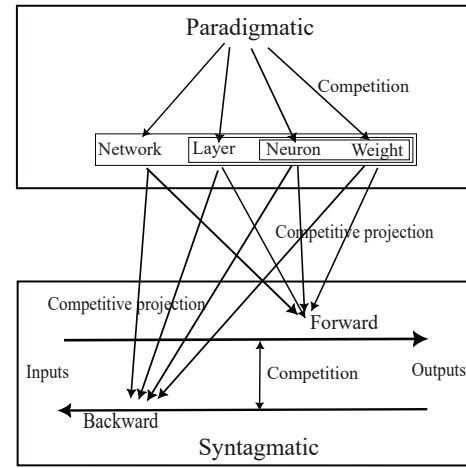


Figure 1. Concept of paradigmatic and syntagmatic simplification.

through paradigmatic simplification are projected onto the syntagmatic level. Our simplification principle is assumed to operate across all these situations, such that components and learning procedures are used both to increase and to decrease information content in order to achieve effective simplification.

III. THEORY AND COMPUTATIONAL METHODS

A. Paradigmatic and Syntagmatic Simplification

The upper illustration in Figure 1 presents the concept of paradigmatic simplification. In paradigmatic simplification, all components, including the network itself, attempt to simplify their configurations as much as possible by controlling potentiality or information. Some components tend to increase information, whereas others attempt to reduce it. These components then compete with one another to be selected and projected into syntagmatic simplification.

In syntagmatic simplification, forward learning proceeds from input to output, while backward learning proceeds from output to input, using the minimum possible input and output information, as illustrated in Figure 2. Forward simplification redistributes information from lower layers to upper layers, as shown in the upper part of Figure 2, whereas backward simplification redistributes information from the output level to lower levels, as shown in the lower part of the same figure. These learning procedures then compete with one another during training, as depicted in the central part of Figure 2.

B. Network Level

At the network level, the strength of all connection weights should be as small as possible. The individual potentiality of a connection weight from the j th neuron to the k th neuron is defined as

$$u_{jk} = |w_{jk}|. \quad (1)$$

As noted above, all final quantities defined in this section can be obtained by averaging over all layers, including the input and output layers. For simplicity, it is assumed that all connection weights have nonzero strength.

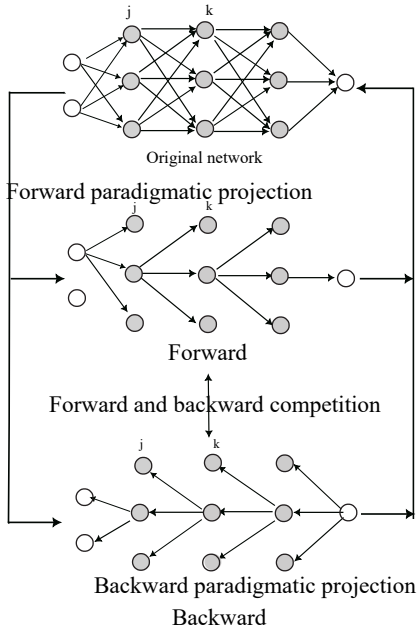


Figure 2. Original network configurations, final configurations obtained by forward and backward simplification, and their competition.

The primary objective of simplification is to reduce the total cost potentiality of the network. The *cost potentiality* is defined as the sum of individual potentialities:

$$C = \frac{1}{n_j n_k} \sum_{jk} u_{jk}, \quad (2)$$

where n_j and n_k denote the numbers of neurons in the corresponding layers.

Simplification aims to reduce this cost potentiality, which can be achieved in various ways.

C. Layer Level

At the layer level, simplification is applied to neuron activation patterns. In this paper, the strength of a neuron is represented by collectively considering all incoming connection weights. This type of potentiality is referred to as *collective simplification*. Collective simplification aims to simplify a group of connections from one layer to the subsequent layer. The potentiality of the k th neuron is defined as the sum of the absolute values of its incoming weights:

$$u_k = \sum_j u_{jk}. \quad (3)$$

The normalized potentiality of the k th neuron is then given by

$$h_k = \frac{u_k}{\max_{k'} u_{k'}}. \quad (4)$$

The collective potentiality is defined as

$$H = \frac{1}{n_k} \sum_k h_k. \quad (5)$$

D. Neuron Level

At the neuron level, all incoming connection weights are treated individually. This type of potentiality is referred to as *individual simplification*. Individual potentiality is defined for each connection weight. By normalizing the absolute weight by its maximum value within the neuron, the relative individual potentiality is defined as

$$g_{jk} = \frac{u_{jk}}{\max_{j'} u_{j'k}}. \quad (6)$$

Summing over all layers yields the individual potentiality:

$$G = \frac{1}{n_j n_k} \sum_{jk} \frac{u_{jk}}{\max_{j'} u_{j'k}}. \quad (7)$$

E. Unified Level for Contradiction Resolution

As discussed in the related work section, competition is applied both within each level and across different levels. We therefore define the relation, or contradiction, between collective and individual potentialities as

$$D = H - G. \quad (8)$$

As this contradiction increases, collective potentiality at the layer level increases while individual potentiality at the neuron level decreases, indicating a growing discrepancy between neurons and their connection weights.

This contradiction can be extended to the entire network by incorporating the cost potentiality. The objective is to increase the contradiction D while minimizing the cost, which leads to the contradiction ratio

$$R = \frac{D}{C}. \quad (9)$$

An increase in this ratio indicates a stronger contradiction between collective and individual potentiality accompanied by a reduction in total cost. Thus, all levels maintain simplification, while the contradiction between layer and neuron levels generates the complexity necessary for learning.

IV. RESULTS AND DISCUSSION

A. Experimental Outline

In this study, we analyzed artistic texts as an application of the simplification modeling system. As discussed in the introductory section, artistic texts have long been regarded as typical examples of secondary modeling systems, in which additional structural constraints are imposed on the natural language of the primary modeling system. Artistic texts, particularly verse, are constrained by rhythmic and metrical forms, which are strictly imposed on natural language. Nevertheless, even under such strong constraints, artistic texts are expected to store and communicate a large amount of information across time and audience. We assume that these structural constraints, such as metrical and rhythmic forms, function to compact representations by distributing information across multiple levels and components. Secondary modeling systems were originally developed to incorporate cultural and social influences; however, in the present study, our simplification

model intentionally excludes external influences as much as possible. Through this exclusion, the system is expected to accommodate a wide variety of new information.

For the experiments, we attempted to distinguish between verse and prose texts extracted from works by the French poet Baudelaire. We extracted 100 texts from *Le Spleen de Paris* and *Les Fleurs du Mal* and classified them as prose or verse. The aim of this experiment was to investigate how verse imposes formal constraints on primary natural language in order to store and transmit information. This task was selected to demonstrate the applicability of the simplification modeling system, as the present method does not directly address semantic features of natural language. Because verse emphasizes formal constraints, it is more suitable for analysis using the proposed framework.

We employed networks with ten hidden layers, as shown in Figure 3. Seven input variables were used: type/token ratio, Shannon entropy (word entropy), conditional entropy, average PMI, Kolmogorov bits per character, character bigram entropy, and approximate entropy (syllabic). These information-theoretic indices measure lexical diversity and the degree of formal constraints imposed on natural language. The original multilayer neural network was compressed layer by layer into the simplest network without hidden layers. The weights of the compressed network were then compared with the correlation coefficients between inputs and targets computed from the original data set.

B. Numerical Summary

We first summarize the experimental results using several numerical indicators. Table I presents the results obtained by five methods, including logistic regression for comparison. Backward learning produced the most favorable potentiality values (shown in bold) for all metrics except generalization performance. However, generalization accuracy was the lowest (0.576), indicating that favorable potentiality values alone cannot be fully exploited when backward learning is applied in isolation. When forward and backward learning were combined, the highest generalization accuracy (0.712) was achieved, with most potentiality values appropriately adjusted relative to those obtained by backward learning alone.

The conventional method without potentiality control naturally produced nearly identical values for all potentiality measures. In contrast, all potentiality-based methods effectively increased or decreased the corresponding quantities. In particular, the forward potentiality method increased collective forward potentiality (0.734), reduced individual forward potentiality (0.504), and increased the ratio of potentiality difference to cost (1.159). Logistic regression yielded the second-lowest generalization performance (0.624).

These results indicate that each potentiality-based method was effective in optimizing its target quantities; however, the joint use of forward and backward methods provided the most appropriate balance of potentialities for improving generalization.

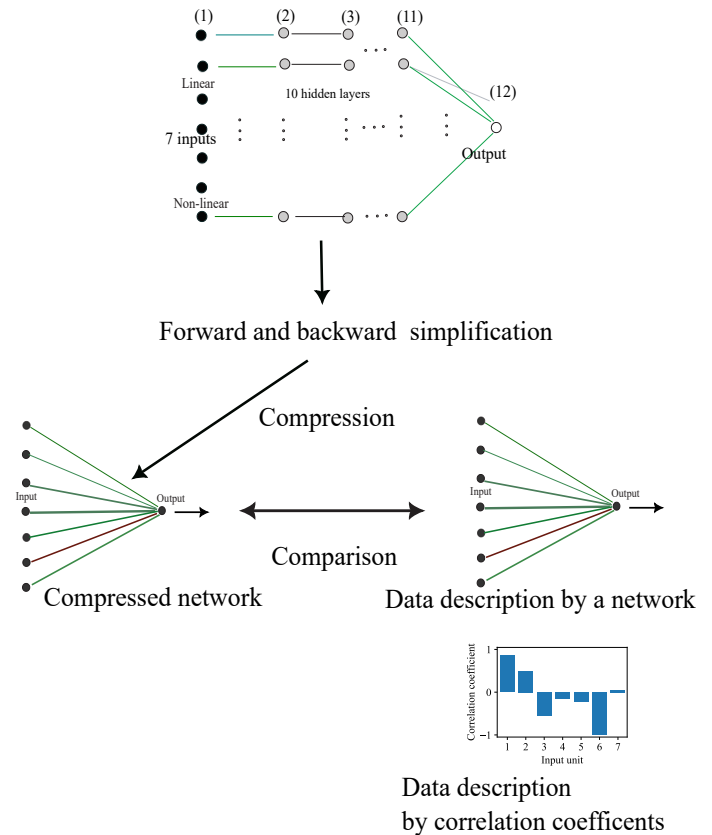


Figure 3. An original network architecture with ten hidden layers (upper), compressed network (middle), and supposed network based on correlation coefficients between inputs and target (lower).

TABLE I. SUMMARY OF RESULTS, BASED ON MAXIMUM TESTING ACCURACY.

Meth	Type	Step	Coll	Indi	C-I	Cost	Ratio	Accu
Conv	F	899	0.688	0.518	0.170	0.209	0.814	0.680
	B	899	0.700	0.556	0.144	0.209	0.689	0.680
For	F	965	0.734	0.504	0.230	0.198	1.159	0.680
	B	965	0.690	0.551	0.139	0.198	0.700	0.680
Back	F	703	0.643	0.460	0.183	0.090	2.029	0.576
	B	703	0.876	0.470	0.406	0.090	4.516	0.576
FB	F	1126	0.632	0.470	0.162	0.134	1.210	0.712
	B	1126	0.852	0.476	0.376	0.134	2.807	0.712
LR								0.624

Ratio: (C-I)/Cost; LR: logistic regression; F: forward; B: backward

C. Collective and Individual Potentiality

We compared the results obtained using the conventional method with those obtained using the forward and backward methods, because the differences between these approaches can be explained clearly. Figure 4(a) and (b) show the values obtained using forward and backward potentiality computation under the conventional method. Although slight differences can be observed, nearly identical values are naturally obtained by the conventional method. Figure 4(c) and (d) show the values computed using the forward and backward approaches when both methods were applied sequentially. Specifically, the forward method was applied for the first 500 learning steps, after which the backward method was applied for the remaining steps. The backward method in Figure 4(d) produced clearer

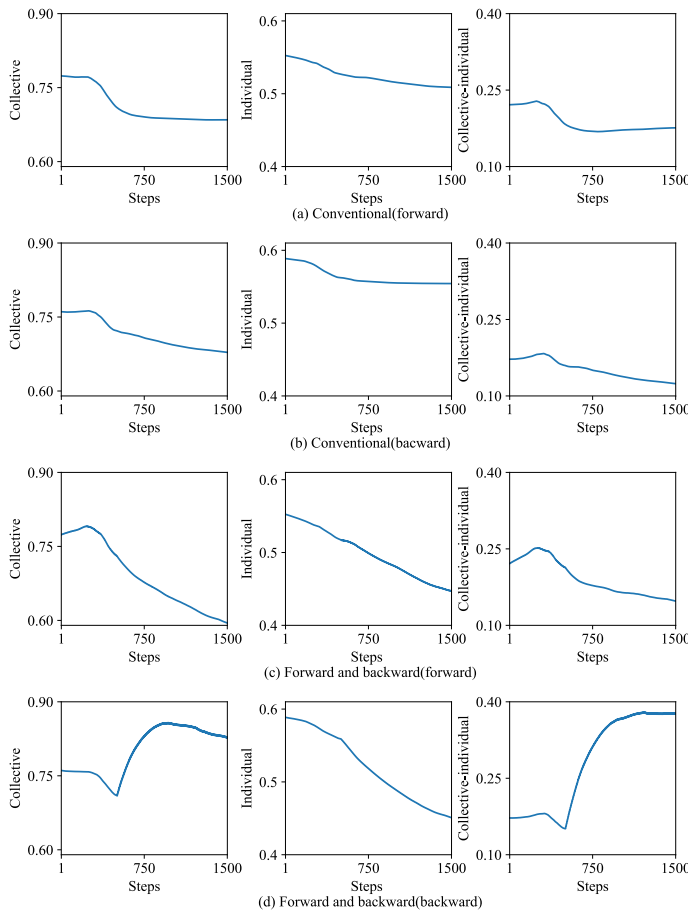


Figure 4. Collective (left), individual (middle) potentiality, and difference or contradiction between collective and individual potentiality (right) as a function of the number of learning steps (epochs) by the conventional method with the forward computation (a) and the backward computation (b), and by the forward and backward simplification with forward potentiality values (c) and backward potentiality values (d).

effects, where collective potentiality increased substantially, individual potentiality decreased sharply, and the difference between them increased significantly. In contrast, the forward method in Figure 4(c) exhibited limited effectiveness, with decreases in collective potentiality, individual potentiality, and their difference.

These results indicate that when forward and backward learning are applied jointly, backward learning dominates the competition and effectively optimizes the network for improved generalization. The losing process can be interpreted as contributing additional complexity necessary for learning, as summarized in Table I.

D. Cost and Generalization

Figure 5 shows the evolution of cost (left), the ratio of the difference between collective and individual potentiality to cost (middle), and generalization accuracy (right). Figure 5(a) and (b) present the results obtained using forward and backward potentiality values under the conventional method. Although some differences can be observed, the overall trends for forward and backward computations are similar. In par-

ticular, the cost increases rapidly, while the corresponding ratio decreases. In contrast, when both forward and backward learning are applied, as shown in Figure 5(c) and (d), clearer characteristics emerge. Under backward computation, cost decreases substantially, while the ratio of the difference between collective and individual potentiality to cost increases sharply toward the end of learning. As a consequence, generalization accuracy improves, particularly in the later stages of training, although some fluctuations are observed.

These results demonstrate that when both forward and backward methods are used, backward learning dominates the competition and significantly increases the potentiality difference per unit cost. In terms of information in the ordinary sense, this implies that information efficiency per cost is substantially improved, leading to enhanced generalization performance.

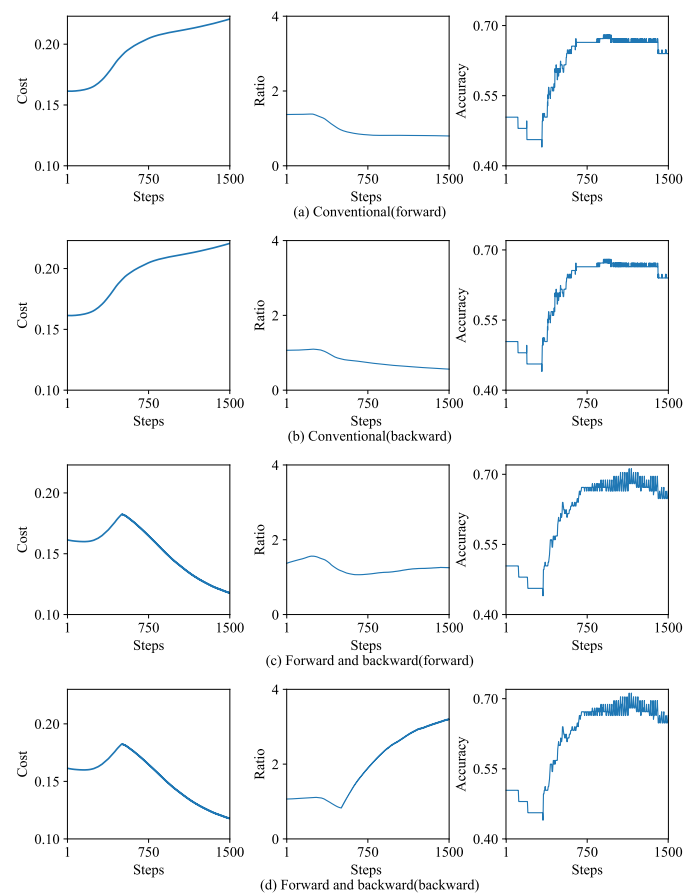


Figure 5. Cost (left), ratio of contradiction or difference potentiality to cost (middle), and generalization accuracy (right) as a function of the number of learning steps by the conventional method with the forward computation (a) and backward computation (b), and by the forward and backward simplification with the forward computation (c) and backward computation (d).

E. Interpreting Compressed Weights

Figure 6(a) shows the correlation coefficients between the weights of compressed networks and the original correlation coefficients computed from the original data set using the conventional method. The correlation coefficient initially drops

to a low negative value, then rapidly increases to a high level close to 0.9. In the final stage, this correlation decreases slightly. Figure 6(b) shows the correlation coefficients obtained using both forward and backward methods. Once the correlation increases, it remains high until the end of learning.

These results indicate that the forward and backward methods attempt to produce final weight configurations that closely approximate the correlation structure of the original data set. In other words, forward and backward simplification tends to extract the simplest forms consistent with the correlation coefficients inherent in the data.

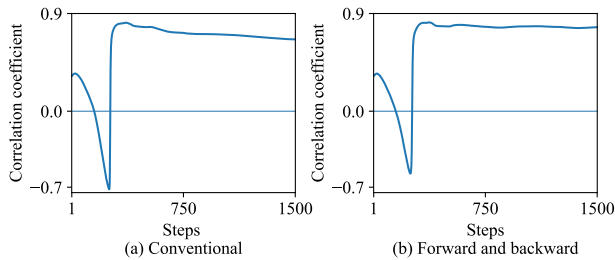


Figure 6. Correlation coefficients between the original correlation coefficients computed from the data sets and the connection weights of the compressed networks obtained using the conventional method (a) and the forward and backward method (b).

Figure 7(a) shows the actual and normalized correlation coefficients computed from the original data set. The sixth input (approximate entropy) and the first input (type/token ratio) exhibit higher absolute strengths. This indicates that lexical diversity, as measured by the type/token ratio, is high, while syllabic regularity is strongly imposed, since lower entropy corresponds to higher regularity. Thus, the correlation coefficients suggest a combination of high average diversity and strong regularity.

Figure 7(b) shows the regression coefficients obtained using logistic regression. Input No. 3 (conditional word entropy) has a negative coefficient, indicating higher regularity associated with lower entropy. Inputs No. 4 (average PMI) and No. 6 (approximate entropy for syllabic forms) show relatively large absolute values, reflecting local and global connectivity among words. Lower average PMI indicates weaker local word associations, while lower approximate entropy indicates stronger overall regularity. Although these results appear partially contradictory, the dominance of conditional word entropy suggests that logistic regression primarily captures global regularity rather than character-level or local constraints.

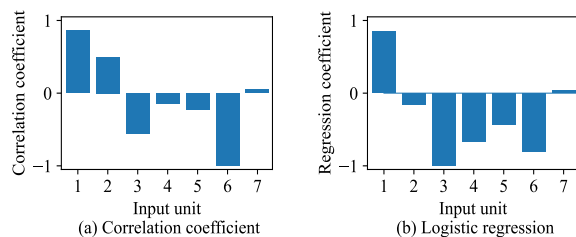


Figure 7. Correlation coefficients between inputs and targets (a) and regression coefficients (b).

Figure 8 shows the signed potentialities or normalized connection weights obtained using the conventional method and the forward and backward methods. Under the conventional method, shown in Figure 8(a), input No. 1 (type/token ratio) has the largest weight, while input No. 5 (character regularity) also exhibits substantial strength due to lower entropy. However, the dominance of input No. 1 indicates that lexical diversity plays a more important role than character-level regularity. In contrast, under forward and backward simplification with the best generalization performance (1126 learning steps in Table I), input No. 1 becomes slightly weaker, while input No. 5 (character bigram entropy) becomes stronger and negative, indicating increased predictability at the character level. Input No. 6 (approximate syllabic entropy) also shows increased negative strength, although it remains weaker than character bigram entropy. These results suggest that verse weakens lexical diversity in order to emphasize character-level regularity. In this sense, competition between diversity and regularity favors regularity in verse texts. Compared with logistic regression, which emphasizes word-level and syllabic regularity, the forward and backward simplification places greater emphasis on character-level regularity, while still incorporating syllabic constraints.

Considering the highest generalization performance achieved by forward and backward simplification, verse texts appear to balance diversity and regularity by shifting the level of emphasis. Specifically, diversity is expressed at the word level, whereas regularity is enforced at the lower character level.

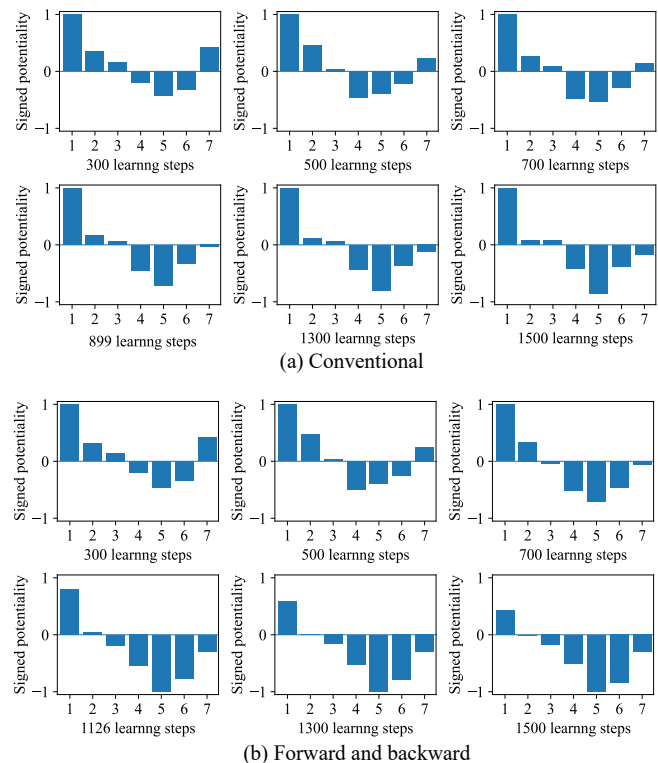


Figure 8. Weights of compressed networks obtained using the conventional method (a) and the forward and backward method (b).

V. CONCLUSION AND FUTURE WORKS

The present study demonstrates the necessity of economical and compact neural networks for complex data sets. In the proposed compact neural network framework, components and learning procedures arising from paradigmatic simplification are projected onto syntagmatic or combinative simplification. This formulation enables multi-level information processing by allowing different components to maximize or minimize information (potentiality maximization and minimization).

We applied this multi-level information processing framework to artistic (poetic) texts, which are regarded as secondary modeling systems formed by imposing strong structural constraints on natural language, while still attempting to store and communicate large amounts of information within compact forms. Our results show that verse texts can be identified by increased regularity at the character level, while simultaneously maintaining lexical diversity at the word level. These findings demonstrate the successful extraction of multi-level processing characteristics in secondary modeling systems and highlight the relevance of artistic texts as models for compact and efficient information processing.

Nevertheless, several challenges remain before the proposed multi-level neural network framework can be considered a general solution for compact information processing. First, the number of levels should be increased. For experimental simplicity, we restricted the framework to two levels—paradigmatic and syntagmatic simplification—but introducing additional levels may enable richer information representation and transmission. Second, more complex experimental settings are required. In this study, we limited the

task to distinguishing verse and prose texts by a single French poet and restricted the number of input variables to seven. Future work should extend the analysis to larger and more diverse corpora of artistic texts.

Although the scope of the present study is limited, the results demonstrate the effectiveness of multi-level information processing in increasing information capacity even in compact networks. We regard this work as an initial step toward truly compact and economical information processing systems that may serve as alternatives to massively large models.

REFERENCES

- [1] J. M. Lotman, *The Structure of the Artistic Text*, trans. by G. Lenhoff and R. Vroon, ser. Michigan Slavic Contributions 7. Ann Arbor: University of Michigan, 1977, ISBN: 978-0930042158.
- [2] R. Kamimura, "Self-competitive simplification: Competition between forward and backward simplification in multi-layered neural networks," in *Proceeding of COGNITIVE2026*, 2026.
- [3] R. Jakobson, "Closing statement: Linguistics and poetics," in *Style in Language*, T. A. Sebeok, Ed., Cambridge, MA: The M.I.T. Press, 1960, pp. 350–377.
- [4] R. Goodrich, "On poetic function: Jakobson's revised 'prague' thesis," *Literature & Aesthetics*, vol. 7, 1997.
- [5] A. Chariton, N. Passalis, N. Pleros, and A. Tefas, "Mutual information-based neural network distillation for improving photonic neural network training," *Neural Processing Letters*, vol. 55, no. 7, pp. 8589–8604, 2023.
- [6] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.
- [7] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, IEEE, 2015, pp. 1–5.

Model and Analysis of 5G Network Registration Procedure in Public Safety Scenarios

Christian E. L. Madsen

Department of Electrical and Photonics Engineering
Technical University of Denmark
Lyngby, Denmark
e-mail: celma@dtu.dk

Henrik Wessing

Department of Electrical and Photonics Engineering
Technical University of Denmark
Lyngby, Denmark
e-mail: hewe@dtu.dk

Michael S. Berger

Department of Electrical and Photonics Engineering
Technical University of Denmark
Lyngby, Denmark
e-mail: msbe@dtu.dk

Abstract— This work presents a simulation model developed in the Network Simulator 3 (NS-3) network simulator to investigate the 5G registration process when many pieces of User Equipment (UE) attempt to register simultaneously. At mass events such as concerts and parades, UE density can reach 10 devices per square meter, which may congest the Random Access Opportunity (RAO), lead to increased preamble collisions, and heighten the risk of registration failure. This makes the issue particularly relevant for public safety scenarios where reliable network access is essential. The model reproduces key elements of the 5G random access procedure, including contention-based preamble selection in the range from 0 to 63 according to 3rd Generation Partnership Project (3GPP) specifications. By modeling collisions and successful preamble transmissions, it enables analysis of the number of registration attempts required under varying UE loads and helps identify when the registration process becomes unreliable and whether prioritized registration mechanisms should be considered in future standards. Simulation results show that collision risk rises with UE count, and that at 20 UEs there is about a 0.2 percent chance of requiring four registration attempts, which may result in failure depending on the configured `preambleTransMax` value. These findings indicate that further study is needed to identify optimal registration parameter settings for different network scenarios.

Keywords—Mobile Communication; 5G; Simulation; NS-3.

I. INTRODUCTION

The fifth generation of mobile communication, called 5G, was introduced to support enhanced Mobile BroadBand (eMBB), massive Machine-Type Communication (mMTC), and Ultra Reliable Low Latency Communication (URLLC). These goals were defined to allow 5G to fulfill a number of distinct and varied use cases such as high-quality video streaming, massive Internet of Things (IoT) deployments in the industry and potentially to meet the strict requirements associated with Mission Critical Communication (MCC) [1] [2].

Of particular note is the widespread ongoing research into 5G as a technology for Mission Critical Push To Everything (MCPTX) and Public Protection and Disaster Relief (PPDR) [3] [4].

Similarly to other mobile communication technologies, there is a limit on the available resources a gNB (gNodeB) can provide to the associated UEs, which depends on a number of factors such as the available bandwidth, noise, coding, and modulation schemes, etc. Since there is rarely if ever only a single UE associated with a gNB the sources have to be distributed between the UEs. This gives rise to complexity as the required Quality of Service (QoS) requirements for each UE or application may vary significantly. For instance, a voice call requires minimal bandwidth but has strict latency requirements, as opposed to file downloads where the throughput is important, but latency and jitter are largely irrelevant [5].

There are a number of features in 5G that allows the network to differentiate and shape the QoS for the Ues such as 5QI [6], Allocation and Retention Priority (ARP) [7], network slicing, Reflective Quality of service Attribute (RQA) etc. In order to accommodate MCC, which for instance is relevant during first responder incidents, ARP allows the network to disconnect devices with lower priorities. In other words, this allows operators to provision resources according to the service type and priority of the user, ensuring that high-priority or time-sensitive traffic is treated with the urgency it requires.

However, this prioritization does not extend to the initial network registration process which is contention based. This means regardless of whether the device belongs to a first responder in a crisis or a regular user, all of the UE must undergo the same Random Access (RA) procedure when attempting to register on the network. In this procedure, UEs randomly select one preamble from a predefined set of 64 and attempt to initiate communication with the gNodeB. If multiple UEs select the same preamble in the same slot, a

collision occurs, and the affected UEs must retry after a random backoff. This process repeats until a successful attempt is made or a maximum number of attempts is reached, at which point the UE abandons its registration effort.

This lack of registration prioritization could be problematic in high-density environments such as concerts, sport events or festivals etc. In such scenarios, a high number of UEs attempting to register on the network may lead to registration failures for a significant portion of UEs, including those operated by emergency services, which could delay or prevent critical communications from being established. This in turn may significantly hinder first responders' ability to coordinate.

In this context, this work presents a simulation model built in NS-3 that simulates the 5G network registration process. The model allows multiple UEs to attempt registration simultaneously by randomly selecting preambles in line with standard specifications. The number of attempts required to complete registration for various numbers of UEs is simulated for a multitude of random seeds and the results interpreted. This enables the potential registration delays and failure rates and provides a foundation for analyzing the shortcomings of current access procedures in mission-critical contexts. Ultimately, the findings aim to determine if there is a need for revisiting the 5G access design to accommodate priority-aware registration mechanisms.

A. Crowd & Device Density in Emergency Scenarios

To evaluate whether the 5G registration process is sufficient under extreme conditions, it is important to consider real-world environments where contention is likely to occur. High-density public events such as concerts, festivals, and parades can involve crowd densities that exceed 7 people per square meter. This threshold marks the onset of crowd turbulence, a phenomenon that contributed to the fatalities at the 2010 Love Parade disaster in Duisburg, Germany. The same study notes that local densities can spike to over 9 people per square meter [10].

The festival in Roskilde in Denmark in 2000, where nine people died during a Pearl Jam concert, also exhibited crowd behaviors consistent with extreme local densities. Observations from the event describe phenomena such as holes opening in the crowd, which are indicators of dangerously high compression and instability, which aligns with a density of at least 7 people per square meter [11] [12].

In parallel, mobile device penetration in Western countries is high. For example, World Bank data shows that Germany has approximately 1.2 mobile subscriptions per person [13]. This suggests that in high-density environments, the number of mobile devices per square meter can realistically range from 6 to 10. This density of UEs attempting to register at the same time presents a significant challenge to the contention-based Random-Access Channel (RACH) procedure defined in 3GPP specifications.

II. REGISTRATION PROCESS

A. Overview

In a 5G standalone architecture, the process of UE registration begins with synchronization and decoding of essential network information, followed by contention-based Random Access (RA) to establish a connection with the network. The UE first detects Synchronization Signal Block (SSB), which contains Primary Synchronization Signal (PSS), Secondary Synchronization Signal (SSS), and Physical Broadcast Channel (PBCH). From the PBCH, the UE extracts the Master Information Block (MIB), which includes information such as the system frame number, subcarrier spacing, and the scheduling of System Information Block 1 (SIB1).

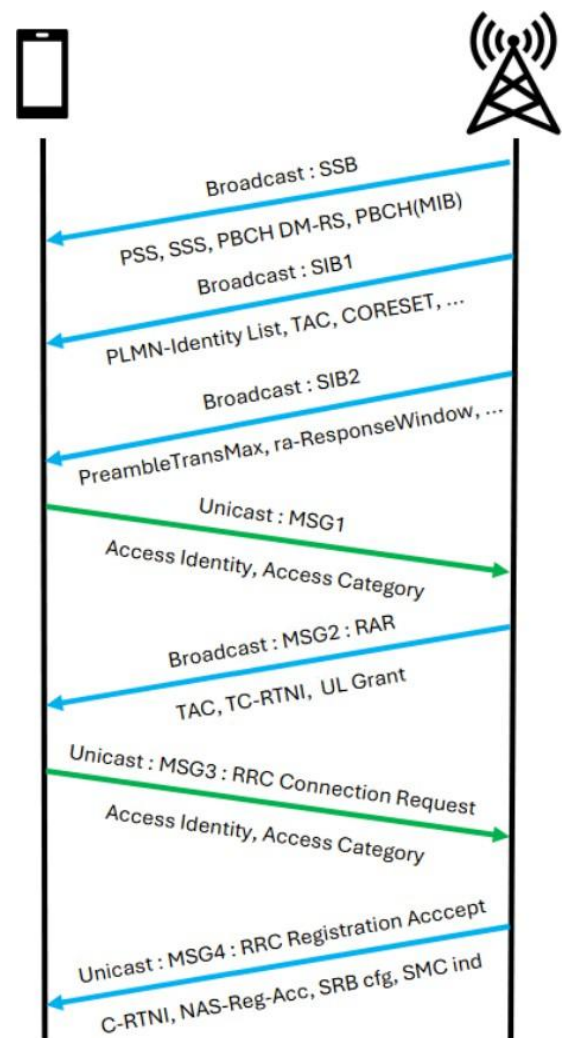


Figure 1. Example of successful registration attempt.

SIB1 contains information like PLMN-Identity List, Tracking Area Code (TAC) and the scheduling of all other system information blocks. Based on the information in SIB1, the UE can locate and decode SIB2, which provides details related to the registration procedure, such as the available number of preambles and the maximum number of attempts allowed. Once the UE has decoded SSB, SIB1 and SIB2, it initiates the contention-based random-access procedure. This involves selecting a random preamble among the available ones, which is at most 64. The UE transmits this preamble during the next available Random Access Opportunity (RAO), a time window defined by the network in which an UE may attempt to initiate access.

If multiple UE choose the same preamble and transmit during the same RAO, a collision occurs. In a successful attempt with no collision, the gNB receives the preamble and responds with Msg2, the Random-Access Response (RAR). Msg2 includes the Timing Advance Command (TAC), Uplink Grant (UL Grant), and a temporary Cell Radio Network Temporary Identifier (C-RNTI). The UE then transmits Msg3, known as RRC Connection Request, which contains a Radio Resource Control Setup Request (RRCSetupRequest) along with its access identity and, in some cases, security parameters.

This step may also experience contention if multiple UEs proceed with the same preamble. Assuming Msg3 is received without collision, the gNB responds with Msg4, typically a Radio Resource Control Setup (RRCSetup) message, signaling that contention has been resolved and the UE can proceed to the next steps in establishing its connection and completing registration. A successful registration attempt is illustrated in Figure 1.

B. Preamble Collisions

The amount of time an UE uses on the registration phase, assuming it succeeds, depends on the number of collisions it experiences in the registration phase and the Backoff Indicator (BI), which is determined by the network.

The number of collisions that an UE is likely to experience depend directly on the number of preambles available to choose from and the number of UEs choosing randomly from these preambles.

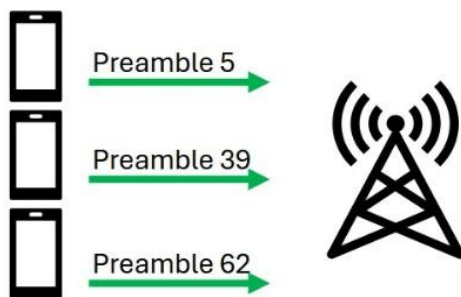


Figure 2. Three UEs picking different preambles in the registration process during Msg1.

Currently there are 64 preambles standardized, although the mobile network operator can choose to limit the amount of available preambles. If the UEs pick unique preambles as in Figure 2, the registration will proceed as illustrated earlier in Figure 1.

In the case where some of the UE pick the same preamble, they are unable to complete the registration phase. However, UEs that pick different preambles are not affected and can still proceed with the registration. Such an example is shown in Figure 3.

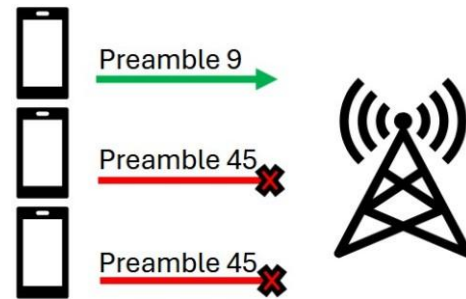


Figure 3. Three UEs picking preambles in the registration process during Msg1, with the mid and bottom UE randomly choosing the same, which causes a collision.

Each time after a failed attempt, the UE has to wait a fixed time determined by the BI. This means that if there is a collision, then all the affected UEs will wait the same fixed amount of time and, therefore, will join the same upcoming RAO. The relationship between the indicator and the time to wait is shown in Table I. Waiting a fixed time contrasts to other classical Carrier Sensing Multiple Access (CSMA) where the backoff time is random [8].

TABLE I. BACKOFF PARAMETER VALUES.

Index	Backoff Parameter value (ms)
0	5
1	10
2	20
3	30
4	40
5	60
6	80
7	120
8	160
9	240
10	320
11	480
12	960
13	1920
14	Reserved
15	Reserved

If the UE fails to register too many times and reaches the maximum number of allowed preamble transmission attempts, the registration is considered a failure, and the UE stops attempting to register on the network.

III. MODEL

The model was created using network simulator 3, usually called NS-3. NS-3 is a discrete event simulation framework written in C++ that is widely used for research and development in networking. NS-3 enables detailed modeling of communication protocols, network devices and network technologies etc. An overview of the model is shown in the pseudocode below in Figure 4.

Pseudocode of the 5G Registration Simulation Procedure

```

1: for each UE do
2:   Set attempt count to 0
3:   Schedule first registration attempt at RAO
4: end for
5: while An unregistered UE exist do
6:   for Each unregistered UE do
7:     randomly select preamble  $p_i$ 
8:   end for
9:   for each preamble,  $p_i$  do
10:    if only one UE selected preamble  $p_i$  then
11:      Complete UE registration
12:    else
13:      for each UE that selected preamble  $p_i$  do
14:        Increase the attempt count for each UE by one.
15:        Schedule new attempt after fixed backoff.
16:      end for
17:    end if
18:  end for
19: end while
20: Collect and save statistics

```

Figure 4. Pseudocode for the simulation.

The goal of the simulation model and analysis is to determine when and if registration becomes problematic due to congestion. The focus is therefore on the number of registration attempts, collisions and their distributions under varying load conditions, without regard for the exact content of the exchanged messages. Although the initial synchronization and collection of network information via MIB, SIB1, SIB2 etc. are important, they do not experience contention and therefore aren't included in the simulation model. To determine the threshold where congestion becomes problematic the amount and distribution of collisions during the registration process is examined. In this context it's also important to determine if any of the UEs reach the maximum transmission threshold.

To validate that the output of the model is sensible, first a simple scenario is created with 10 UEs and the simulation runs for 100 different random seeds. After each simulation the number of devices that have used 1, 2, ... N attempts are logged and finally the combined result for all 100 seeds analyzed. Using this method, the results for 10 UE are obtained and can be seen in Figure 5. From Figure 5, the number of attempts required to register the UE on the network were 1, 2 and 3 for the 100 seeds.

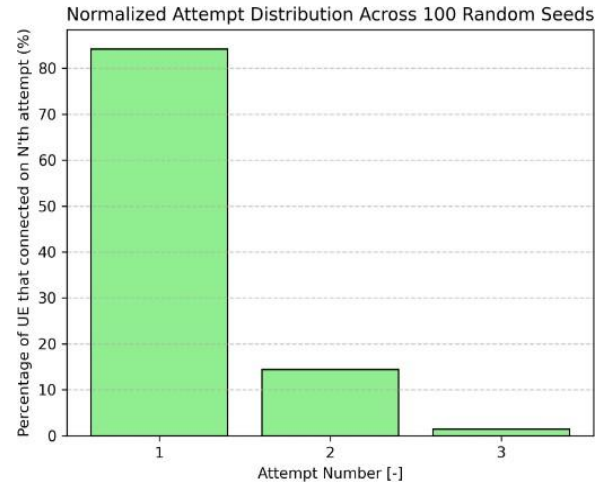


Figure 5. Probability of completing registration in the N-th attempt for 10 UEs.

Above 80 and 10 percent of the time, 1 and 2 attempts were sufficient. respectively, while around 2% of the time 3 attempts were needed. This is sensible as the chance of collision is not yet that high and therefore most UE register on the first attempt. There are also very few cases where 3 attempts are used. This is sensible, since the vast majority of devices complete registration in the first attempt, meaning the risk of collision in the second attempt is much smaller. Based on the same data used to generate Figure 5, a boxplot is made to get a better understanding of the underlying distribution. This is shown in Figure 6.

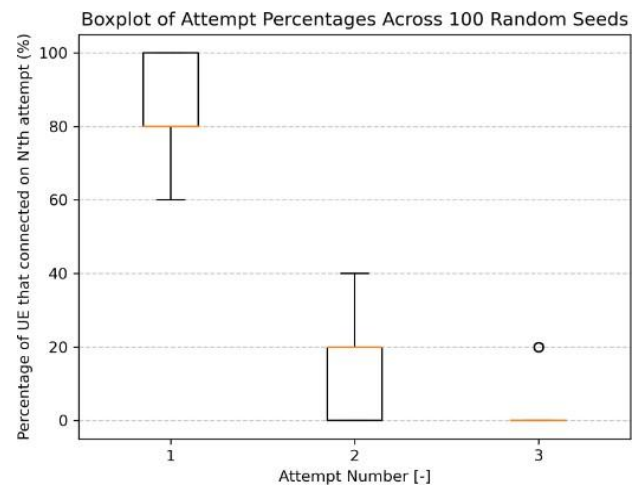


Figure 6. Boxplot of the probability of completing registration in the N-th attempt for 10 UEs.

As expected, the whiskers and the box edges are all a multiple of 10, because there are 10 UEs, so each UE corresponds to 10% in this scenario. In addition, as expected based on the low percentage for three attempts in Figure 5, it can be seen that Figure 6 contains an outlier.

IV. RESULTS AND DISCUSSION

To investigate the dynamics of collision and retransmission during the registration phase, multiple simulation scenarios were constructed, each with a varying number of concurrently registering UEs. These scenarios allowed for controlled observation of how increasing contention influences registration reliability. Particular attention was given to the number of attempts required for successful registration, as there is a maximum number of allowed attempts.

In order to conclude whether the amount of transmission is problematic, it is necessary to define the maximum amount of allowed preamble transmissions. The allowed values of the network parameter called PTM (PreambleTransMax) are defined in TS 38.331 [9] and shown in Table II below.

TABLE II. ALLOWED VALUES FOR PREAMBLETRANSMAX.

Enum Value	Maximum Attempts
n3	3
n4	4
n5	5
n6	6
n7	7
n8	8
n10	10
n20	20
n50	50
n100	100

It is the mobile network operator who decides which value of PTM is used in the network. With these values in mind the simulation is carried out for 10, 20, 30, 40, 60, 80 and 100 UE respectively to examine the impact on the number of registration attempts required. The results can be seen in Figure 7 below.

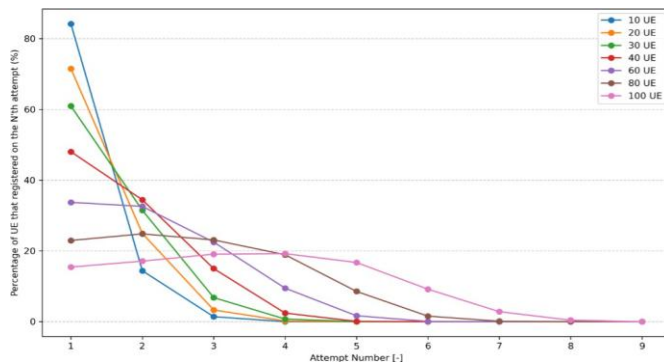


Figure 7. Probability of completing registration in the N-th attempt for 10, 20, 30, 40, 60, 80 and 100 UEs.

Based on Figure 7, it can be concluded that for 60 UE and lower, the chance of registering in one attempt is the highest. As the number of UEs increases beyond this number to 80 UEs, the risk of collision in the first attempt becomes so high that fewer devices register in the first attempt than the second one. The distribution of the number of attempts required can be seen in Figure 8 and is also available in Table III.

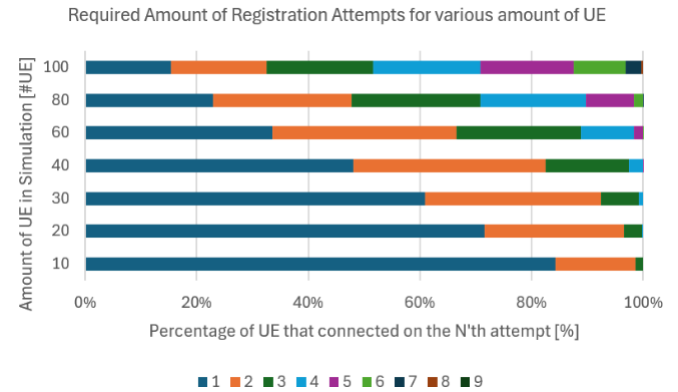


Figure 8. Distribution of the number of attempts required to register on the network for 10, 20, 30, 40, 60, 80 and 100 UEs.

From Figure 8, it can be seen that for 30 UE there is a risk of around 1% using 4 attempts, which is above the minimum number of preambles that can be configured, as seen in Table II. This behavior illustrates how an increase in UE density can lead to a nonlinear impact on access reliability.

TABLE III. PERCENTAGE OF UES ATTEMPT REQUIRED UNDER VARYING CONCURRENT LOAD.

Tries/UE	10	20	30	40	60	80	100
1	84.20	71.55	60.97	48.05	33.51	22.95	15.43
2	14.40	24.95	31.50	34.45	33.03	24.81	17.11
3	1.40	3.30	6.80	15.00	22.33	23.10	19.09
4	0.00	0.20	0.73	2.43	9.40	18.91	19.23
5	0.00	0.00	0.00	0.08	1.66	8.55	16.73
6	0.00	0.00	0.00	0.00	0.07	1.56	9.18
7	0.00	0.00	0.00	0.00	0.00	0.11	2.82
8	0.00	0.00	0.00	0.00	0.00	0.00	0.41

The results highlight the importance of anticipating access contention in scenarios with simultaneous or synchronized UE activation. In practice, such situations may arise during coordinated service restoration, mass device reboots, or mass event emergencies where deterministic performance is critical.

V. FUTURE WORKS

The current NS-3 model focuses on the core features and phenomenon in the registration process in 5G and therefore relies on some assumptions. In the future, we plan to look into the following aspects:

- *Physical Environment:* The current model does not take into account environmental factors like reflections or shadowing, which can significantly alter the time of reception and the quality of the signal received at the gNB. In the future, it would therefore be good to include a characterization of the physical environment.
- *Channel Environment:* For large numbers of UEs, it is expected that there will be significant interference between devices, which in theory could lead to none of the preambles arriving at the destination. A more complex model could take this into account.

- *Probabilistic Collisions*: There exist situations where multiple UE can select the same preamble, but the gNB can still decode one of the messages, for instance if one signal is stronger due to reflections. In the future, it would be good to simulate the chance of a message being decoded during Msg1 and Msg3 despite multiple UE selecting the same preamble.
- *UE Join/Leave Model*: In the current model, it is assumed a number of UE attempts to join during the same RAO and that no other terminals join this process. In a real environment, there is a risk of new devices joining, which both depends on the UE density and user behavior.
- *Practical experiment*: It is planned to conduct practical experiments in a real-world test environment. The experiments can help validate key assumptions in the model, reveal implementation-specific behavior, and provide a more complete understanding of how collisions and retries manifest under real operating conditions.
- *Prioritized Access*: Techniques to improve the reliability of the registration process for mission-critical devices may be improved through prioritization mechanisms. This could include strategies such as reserving a subset of preambles exclusively for high-priority UEs or applying differentiated RAO scheduling and are worth exploring.

VI. CONCLUSION

In this study, the behavior of the 5G registration process was evaluated under conditions where multiple UEs join the same random-access opportunity using an NS-3 based simulation model. It was demonstrated that even for few UEs, contention for resources during the random-access procedure can result in requiring multiple registration attempts due to collisions. It was found that the number of attempts required to register increased as a function of the number of UE and that when 20 UEs attempted registration concurrently, a statistical probability of approximately 0.2% was observed for requiring four registration attempts. This value is important as that exceeds the minimum selectable value for `preambleTransMax` as defined in the 3GPP specifications. These results suggest that the existing configuration and handling of simultaneous registration attempts may be insufficient to meet the reliability demands of mission-critical applications in high load scenarios. The access reliability of a 5G system depends both on the choice of the `preambleTransMax` value and on the collision probability, which is influenced by the number of UEs, RAOs and the available number of preambles.

Based on the findings of this study, it is recommended that further work investigate how these parameters can be tuned to match the needs of different deployment scenarios, particularly those with mission critical communication requirements. The results also highlight the importance of

the registration phase, which should be carefully considered in the design of future systems such as 6G.

ACKNOWLEDGMENT

This work was supported by Innovation Fund Denmark and the Center for Emergency Communication at Danish Emergency Management Agency.

REFERENCES

- [1] Z. Zhu, X. Li, and Z. Chu, "Three Major Operating Scenarios of 5G: eMBB, mMTC, URLLC," in *Intelligent Sensing and Communications for Internet of Everything*, 2022, pp. 15–76.
- [2] R. Ren, J. Wang, J. Yu, X. Zhu, X. Wan, and H. Lu, "Joint Resource Allocation for Multiplexing eMBB, URLLC and mMTC Traffics Based on DRL," in *Proc. IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, 2024.
- [3] M. Volk and J. Sterle, "5G Experimentation for Public Safety: Technologies, Facilities and Use Cases," *IEEE Access*, vol. 9, pp. 47537–47550, 2021, doi: 10.1109/ACCESS.2021.3064405.
- [4] S. Gamboa, T. R. Henderson, W. Garey, C. Liu, and R. Rouil, "Towards System Level Simulations of Public Safety Applications over 5G NR Sidelink," in *Proc. IEEE World Forum on Public Safety Technology (WFPST)*, Herndon, VA, USA, 2024, pp. 1–6.
- [5] Y. Wang, W. Liu, and L. Fang, "Adaptive Modulation and Coding Technology in 5G System," in *Proc. Int. Wireless Communications and Mobile Computing Conf. (IWCMC)*, Limassol, Cyprus, 2020.
- [6] 3GPP, "Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) Protocol for 5G System (5GS); Stage 3," TS 24.501, 3rd Generation Partnership Project.
- [7] 3GPP, "Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture," TS 23.203, 3rd Generation Partnership Project.
- [8] 3GPP, "Technical Specification Group Radio Access Network; NR; Medium Access Control (MAC) Protocol Specification," TS 38.321, 3rd Generation Partnership Project.
- [9] 3GPP, "Technical Specification Group Radio Access Network; NR; Radio Resource Control (RRC) Protocol Specification," TS 38.331, 3rd Generation Partnership Project.
- [10] D. Helbing, A. Johansson, and H. Z. Al-Abideen, "Dynamics of Crowd Disasters: An Empirical Study," *Phys. Rev. E*, vol. 75, no. 4, Art. no. 046109, 2007.
- [11] D. Helbing and P. Mukerji, "Crowd Disasters as Systemic Failures: Analysis of the Love Parade Disaster," *Saf. Sci.*, vol. 50, no. 1, pp. 166–180, 2012.
- [12] A. Johansson, D. Helbing, and P. K. Shukla, "Empirical Observations of Crowd Turbulence at the Roskilde Festival," *arXiv:2303.03977*, 2023.
- [13] World Bank, "Mobile Cellular Subscriptions (per 100 People) – Germany." [Online]. Available: <https://data.worldbank.org/indicator/>. Accessed: Jul. 3, 2026.

Evolutionary Transfer Optimization for Diversified Investment Using Predictive Distribution of Asset Returns

Kiyoharu Tagawa 

Department of Informatics
Faculty of Informatics, Kindai University
Higashi-Osaka, Osaka, Japan
e-mail: tagawa@info.kindai.ac.jp

Yukiko Orito

Department of Management Science
College of Engineering, Tamagawa University
Machida, Tokyo, Japan
e-mail: orito.y@eng.tamagawa.ac.jp

Abstract—Bayesian inference is a powerful mathematical tool that enables Artificial Intelligence (AI) to make probabilistic predictions and decisions in environments with incomplete data or high levels of noise. In this paper, the traditional portfolio optimization problem is extended to a data-driven one based on Bayesian inference. Specifically, a probabilistic model of asset returns is given by the predictive distribution and updated whenever a new data set of asset returns becomes available. Thus, each time the predictive distribution is updated, it becomes necessary to search for a new solution to the portfolio optimization problem. In order to solve the data-driven portfolio optimization problem iteratively, a novel evolutionary transfer optimization method is proposed and then its performance is evaluated using actual asset data from the Tokyo Stock Price Index (TOPIX17).

Keywords—Artificial intelligence; Bayesian inference; portfolio optimization; evolutionary transfer optimization.

I. INTRODUCTION

Evolutionary Transfer Optimization (ETO) is a paradigm that integrates Evolutionary Algorithm (EA) with knowledge learning and transfer to achieve higher optimization efficiency and performance [1]. Specific ETO techniques include various EAs for dynamic optimization problems [2][3], EAs for optimization problems in uncertain environments [4], EAs for complex and large-scale problems [5], and EAs based on surrogate models [6][7]. Recently, inspired by the success of neural network transfer learning, attention has shifted toward ETO techniques. These ETO methods leverage information from solutions or parameter values of specific optimization problems to solve similar problems more efficiently.

Portfolio optimization is the process of determining the best portion of investment in different assets according to some objective. Since portfolio optimization is one of the core issues in the field of finance, numerous reports have been published on both its theoretical and practical aspects [8][9]. It should be noted that the problem setting differs between investment and speculation. In speculative investment, asset returns are treated as time-series data. Consequently, Artificial Intelligence (AI) techniques, such as Transformers [10] and Bayesian networks [11], are employed for their prediction. On the other hand, in asset investment, the asset returns are treated as random variables from the Efficient-Market Hypothesis (EMH) [12]. Therefore, in portfolio optimization problems, probabilistic models are usually used to predict asset returns [8][9].

This paper extends the traditional portfolio optimization problem based on the EMH to a data-driven one. Specifically, the predictive distribution in Bayesian inference [13] is used as a probabilistic model for asset returns. In the real world, as new data on asset returns is observed, predictions of them are constantly changing. The predictive distribution allows the probabilistic model for asset returns to be updated according to new data without losing information inherited from historical data. Furthermore, to update the predictive distribution in response to new observed data on asset returns and solve the data-driven portfolio optimization problem iteratively, a novel ETO method using a powerful EA, namely an Adaptive Differential Evolution (ADE) algorithm [14], is proposed. The usefulness of the proposed ETO method is verified through the diversified investment using actual data on asset returns from the Tokyo Stock Price Index (TOPIX17) [15].

The rest of this paper is organized as follows. Section II formulates a portfolio optimization problem based on EMH. Section III extends the portfolio optimization problem to the data-driven one by using the predictive distribution of asset returns. Section IV proposes a novel ETO method for solving the data-driven portfolio optimization problem. Section V evaluates the performance of the proposed ETO method by using actual asset data. Section VI concludes this paper.

II. PORTFOLIO OPTIMIZATION

Let $x_i = \mathbb{R}$, $i = 1, \dots, D$ be the investment ratio for i -asset. The portfolio is defined as $\mathbf{x} = (x_1, \dots, x_i, \dots, x_D) \in \mathbb{R}^D$. One unit investment in i -asset yields a return $\xi_i \in \mathbb{R}$. From the EMH, the return $\xi_i \in \mathbb{R}$ of i -asset is considered to be a random variable following a Gaussian distribution. Thus, the vector $\boldsymbol{\xi} \in \mathbb{R}^D$ of them obeys a multivariate Gaussian distribution:

$$\boldsymbol{\xi} = (\xi_1, \dots, \xi_i, \dots, \xi_D) \sim \mathcal{N}(\boldsymbol{\xi} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (1)$$

where $\boldsymbol{\mu} \in \mathbb{R}^D$ denotes the mean vector, $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ is the covariance matrix, and $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is called the precision matrix.

From the asset returns $\boldsymbol{\xi} \in \mathbb{R}^D$ in (1), the return of the portfolio $\mathbf{x} \in \mathbb{R}^D$ is evaluated as $r = \boldsymbol{\xi}^T \mathbf{x}$. The return $r \in \mathbb{R}$ of the portfolio is also a random variable. Let $r_o \in \mathbb{R}$ be the target value for the return $r \in \mathbb{R}$. In the diversified investment considered in this paper, the optimal portfolio $\mathbf{x} \in \mathbb{R}^D$ that minimizes the risk $\alpha(\mathbf{x})$, or the probability of the return $r \in \mathbb{R}$

falling below the target value $r_o \in \mathbb{R}$, is sought. Therefore, the portfolio optimization problem is formulated as

$$\begin{cases} \min_{\mathbf{x}} & \alpha(\mathbf{x}) = \Pr(r = \boldsymbol{\xi}^T \mathbf{x} \leq r_o) \\ \text{sub. to} & x_1 + \dots + x_i + \dots + x_D = 1 \\ & 0 \leq x_i \leq 1, i = 1, \dots, D \end{cases} \quad (2)$$

where $\Pr(\mathcal{E})$ is the probability that event $\mathcal{E}$ occurs.

III. PROBLEM FORMULATION

A. Predictive Distribution

It is assumed that a new data set of asset returns $\boldsymbol{\xi}_n \in \mathbb{R}^D$, $n = 1, \dots, N$ is observed over a period t as

$$\Theta_t = \{\boldsymbol{\xi}_n \in \mathbb{R}^D \mid n = 1, \dots, N\}. \quad (3)$$

The observed data set in (3) can be obtained periodically. However, the data size N is too small to statistically evaluate the risk $\alpha(\mathbf{x})$ in (2). Hence, a probabilistic model, specifically the predictive distribution, is used to obtain sufficient data for evaluating the risk. The observed data on asset returns $\boldsymbol{\xi}_n \in \Theta_t$ in (3) obeys the Gaussian distribution in (1). Thus, the conjugate prior distribution for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ of the distribution in (1) is a Gaussian-Wishart distribution:

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \mathbf{m}, \beta, \nu, \mathbf{W}) \quad (4)$$

where $\mathbf{m} \in \mathbb{R}^D$, $\beta \in \mathbb{R}$, $\nu > D - 1$, and $\mathbf{W} \in \mathbb{R}^{D \times D}$ [13].

Let $p(\boldsymbol{\xi} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda})$ be the Probability Distribution Function (PDF) of the Gaussian distribution in (1). Let $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ be the PDF of the prior distribution in (4). From Bayes' theorem, the PDF of the posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \Theta_t)$ is

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \Theta_t) \propto \prod_{n=1}^N p(\boldsymbol{\xi}_n \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}). \quad (5)$$

From (5), the posterior distribution becomes a Gaussian-Wishart distribution depending on $\boldsymbol{\xi}_n \in \Theta_t$ in (3) as

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \hat{\mathbf{m}}, \hat{\beta}, \hat{\nu}, \hat{\mathbf{W}}) \quad (6)$$

where the parameters also depend on those in (4) as

$$\begin{cases} \hat{\beta} &= \beta + N \\ \hat{\mathbf{m}} &= \frac{1}{\hat{\beta}} \left(\sum_{n=1}^N \boldsymbol{\xi}_n + \beta \mathbf{m} \right) \\ \hat{\mathbf{W}}^{-1} &= \sum_{n=1}^N \boldsymbol{\xi}_n \boldsymbol{\xi}_n^T + \beta \mathbf{m} \mathbf{m}^T \\ &\quad - \hat{\beta} \hat{\mathbf{m}} \hat{\mathbf{m}}^T + \mathbf{W}^{-1} \\ \hat{\nu} &= \nu + N. \end{cases} \quad (7)$$

By marginalizing out the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ of Gaussian distribution $p(\boldsymbol{\xi} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda})$ with the posterior one in (5) as

$$p(\hat{\boldsymbol{\xi}} \mid \Theta_t) = \iint p(\hat{\boldsymbol{\xi}} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \Theta_t) d\boldsymbol{\mu} d\boldsymbol{\Lambda}, \quad (8)$$

the PDF of the predictive distribution $p(\hat{\boldsymbol{\xi}} \mid \Theta_t)$ is obtained. From (8), the predictive distribution becomes a multivariate Student's t -distribution depending on $\boldsymbol{\xi}_n \in \Theta_t$ in (3) as

$$\hat{\boldsymbol{\xi}} \sim \mathcal{S}(\hat{\boldsymbol{\xi}} \mid \boldsymbol{\mu}_s, \boldsymbol{\Lambda}_s, \nu_s) \quad (9)$$

where $\hat{\boldsymbol{\xi}} \in \mathbb{R}^D$ is the predicted value for the return $\boldsymbol{\xi} \in \mathbb{R}^D$ in (1). The parameters are obtained from those in (6) as

$$\begin{cases} \boldsymbol{\mu}_s &= \hat{\mathbf{m}} \\ \boldsymbol{\Lambda}_s &= \frac{(1 - D - \hat{\nu}) \hat{\beta}}{1 + \hat{\beta}} \hat{\mathbf{W}} \\ \nu_s &= 1 - D + \hat{\nu}. \end{cases} \quad (10)$$

B. Data-Driven Portfolio Optimization

By using the predictive distribution in (9), the predicted data set of asset returns is generated randomly as

$$\hat{\Theta}_t = \{\hat{\boldsymbol{\xi}}_m \in \mathbb{R}^D \mid m = 1, \dots, M\} \quad (11)$$

where the data size M , $M > N$ is large enough.

From the predicted data set of asset returns $\hat{\boldsymbol{\xi}}_m \in \hat{\Theta}_t$ in (11), the Empirical Cumulative Distribution Function (ECDF) for the return $r \in \mathbb{R}$ of the portfolio $\mathbf{x} \in \mathbb{R}^D$ is defined as

$$F(r \mid \mathbf{x}, \hat{\Theta}_t) = \frac{1}{M} \sum_{m=1}^M \mathbb{I}(\hat{\boldsymbol{\xi}}_m^T \mathbf{x} \leq r) \quad (12)$$

where $\mathbb{I}(\mathcal{E}) = 1$ if $\mathcal{E}$ is satisfied; otherwise $\mathbb{I}(\mathcal{E}) = 0$.

By using the ECDF in (12), the risk $\alpha(\mathbf{x})$ in (2) can be estimated from the predicted data set $\hat{\Theta}_t$ in (11) as

$$\alpha(\mathbf{x}) = \Pr(r = \boldsymbol{\xi}^T \mathbf{x} \leq r_o) \simeq F(r_o \mid \mathbf{x}, \hat{\Theta}_t) \quad (13)$$

From (13), the portfolio optimization problem in (2) is formulated as a data-driven optimization problem as

$$\begin{cases} \min_{\mathbf{x}} & \alpha(\mathbf{x}) = F(r_o \mid \mathbf{x}, \hat{\Theta}_t) \\ \text{sub. to} & x_1 + \dots + x_i + \dots + x_D = 1 \\ & 0 \leq x_i \leq 1, i = 1, \dots, D. \end{cases} \quad (14)$$

The ECDF in (12) is composed of the predicted data set $\hat{\Theta}_t$ in (11) which is generated by using the predictive distribution in (9). As will be discussed later, the predictive distribution $p(\hat{\boldsymbol{\xi}} \mid \Theta_t)$ in (8) incorporates information not only from the observed data $\boldsymbol{\xi}_n \in \Theta_t$ but also from all past data. Therefore, the optimal solution of the portfolio optimization problem in (14) depends on all the data observed up to the period t .

IV. EVOLUTIONARY TRANSFER OPTIMIZATION

A. Adaptive Differential Evolution

A powerful ADE algorithm called jDE [14] is used to solve the portfolio optimization problem in (14), which is defined by the data set $\hat{\Theta}_t$. ADE holds a set of individuals called the population $\mathbf{P}_{t,g}$. In each generation g , the current population $\mathbf{P}_{t,g}$ is updated to the next population $\mathbf{P}_{t,g+1}$. The termination condition of ADE is given by the maximum number of generations $\bar{g}$. Each individual $\mathbf{v}_k = (\mathbf{x}_k, F_k, CR_k) \in \mathbf{P}_{t,g}$, $k = 1, \dots, N_P$ in the population is composed of three

elements: $x_k \in \mathbb{R}^D$ is a candidate solution to the portfolio optimization problem in (14); $F_k \in [0, 1]$ and $CR_k \in [0, 1]$ are used for the self-adaptive control parameters of ADE.

In order to decide the individuals $\hat{v}_k \in P_{t,g+1}$ in the next generation based on the current generation's ones $v_k \in P_{t,g}$, $k = 1, \dots, N_P$, the following procedure is applied to each of the individuals $v_k \in P_{t,g}$, $k = 1, \dots, N_P$ in turn where $\varepsilon_j \in [0, 1]$ denotes the j th uniform random number.

Step 1: Three individuals $v_a, v_b, v_c \in P_{t,g}$, $a \neq b \neq c \neq k$ are selected at random from the current population $P_{t,g}$ excluding the target one $v_k = (x_k, F_k, CR_k)$.

Step 2: Scale factor $F \in [0, 1]$ is determined as

$$F = \begin{cases} 0.1 + 0.9\varepsilon_1 & \text{if } \varepsilon_2 < 0.1 \\ F_k & \text{otherwise.} \end{cases} \quad (15)$$

Step 3: Crossover rate $CR \in [0, 1]$ is determined as

$$CR = \begin{cases} \varepsilon_3 & \text{if } \varepsilon_4 < 0.1 \\ CR_k & \text{otherwise.} \end{cases} \quad (16)$$

Step 4: Elements $z_{k,i} \in \mathbb{R}$, $i = 1, \dots, D$ of a new solution $z_k \in \mathbb{R}^D$ are generated from the four solutions as

$$z_{k,i} = \begin{cases} x_{a,i} + F(x_{b,i} - x_{c,i}) & \text{if } \varepsilon_{4+i} < CR \\ x_{k,i} & \text{otherwise.} \end{cases} \quad (17)$$

Furthermore, the new solution $z_k = (z_{k,1}, \dots, z_{k,D})$ is normalized for the equality constraint in (14) as

$$z_{k,i} = \frac{z_{k,i}}{z_{k,1} + \dots + z_{k,D}}, \quad i = 1, \dots, D. \quad (18)$$

Step 5: Let $u_k = (z_k, F, CR)$ be a new individual.

Step 6: The next individual $\hat{v}_k \in P_{t,g+1}$ is selected as

$$\hat{v}_k = \begin{cases} u_k & \text{if } \alpha(z_k) \leq \alpha(x_k) \\ v_k & \text{otherwise} \end{cases} \quad (19)$$

where $u_k = (z_k, F, CR)$ and $v_k = (x_k, F_k, CR_k)$.

After repeating the above procedure until all individuals $\hat{v}_k \in P_{t,g+1}$ have been determined, the population is updated as $P_{t,g} = P_{t,g+1}$, $g = g + 1$. The best solution in the final population $P_{t,\bar{g}}$ is adopted as the solution to the portfolio optimization problem in (14), or the portfolio $x_t^* \in \mathbb{R}^D$.

B. Proposed ETO Method

In the proposed ETO method, ADE is applied iteratively to the portfolio optimization problem in (14) that is defined by the predicted data set $\hat{\Theta}_t$ in (11) for $t = 1, 2, \dots$. The data set Θ_t is generated by the predictive distribution $p(\hat{\xi} | \Theta_t)$ in (8). Each time a new data set Θ_t in (3) is observed in period t , the predictive distribution $p(\hat{\xi} | \Theta_t)$ is updated via the prior and posterior distributions. If newly observed asset returns $\xi_n \in \Theta_{t+1}$, $n = 1, \dots, N$ are considered to follow the current predictive distribution $p(\hat{\xi} | \Theta_t)$, some of superior individuals from the final population $P_{t,\bar{g}}$ are transferred to the next initial one $P_{t+1,0}$, and then ADE is applied to the portfolio optimization problem with the next data set $\hat{\Theta}_{t+1}$.

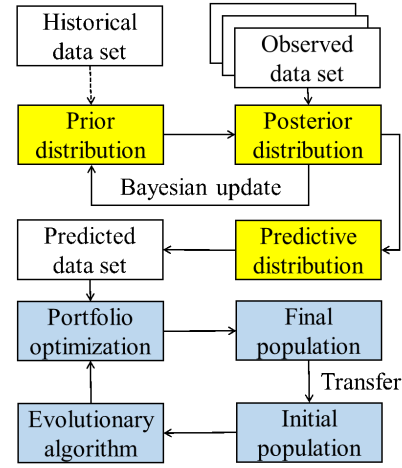


Figure 1. Flowchart of proposed ETO method.

Figure 1 shows the flowchart of the proposed ETO method to solve the data-driven portfolio optimization problem. The procedure of the ETO method is described as follows:

Step 1: Let $t = 1$. The individuals $v_k \in P_{t,0}$, $k = 1, \dots, N_P$ are generated randomly as the initial population.

Step 2: The first prior distribution $p(\mu, \Lambda)$ is obtained by the likelihood estimation [16] from the historical data.

Step 3: From the observed data set Θ_t in (3) and $p(\mu, \Lambda)$, the posterior distribution $p(\mu, \Lambda | \Theta_t)$ in (5) is obtained.

Step 4: From the posterior distribution in (5), the predictive one $p(\hat{\xi} | \Theta_t)$ for $\hat{\xi} \in \mathbb{R}^D$ is obtained as shown in (8).

Step 5: Using the predictive distribution in (9), the predicted data set of asset returns $\hat{\xi}_m \in \hat{\Theta}_t$ in (11) is generated.

Step 6: Using the ECDF in (12), the portfolio optimization problem in (14) is defined by the data set $\hat{\Theta}_t$ in (11).

Step 7: ADE is applied to the portfolio optimization problem. As a result, the best portfolio $x_t^* \in \mathbb{R}^D$ is obtained.

Step 8: Update the prior distribution in (4) by the posterior one in (6) as $m = \hat{m}$, $\beta = \hat{\beta}$, $\nu = \hat{\nu}$, and $W = \hat{W}$.

Step 9: If observed asset returns $\xi_n \in \Theta_{t+1}$ are considered to follow the current predictive distribution $p(\hat{\xi} | \Theta_t)$, the top $\tau \times 100\%$ individuals from the final population $P_{t,\bar{g}}$ are transferred to the next initial one $P_{t+1,0}$. Other individuals in $P_{t+1,0}$ are generated randomly. Otherwise, all individuals in the initial population $P_{t+1,0}$ are generated randomly.

Step 10: Let $t = t + 1$. Go back to Step 3.

C. Kolmogorov-Smirnov Test

In order to test whether the newly observed asset returns follow the current predictive distribution in Step 9 of the ETO method, the Kolmogorov-Smirnov test (K-S test) is used.

Let $x_t^* \in \mathbb{R}^D$ be the best solution obtained by ADE to the portfolio optimization problem in (14), which is defined by the predicted data set $\hat{\Theta}_t$. Using the same solution $x_t^* \in \mathbb{R}^D$, two ECDFs of the return $r \in \mathbb{R}$ are formed from the predicted asset returns $\hat{\xi}_m \in \hat{\Theta}_t$ and the observed ones $\xi_n \in \Theta_{t+1}$. From

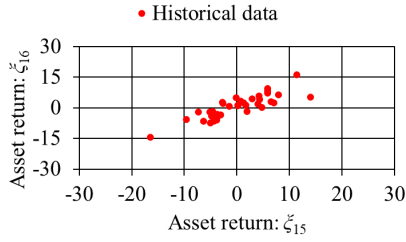


Figure 2. Historical data for prior distribution.

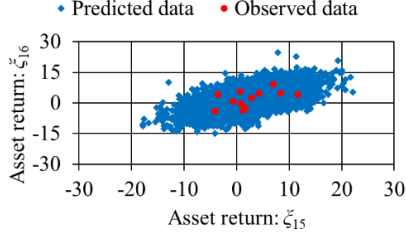
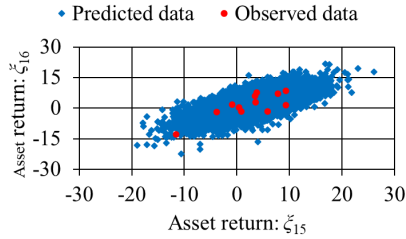
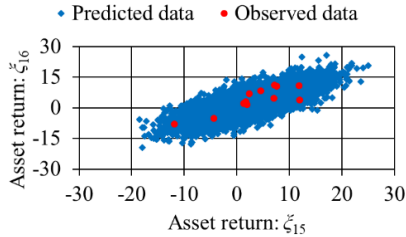

 (a) Period: $t = 1$

 (b) Period: $t = 2$

 (c) Period: $t = 3$

 Figure 3. Predicted $\hat{\xi}_m \in \hat{\Theta}_t$ and observed $\xi_n \in \Theta_{t+1}$.

the two ECDFs of $r \in \mathbb{R}$, the null hypothesis (H_0) and the alternative hypothesis (H_1) of the K-S test are defined as

$$H_0 : F(r_n | x_t^*, \hat{\Theta}_t) = F(r_n | x_t^*, \Theta_{t+1}), \quad n = 1, \dots, N$$

$$H_1 : F(r_n | x_t^*, \hat{\Theta}_t) \neq F(r_n | x_t^*, \Theta_{t+1}), \quad n = 1, \dots, N$$

where $r_n = \xi_n x_t^{*T} \in \mathbb{R}$ is the return of the portfolio $x_t^* \in \mathbb{R}^D$ evaluated by the newly observed asset returns $\xi_n \in \Theta_{t+1}$.

If the null hypothesis is not rejected, the newly observed asset returns $\xi_n \in \Theta_{t+1}$ are considered to follow the predictive distribution $p(\xi | \Theta_t)$ as well as the predicted asset returns $\hat{\xi}_m \in \hat{\Theta}_t$ in (11). On the other hand, if the null hypothesis is rejected, $\xi_n \in \Theta_{t+1}$ are not considered to obey $p(\xi | \Theta_t)$.

V. NUMERICAL EXPERIMENT

A. Experimental Setup

The monthly data from the Tokyo Stock Price Index (TOPIX17) over the past six years (2019 – 2024) are used for asset returns [15]. Therefore, the number of assets is $D = 17$. The first prior distribution $p(\mu, \Lambda)$ in Step 2 of the proposed ETO method is obtained from the monthly data on the first three years. The monthly data from the remaining three years are used for the observed data Θ_t , $t = 1, 2, 3$ in (3). Since monthly data is used, the data size of Θ_t is $N = 12$.

Figure 2 shows two asset returns $\xi_{n,15}$ and $\xi_{n,16} \in \mathbb{R}$ from the historical data on asset returns $\xi_n \in \mathbb{R}^{17}$, $n = 1, \dots, 36$. A positive correlation is observed between them. In the same way, Figure 3 shows the observed data $\xi_n \in \Theta_t$, $N = 12$ in (3) and the predicted data $\hat{\xi}_m \in \hat{\Theta}_t$, $M = 10^4$ in (11) for the periods $t = 1, 2, 3$. From the data size of $\hat{\Theta}_t$, the margin of error in the probability estimate is around $1/\sqrt{M} = 0.01$.

B. Performance of ETO method

The proposed ETO method is applied to the portfolio optimization problems in (14) defined by the three predicted data sets $\hat{\Theta}_t$, $t = 1, 2, 3$ and the target value $r_o = 0.01$.

The population size of ADE is $N_P = 100$. The maximum number of generations is chosen as $\bar{g} = 100$. For the transfer rate τ , five values are selected from the range 0 to 1. If $\tau = 0$ then no individual is transferred from $P_{t,\bar{g}}$ to $P_{t+1,0}$. If $\tau = 1$ then all individuals are transferred from $P_{t,\bar{g}}$ to $P_{t+1,0}$.

Figure 4 compares the two ECDFs of returns $r \in \mathbb{R}$ used for the K-S test: $F(r | x_t^*, \hat{\Theta}_t)$ and $F(r | x_t^*, \Theta_{t+1})$, $t = 1, 2$. The best solution $x_t^* \in \mathbb{R}^D$ is obtained with $\tau = 0.25$.

The null hypothesis of the K-S test has not been rejected for every value of τ at a significance level: 0.05. Therefore,

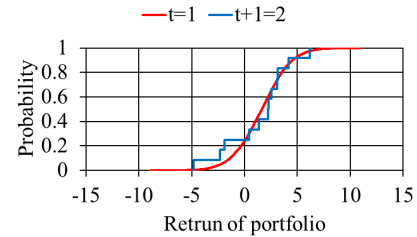
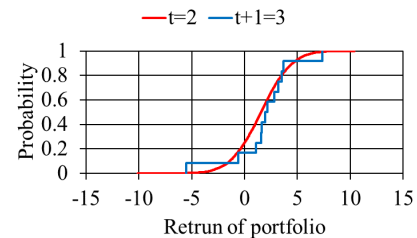

 (a) $F(r | x_1^*, \hat{\Theta}_1)$ and $F(r | x_1^*, \Theta_2)$

 (b) $F(r | x_2^*, \hat{\Theta}_2)$ and $F(r | x_2^*, \Theta_3)$

 Figure 4. Two ECDFs of $r \in \mathbb{R}$ for K-S test.

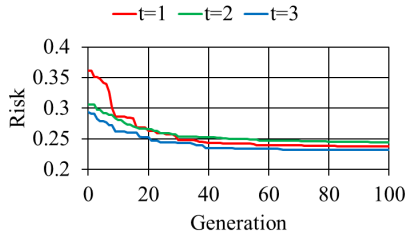
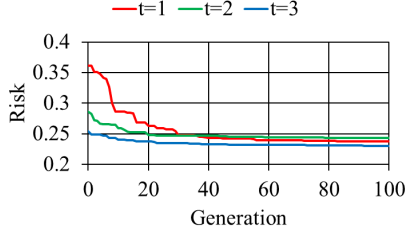
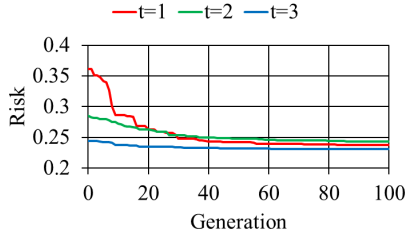

 (a) Transfer rate: $\tau = 0$

 (b) Transfer rate: $\tau = 0.25$

 (c) Transfer rate: $\tau = 1$

Figure 5. Convergence graphs of ADE.

the top $\tau \times 100\%$ individuals from the final population $P_{t,\bar{g}}$ are transferred to the next initial one $P_{t+1,0}$ for $t = 1, 2$.

Figure 5 shows convergence graphs of ADE using three transfer rates. The horizontal axis is the generation g of ADE. The vertical axis is the risk $\alpha(x_k^*)$ of the best solution $x_k^* \in \mathbb{R}^D$ obtained by ADE in each generation. From Figure 5, we can confirm that the best solutions $x_k^* \in \mathbb{R}^D$ in $P_{t,g}$ have been converged sufficiently at the final generation $\bar{g} = 100$.

From Figure 5, the effect of the proposed ETO method is confirmed. The shapes of the convergence graphs for the period $t = 1$ are almost identical. On the other hand, the shapes for the period $t \geq 2$ vary depending on the transfer rate τ . This dependency arises because the solutions $x_k \in \mathbb{R}^D$ in the initial population $P_{t,0}$ fluctuate according to the value of τ .

Table I shows the risks $\alpha(x_t^*)$, $t = 2, 3$ of the best solutions $x_t^* \in \mathbb{R}^D$ averaged over 20 runs. The variance of the risks is also shown in parenthesis. From Table I, the risk $\alpha(x_t^*)$ of the solution $x_t^* \in \mathbb{R}^D$ obtained by ADE appears to depend on the transfer rate τ . Hence, the Kruskal-Wallis test (K-W test) is used to determine whether there is a significant difference between the risks resulting from different τ values. The results of K-W test indicate that the risks in Table I significantly depend on the values of τ in both of $t = 2$ and $t = 3$.

The Steel-Dwass test (S-D test) is designed for comparing

TABLE I. RISK OF BEST SOLUTION

τ	$t = 2$	$t = 3$
0	0.244 (6.72×10^{-7})	0.231 (1.09×10^{-7})
0.25	0.243 (1.71×10^{-7})	0.230 (1.69×10^{-7})
0.5	0.243 (4.26×10^{-7})	0.231 (1.99×10^{-7})
0.75	0.243 (3.39×10^{-7})	0.230 (1.24×10^{-7})
1	0.244 (1.64×10^{-6})	0.230 (3.67×10^{-8})

TABLE II. STEEL-DWASS TEST ABOUT RISK

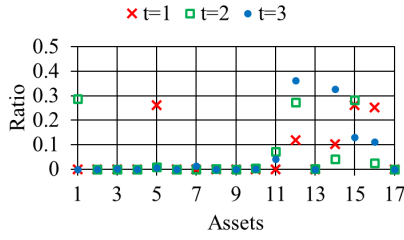
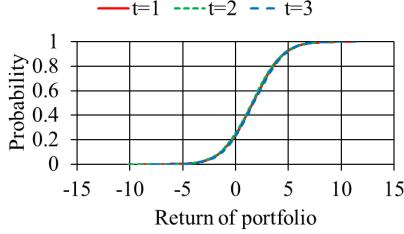
τ	τ	$t = 2$	$t = 3$
0	0.25	33.5 (0.001)	37.975 (0.001)
	0.5	19.475 (0.214)	23.9 (0.048)
	0.75	37.475 (0.001)	25.675 (0.004)
	1	3.925 (0.998)	52.825 (0.000)
0.25	0.5	-14.025 (0.594)	-14.075 (0.466)
	0.75	3.975 (0.938)	-12.3 (0.598)
	1	-29.575 (0.008)	14.85 (0.447)
0.5	0.75	18 (0.263)	1.775 (0.999)
	1	-15.55 (0.455)	28.925 (0.017)
0.75	1	-33.55 (0.006)	27.15 (0.002)

multiple groups to identify which specific groups differ from each other. Table II shows the results of S-D test about the risks $\alpha(x_t^*)$, $t = 2, 3$ in Table I. Table II shows the difference in risk ranking obtained with the left and right τ values. The greater the difference in ranking, the lower the risk obtained with the right τ value. The results of S-D test are also shown by p -value in parenthesis. From Table II, the best solution obtained by the ETO method with $\tau = 0.25$ is clearly better than the best solution obtained with $\tau = 0$. No individual is transferred if $\tau = 0$. However, $\tau = 0.25$ is competitive with $\tau = 0.75$ in terms of the quality of the best solution.

From Table I and Table II, it can be confirmed that the transfer of superior solutions performed by the ETO method enhances the quality of the solution obtained by ADE.

C. Analysis of Portfolio

The best portfolios $x_t^* \in \mathbb{R}^D$, $t = 1, 2, 3$ for the data-driven portfolio optimization problem in (14) are obtained by the proposed ETO method with the transfer rate $\tau = 0.25$. The investment ratios $x_i \in \mathbb{R}$, $i = 1, \dots, D$ provided by the three best portfolios $x_t^* \in \mathbb{R}^D$, $t = 1, 2, 3$ are shown in Figure 6. From Figure 6, the best portfolio $x_t^* \in \mathbb{R}^D$ varies depending on the period t . However, most of the assets with a zero ratio

Figure 6. Best portfolio $\mathbf{x}_t^* \in \mathbb{R}^D$, $D = 17$.Figure 7. ECDF $F(r | \mathbf{x}_t^*, \hat{\Theta}_t)$ by best portfolio $\mathbf{x}_t^* \in \mathbb{R}^D$.

($x_i = 0$) are common to all three portfolios. In other words, these portfolios correctly identify non-performing assets.

Figure 7 shows the ECDFs of returns $r \in \mathbb{R}$ by the three best portfolios: $F(r | \mathbf{x}_t^*, \hat{\Theta}_t)$, $t = 1, 2, 3$. From Figure 7, there is no significant difference in the shapes of ECDFs. Therefore, the risk $\alpha(\mathbf{x}_t^*)$ achieved by the best portfolio $\mathbf{x}_t^* \in \mathbb{R}^D$ does not differ drastically across the three periods $t = 1, 2, 3$.

From Figure 6 and Figure 7, the structure of the best solution $\mathbf{x}_t^* \in \mathbb{R}^D$ to the data-driven portfolio optimization problem examined in this paper does not change significantly over the three periods. As a result, the transfer of superior solutions performed by the ETO method has worked effectively.

VI. CONCLUSION AND FUTURE WORK

By using the predictive distribution for asset returns, the traditional portfolio optimization problem has been extended to a data-driven one. As shown in Figure 1, the predictive distribution is updated via the prior and posterior distributions whenever a new observed data set becomes available. As a result, information from all past data is aggregated into the probabilistic model of asset returns. This paper presents a novel ETO method utilizing ADE to effectively address the data-driven portfolio optimization problem. In the ETO method, superior solutions from the final population are transferred to the next initial population. Finally, the usefulness of the ETO method has been demonstrated by using actual asset data.

In future work, we will adaptively determine the transfer rate for the proposed ETO method from the state of the population. Furthermore, the proposed ETO method is expected to be applied to a variety of data-driven optimization problems.

ACKNOWLEDGMENT

This work was partially supported by JSPS (Japan Society for Promotion Science) KAKENHI, 22K12187, 24K15113.

REFERENCES

- [1] K. C. Tan, L. Feng, and M. Jiang, "Evolutionary transfer optimization - a new frontier in evolutionary computation research," *IEEE Computational Intelligence Magazine*, vol. 16, no. 1, pp. 22–33, 2021.
- [2] T. T. Nguyen, S. Yang, and J. Branke, "Evolutionary dynamic optimization - a survey of the state of the art," *Swarm and Evolutionary Computation*, vol. 6, pp. 1–24, 2012.
- [3] S. Das, A. Mandal, and R. Mukherjee, "An adaptive differential evolution algorithm for global optimization in dynamic environments," *IEEE Trans. on Cybernetics*, vol. 44, no. 6, pp. 966–978, 2024.
- [4] Y. Jin and J. Branke, "Evolutionary optimization in uncertain environments - a survey," *IEEE Trans. on Evolutionary Computation*, vol. 9, no. 3, pp. 303–317, 2005.
- [5] M. Iqbal, W. N. Browne, and M. Zhang, "Reusing building blocks of extracted knowledge to solve complex, large-scale Boolean problems," *IEEE Trans. on Evolutionary Computation*, vol. 18, no. 4, pp. 465–480, 2014.
- [6] D. Zhan and H. Xing, "A fast kriging-assisted evolutionary algorithm based on incremental learning," *IEEE Trans. on Evolutionary Computation*, vol. 25, no. 5, pp. 941–955, 2021.
- [7] Y. Liu, J. Liu, J. Ding, S. Yang, and Y. Jin, "A surrogate-assisted differential evolution with knowledge transfer for expensive incremental optimization problems," *IEEE Trans. on Evolutionary Computation*, vol. 28, no. 4, pp. 1039–1053, 2024.
- [8] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann, *Modern Portfolio Theory and Investment Analysis (9th ed.)* John Wiley & Sons, 2014.
- [9] M. Mokhtar, A. Shuib, and D. Mohamad, "Mathematical programming models for portfolio optimization problem: A review," *International Journal of Mathematical and Computational Sciences*, vol. 8, no. 2, pp. 428–435, 2014.
- [10] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep transformer model," *Expert Systems with Applications*, vol. 208, pp. 1–10, 2022.
- [11] A. B. Mard, B. Hnich, A. Lahiani, and S. Mefteh-Wali, "Comprehensive stock market insight: Bayesian networks for multi-output forecasting," *Computational Economics*, vol. 66, no. 5, pp. 4329–4349, 2025.
- [12] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.
- [13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [14] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer, "Self-adaptive control parameters in differential evolution: Comparative study on numerical benchmark problems," *IEEE Trans. on Evolutionary Computation*, vol. 10, no. 6, pp. 646–657, 2006.
- [15] Kabutan, <https://kabutan.jp>, Accessed: 2025/01/20.
- [16] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference*. John Wiley & Sons, 2018.

LiDAR-Based Simultaneous Localization, Mapping, and Multiobject Tracking: Combining Semantic Segmentation and Interacting Multiple Model-Extended Object Tracking

Akio Okumura

Graduate School of Science and Engineering
Doshisha University
Kyotanabe, Kyoto, Japan
e-mail: ctwn0155@mail4.doshisha.ac.jp

Masafumi Hashimoto, Kazuhiko Takahashi

Faculty of Science and Engineering
Doshisha University
Kyotanabe, Kyoto, Japan
e-mail: {mhashimo, katakaha}@mail.doshisha.ac.jp

Abstract—This paper proposes a method for Simultaneous Localization And Mapping (SLAM) and MultiObject Tracking (MOT) using an onboard Light Detection And Ranging (LiDAR) system in dynamic road environments. LiDAR-based SLAM involves Normal Distributions Transform (NDT)-graph SLAM, and MOT combines an Interacting Multiple Model estimator with Extended Object Tracking (IMM-EOT). To effectively handle dynamic scenes, semantic segmentation using RangeNet++ classifies LiDAR measurements into those related to stationary or moving objects. NDT-graph SLAM uses the measurements related to stationary objects for robust mapping and localization, whereas IMM-EOT uses the measurements related to moving and potentially moving objects to simultaneously estimate the motions and shapes of objects exhibiting various motion patterns. The effectiveness of the proposed method is demonstrated through experiments on the SemanticKITTI and KITTI tracking datasets.

Keywords— *LiDAR; NDT-graph SLAM; multiobject tracking; semantic segmentation; IMM extended object tracking.*

I. INTRODUCTION

Many Advanced Driver Assistance Systems (ADAS) and Autonomous Driving Systems (ADS) have been developed for Intelligent Transportation Systems (ITS) [1][2]. Simultaneous Localization And Mapping (SLAM) and MultiObject Tracking (MOT) using onboard sensors—such as cameras, Light Detection And Ranging (LiDAR) sensors, and radars—are key technologies that support ADAS and ADS [3][4]. This study focuses on LiDAR-based SLAM and MOT (SLAMMOT).

Typical LiDAR SLAM methods rely on scan-matching-based SLAM, which employs the iterative closest-point algorithm and the Normal Distributions Transform (NDT) [5]. However, the accuracy of scan-matching-based LiDAR SLAM degrades in dynamic environments with many coexisting moving objects, such as cars and two-wheelers.

Traditionally, LiDAR measurements related to moving objects were removed using the occupancy-grid method [6], which classifies LiDAR measurements into stationary objects (hereinafter “stationary LiDAR measurements”) and moving objects (hereinafter “moving LiDAR measurements”) and uses them for SLAM and MOT, respectively. However, occupancy-grid-based classification often fails to discriminate among different types of moving objects and frequently misdetects potentially moving objects—such as parked vehicles or vehicles stopping at red lights—as stationary objects. These limitations degrade the performance of ADAS and ADS.

Graph SLAM [7] is a popular method used for avoiding errors that accumulate over the travel time of the ego vehicle in scan-matching-based SLAM. A key component of graph SLAM is loop detection, which identifies previously visited locations. Many loop-detection methods have been proposed [5].

Conventional loop-detection methods rely on geometric point features, such as Fast Point Feature Histograms (FPFH) [8] or geometric surface-shape features (e.g., plains and poles) [9]. However, these geometric features cannot reliably distinguish objects with similar shapes; for example, traffic signs, utility poles, and trees are often confused, leading to false loop detection.

Recent studies have sought deep-learning-based object-recognition methods [10]. For example, the semantic segmentation method enhances the accuracy and robustness of LiDAR SLAM, and many semantic SLAM methods have been proposed [11][12][13]. Nevertheless, effectively integrating semantic information into LiDAR SLAM remains challenging.

Conventional MOT methods typically follow a two-step estimation sequence. First, object shapes are estimated as bounding boxes based on LiDAR measurements. Then, object motion parameters, such as the position and velocity, are estimated from the center position of the bounding box using Bayesian filters, including Kalman and particle filters [12] [14]. In this approach, the accuracy of motion parameter estimates strongly depends on the accuracy of shape estimation, and uncertainties (error covariances) in motion parameter and shape estimates are inaccurately calculated. Furthermore, because most conventional MOT methods assume nearly constant object velocities [11], their tracking performance degrades during sudden motion changes, such as sudden starts or stops. These limitations reduce the reliability of MOT in ADAS and ADS.

Extended Object Tracking (EOT) methods [15][16] resolve these limitations by simultaneously estimating motion parameters and shapes of objects, along with their uncertainties. The Interacting Multiple Model (IMM) framework has been combined with EOT (IMM-EOT) to accurately track objects exhibiting various behaviors, including stopping, near-constant-velocity motion, and sudden accelerations or decelerations, in the ITS domain [17][18].

Hence, this paper presents a SLAMMOT method for vehicle-mounted LiDAR systems in dynamic road environments. Extending our previous works [13][18], the proposed method integrates semantic segmentation and IMM-EOT.

In [13], semantic information was used only for loop detection in LiDAR SLAM, and the moving LiDAR measurements were removed using occupancy-grid method. This approach retained potentially moving LiDAR measurements of parked or temporarily stopped vehicles in the environment maps. This paper applies semantic segmentation to remove both moving and potentially moving LiDAR measurements, thereby improving SLAM performance.

Furthermore, the IMM-EOT-based object tracking method reported in our previous work [18] utilized roadside LiDAR. When the LiDAR is fixed at an elevated position (e.g., at the top of a traffic-light pole), objects can be easily detected using background subtraction. Rather than evaluating the tracking performance through simulation experiments alone [18], this paper evaluates the newly proposed method on real-world open datasets. Using vehicle-mounted LiDAR, our MOT scheme integrates semantic segmentation and IMM-EOT. To the best of our knowledge, these have not been simultaneously implemented in LiDAR MOT.

The structure of this article is as follows: Section II overviews the LiDAR SLAMMOT approach, and Sections III and IV explain loop detection and MOT methods, respectively. Section V presents the experimental evaluation using the SemanticKITTI and KITTI tracking datasets [19][20]. The paper concludes with Section VI.

II. OVERVIEW OF LiDAR SLAMMOT

Figure 1 shows a sequence of our developed LiDAR SLAMMOT system. Semantic segmentation is performed based on the point-cloud segmentation model RangeNet++ [21]. LiDAR SLAM combines NDT SLAM with graph SLAM (NDT-graph SLAM), whereas MOT employs IMM-EOT, which integrates the Multiplicative Error Model-Extended Kalman Filter (MEM-EKF) framework [16] with an IMM estimator (IMM-MEM-EKF) [18].

RangeNet++ was selected because it requires less memory and computational cost than voxel-based and point-based semantic segmentation methods [22][23]. Among several EOT methods, including random matrix and random hypersurface model-based methods [15], we selected the MEM-EKF [16] because it estimates the object motion and shape states, along with their uncertainties (error covariances), within the Kalman filter framework.

LiDAR measurements are typically captured by scanning laser beams over the target area. Vehicle movement introduces motion artifacts (distortions) in the LiDAR measurements, degrading the SLAMMOT accuracy. Such motion artifacts are corrected with an unscented Kalman filter-based algorithm [24]. The corrected LiDAR measurements are then processed by RangeNet++ to obtain the semantic information. We consider 12 semantic classes: people, cars, two-wheelers, other vehicles, buildings, roads, fences, vegetation, lawns, trunks, traffic signs, and poles. After mapping the LiDAR measurements with their semantic labels onto a voxel map, the semantic class of each voxel is determined by majority voting of the semantic classes of the measurements.

LiDAR measurements related to moving objects, such as cars, two-wheelers, and people, which degrade the accuracy of NDT SLAM, are removed based on their semantic labels, leaving only the stationary LiDAR measurements for SLAM.

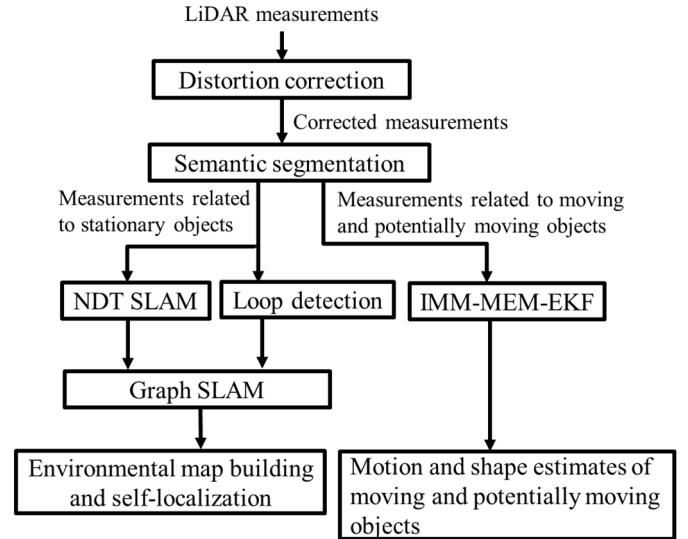


Figure 1. Sequence of LiDAR SLAMMOT.

The accuracy of NDT SLAM also deteriorates over time as errors accumulate. This error is mitigated by the loop-detection method within the graph SLAM framework [7]. Moving and potentially moving LiDAR measurements detected through the semantic segmentation method are applied to object tracking using the IMM-MEM-EKF method.

III. LOOP DETECTION AND RELATIVE POSE ESTIMATE

A. Loop Detection

Detection of loops (revisited locations) is a crucial problem in graph SLAM. Loops are selected from a set of loop candidates. To identify loop candidates, we define two feature descriptors, $F = (f_1, f_2, \dots, f_6)$ and $G = (g_1, g_2, \dots, g_6)$. Here, F and G are obtained from LiDAR measurements captured at the old and current locations of a vehicle, respectively, and f_i and g_i ($i = 1, 2, \dots, 6$) denote the numbers of objects belonging to the i -th semantic class of stationary objects (buildings, fences, poles, traffic signs, vegetation, and trunks). Loop candidates are then identified based on the following similarity indicator:

$$\text{Similarity indicator} = \frac{\sum_{i=1}^6 \{\max(f_i, g_i) - |f_i - g_i|\}}{\sum_{i=1}^6 \max(f_i, g_i)} \quad (1)$$

When the similarity indicator is high, the old locations of the vehicle are identified as loop candidates.

Next, loops are selected from the identified loop candidates. NDT SLAM estimates the relative pose of the vehicle from the stationary LiDAR measurements captured at each loop candidate and at the current location. The overlap indicator between the two sets of stationary LiDAR measurements is then calculated as

$$\text{Overlap indicator} = \frac{1}{n} \sum_{i=1}^n \text{overlap}(d_i) \quad (2)$$

where n denotes the number of stationary LiDAR measurements captured at the current location, and d_i denotes the nearest-neighbor distance between the measurements at the current location and those at each loop candidate. If d_i is shorter than 0.1 m, the overlap function (d_i) is set to one, otherwise, it is set to zero.

Finally, loop candidates with high overlap-indicator values are accepted as loops.

B. Relative Pose Estimate

The relative poses of the vehicle at the current location and at loops are calculated by NDT scan matching and provided as loop constraints in graph SLAM. Because NDT scan matching is a local registration method, an accurate initial estimation of the relative pose (i.e., coarse registration) is required to avoid local minima problems. In this paper, coarse registration is performed based on Semantic and Geometric point Feature Histograms (SGFH) [13].

First, the stationary LiDAR measurements captured at the current vehicle location are mapped onto a voxel grid map and downsampled using a voxel grid filter. As shown in Figure 2, the centroid (called the feature point) of the i -th voxel is denoted as A_i and the feature point of the surrounding j -th voxel ($j = 1, 2, \dots, n$) is denoted as A_j . After calculating the geometric three-angle features α_j , β_j , and γ_j at each feature point, a geometric point feature histogram (33 dimensions in this study) is obtained based on FPFH [8].

A semantic feature histogram is also constructed using the semantic labels of the six stationary semantic classes: buildings, fences, poles, traffic signs, vegetation, and trunks. Semantic labels are assigned to the feature points A_i and A_j ($i, j = 1, 2, \dots, n$). The six-dimensional Semantic Feature Histogram (SFH) related to A_i is then given by

$$SFH(A_i) = SH(A_i) + \frac{1}{n} \sum_{j=1}^n w_j SH(A_j) \quad (3)$$

where $SH(A_i)$ and $SH(A_j)$ denote the SFHs at the feature points A_i and A_j , respectively. The weight w_i is given as the inverse of the distance between A_i and A_j .

Finally, the six-dimensional SFH is combined with the 33-dimensional geometric feature histogram to form the 39-dimensional SGFH descriptor. The same procedure is applied to the stationary LiDAR measurements captured at each loop to obtain the feature points (B_i and B_j) and SGFH descriptors.

Feature points A_i and B_i are then matched using a random sample consensus-based method [25], completing the coarse

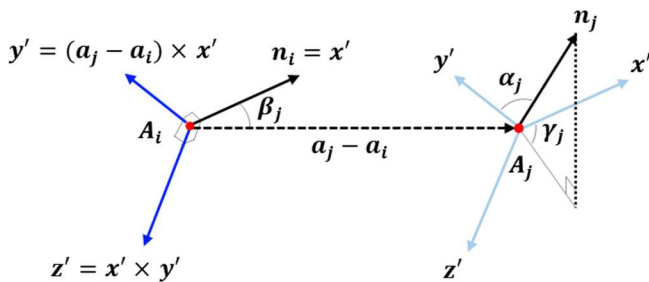


Figure 2. Feature points and angle features in FPFH.

registration. The relative pose between the current and loop locations, refined by NDT scan matching, is subsequently used as a loop constraint in graph SLAM to correct the accumulated error in NDT SLAM.

IV. MULTI-OBJECT TRACKING

A. Motion, Shape, and Measurement Models

The moving LiDAR measurements extracted using the semantic segmentation method are clustered via density-based spatial clustering of applications with noise [26]. To track moving objects, the clustered moving LiDAR measurements are processed using the IMM-MEM-EKF method. To accommodate diverse object motions in real road environments, the following motion modes of target objects are considered:

- Mode 1: Constant-position mode for temporarily stationary objects.
- Mode 2: Constant-velocity mode for an object moving at a nearly constant linear or turning velocity.
- Mode 3: Acceleration mode for objects that suddenly accelerate or decelerate.

The motion state in mode m ($m = 1, 2, 3$) is modeled by the following state equation [18]:

$$\mathbf{r}_{m,t} = \mathbf{f}_m(\mathbf{r}_{m,t-1}, \Delta \mathbf{r}_{m,t-1}) \quad (4)$$

where t and $t-1$ denote the time steps. $\mathbf{r}_m$ denotes the motion-state variable in mode m , and $\Delta \mathbf{r}_m$ is the related disturbance with covariance matrix $\mathbf{C}_m^{\Delta \mathbf{r}}$. $\mathbf{f}_m$ is a function of $\mathbf{r}_m$ and $\Delta \mathbf{r}_m$.

The motion-state variables for modes 1, 2, and 3 are $\mathbf{r}_1 = (x, y)^T$, $\mathbf{r}_2 = (x, y, \theta, v, \omega)^T$, and $\mathbf{r}_3 = (x, y, \theta, v, \dot{v})^T$, respectively, where (x, y) and θ denote the position and moving direction of an object, respectively, in the world coordinate system (Figure 3). Further, v and ω denote the linear and turning velocities, respectively, and $\dot{v}$ denotes the linear acceleration.

The shape state of an object is modeled as follows:

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \Delta \mathbf{p}_{t-1} \quad (5)$$

where $\mathbf{p} = (\phi, l_1, l_2)^T$ denotes the shape-state variable, consisting of the heading angle ϕ and object size (l_1, l_2) . $\Delta \mathbf{p}$ is the related disturbance with covariance matrix $\mathbf{C}^{\Delta \mathbf{p}}$.

Denoting N_t as the number of moving LiDAR measurements of an object at time t ($t = 0, 1, 2, \dots$) and $\mathbf{y}_t^{(i)} = (y_{1t}^{(i)}, y_{2t}^{(i)})^T$ as the i -th measurement ($i = 1, 2, \dots, N_t$) in the

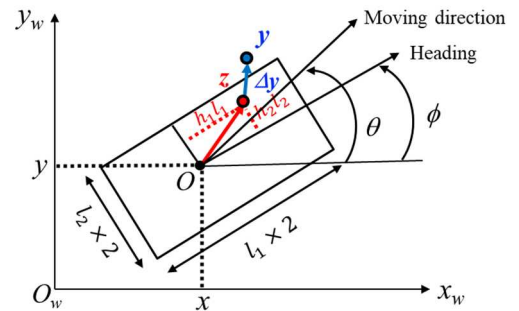


Figure 3. Motion and shape-state variables (top view).

world coordinate system (Figure 3), the measurement model is given by

$$\mathbf{y}_t^{(i)} = \mathbf{H}_m \mathbf{r}_{m,t} + \mathbf{S}_t \mathbf{h}_t^{(i)} + \Delta \mathbf{y}_t^{(i)} \quad (6)$$

where $\mathbf{H}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{H}_2 = \mathbf{H}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$, and

$$\mathbf{S}_t = \begin{pmatrix} S_{1t} \\ S_{2t} \end{pmatrix} = \begin{pmatrix} \cos \phi_t & -\sin \phi_t \\ \sin \phi_t & \cos \phi_t \end{pmatrix} \begin{pmatrix} l_{1t} & 0 \\ 0 & l_{2t} \end{pmatrix}.$$

$\mathbf{h}_t^{(i)} = (h_1^{(i)}, h_2^{(i)})^T$ are uniformly distributed random variables in the interval $-1 \leq h_1^{(i)} \leq 1$ and $-1 \leq h_2^{(i)} \leq 1$. The variable $\mathbf{h}_t^{(i)}$ is assumed as Gaussian white noise with zero mean and covariance $\mathbf{C}^h = \mathbf{E}_2 / 3$, where $\mathbf{E}_2$ is the two-dimensional identity matrix. $\Delta \mathbf{y}_t^{(i)}$ denotes the measurement noise with related covariance matrix $\mathbf{C}^{\Delta y}$.

B. Motion Estimation of Object

Let $\hat{\mu}_{m,t-1}^{(N_{t-1})}$ be the probability of mode m ($m = 1, 2, 3$) occurring at time $t - 1$. The conditional state estimate and its related error covariance for mode m are denoted by $\hat{\mathbf{r}}_{m,t-1}^{(N_{t-1})}$ and $\mathbf{C}_{m,t-1}^{r(N_{t-1})}$, respectively. These quantities are mixed as

$$\left. \begin{aligned} \bar{\mathbf{r}}_{m,t-1}^{(N_{t-1})} &= \sum_{n=1}^3 \alpha_{mn} \hat{\mathbf{r}}_{n,t-1}^{(N_{t-1})} \\ \bar{\mathbf{C}}_{m,t-1}^{r(N_{t-1})} &= \sum_{n=1}^3 \alpha_{mn} [\mathbf{C}_{n,t-1}^{r(N_{t-1})} + (\hat{\mathbf{r}}_{n,t-1}^{(N_{t-1})} - \bar{\mathbf{r}}_{m,t-1}^{(N_{t-1})})(\bullet)^T] \end{aligned} \right\} \quad (7)$$

where N_{t-1} is the number of moving LiDAR measurements obtained at time $t - 1$. The mixing weight α_{mn} ($m, n = 1, 2, 3$) is given by $\alpha_{mn} = T_{mn} \hat{\mu}_{m,t-1}^{(N_{t-1})} / \sum_{m=1}^3 T_{mn} \hat{\mu}_{m,t-1}^{(N_{t-1})}$. T_{mn} is the transition probability from mode m to n . In this study, the transition probabilities are set to $T_{mn} = 0.9$ for $m = n$ and $T_{mn} = 0.05$ for $m \neq n$.

Using the EKF, the prediction and its related error covariance for mode m at time t are computed as follows:

$$\left. \begin{aligned} \hat{\mathbf{r}}_{m,t}^{(0)} &= \mathbf{f}_m(\bar{\mathbf{r}}_{m,t-1}^{(N_{t-1})}) \\ \mathbf{C}_{m,t}^{r(0)} &= \nabla \mathbf{F}_{m,t-1} \bar{\mathbf{C}}_{m,t-1}^{r(N_{t-1})} \nabla \mathbf{F}_{m,t-1}^T + \nabla \mathbf{G}_{m,t-1} \mathbf{C}_w^r \nabla \mathbf{G}_{m,t-1}^T \end{aligned} \right\} \quad (8)$$

where $\nabla \mathbf{F}$ and $\nabla \mathbf{G}$ are the Jacobians of the motion model $\mathbf{f}_m$ pertaining to the state $\bar{\mathbf{x}}_{m,t-1}$ and noise $\Delta \mathbf{v}_{m,t-1}$, respectively.

When the i -th moving LiDAR measurement ($i = 1, 2, \dots, N_t$) is obtained at time t , the motion estimate $\hat{\mathbf{r}}_{m,t}^{(i)}$, its related error covariance $\mathbf{C}_{m,t}^{r(i)}$, and the likelihood $\lambda_{m,t}^{(i)}$ of mode m are updated as follows:

$$\left. \begin{aligned} \hat{\mathbf{r}}_{m,t}^{(i)} &= \hat{\mathbf{r}}_{m,t}^{(i-1)} + \mathbf{C}_{m,t}^{ry(i)} (\mathbf{C}_{m,t}^{y(i)})^{-1} \tilde{\mathbf{y}}_{m,t}^{(i)} \\ \mathbf{C}_{m,t}^{r(i)} &= \mathbf{C}_{m,t}^{r(i-1)} - \mathbf{C}_{m,t}^{ry(i)} (\mathbf{C}_{m,t}^{y(i)})^{-1} (\mathbf{C}_{m,t}^{ry(i)})^T \end{aligned} \right\} \quad (9)$$

$$\lambda_{m,t}^{(i)} = \frac{1}{2\pi \sqrt{|\mathbf{C}_{m,t}^{y(i)}|}} \exp\left[-\frac{1}{2} (\tilde{\mathbf{y}}_{m,t}^{(i)})^T (\mathbf{C}_{m,t}^{y(i)})^{-1} \tilde{\mathbf{y}}_{m,t}^{(i)}\right] \quad (10)$$

where $\tilde{\mathbf{y}}_{m,t}^{(i)}$ and $\mathbf{C}_{m,t}^{y(i)}$ denote the error in the predicted measurement and its related covariance, respectively:

$$\left. \begin{aligned} \tilde{\mathbf{y}}_{m,t}^{(i)} &= \mathbf{y}_t^{(i)} - \mathbf{H}_m \hat{\mathbf{r}}_{m,t}^{(i-1)} \\ \mathbf{C}_{m,t}^{y(i)} &= \mathbf{H}_m \mathbf{C}_{m,t}^{r(i-1)} \mathbf{H}_m^T + \hat{\mathbf{S}}_t^{(i-1)} \mathbf{C}^h (\hat{\mathbf{S}}_t^{(i-1)})^T \\ &\quad + \text{trace}\{\mathbf{C}_t^{p(i-1)} (\hat{\mathbf{J}}_{bt}^{(i-1)})^T \mathbf{C}^h (\hat{\mathbf{J}}_{at}^{(i-1)})\} + \mathbf{C}_m^{\Delta r} \end{aligned} \right\} \quad (11)$$

where $\hat{\mathbf{J}}_{at}^{(i-1)}$ and $\hat{\mathbf{J}}_{bt}^{(i-1)}$ ($a, b = 1, 2$) are the Jacobians of S_{1t} and S_{2t} , respectively.

Equations (9) and (11) follow the derivations in [16]. In terms of likelihood $\lambda_{m,t}^{(i)}$, the mode probability for mode n is updated as

$$\hat{\mu}_{n,t}^{(N_t)} = \frac{\bar{\mu}_{n,t}^{(N_{t-1})} \Lambda_{n,t}}{\sum_{m=1}^3 \bar{\mu}_{m,t}^{(N_{t-1})} \Lambda_{m,t}} \quad (12)$$

where $\Lambda_{m,t} = \prod_{i=1}^{N_t} \lambda_{m,t}^{(i)}$ and $\bar{\mu}_{n,t}^{(N_{t-1})} = \sum_{m=1}^3 T_{mn} \hat{\mu}_{m,t-1}^{(N_{t-1})}$

Finally, the overall motion estimate and its related error covariance are obtained as

$$\left. \begin{aligned} \hat{\mathbf{r}}_t^{(N_t)} &= \sum_{m=1}^3 \hat{\mu}_{m,t}^{(N_t)} \hat{\mathbf{r}}_{m,t}^{(N_t)} \\ \mathbf{C}_t^{r(N_t)} &= \sum_{m=1}^3 \hat{\mu}_{m,t}^{(N_t)} [\mathbf{C}_{m,t}^{r(N_t)} + (\hat{\mathbf{r}}_{m,t}^{(N_t)} - \hat{\mathbf{r}}_t^{(N_t)})(\bullet)^T] \end{aligned} \right\} \quad (13)$$

Tracking in multiobject environments requires data association (one-to-one matching between tracked objects and clustered moving LiDAR measurements). To this end, the object shape (assumed rectangular) is extracted from the moving LiDAR measurements using the rotating-calipers method [27]. The Intersection over Union (IoU) between the extracted rectangle and the rectangle estimated by the IMM-MEM-EKF method is then calculated. Based on the IoU, data association is performed using a simple online real-time tracking algorithm [28].

C. Shape Estimation of Object

The predicted shape of the object and its related error covariance at time t are given by

$$\left. \begin{aligned} \hat{\mathbf{p}}_t^{(0)} &= \hat{\mathbf{p}}_{t-1}^{(N_{t-1})} + \mathbf{A} \hat{\mathbf{r}}_{t-1}^{(N_{t-1})} \\ \mathbf{C}_t^{p(0)} &= \mathbf{C}_{t-1}^{p(N_{t-1})} + \mathbf{C}^{\Delta p} + \mathbf{A} \mathbf{C}_{t-1}^{r(N_{t-1})} \mathbf{A}^T \end{aligned} \right\} \quad (14)$$

where the matrix $\mathbf{A}$ is used for extracting the rotational velocity $\hat{\omega}_{t-1}^{(N_{t-1})}$.

Equation (5) assumes that the object maintains a constant heading angle. The final term in (14) incorporates the correlation between the heading angle and rotational velocity.

For the i -th moving LiDAR measurement ($i = 1, 2, \dots, N_t$) obtained at time t , the shape estimate and its related error covariance are computed as follows:

$$\left. \begin{aligned} \hat{\mathbf{p}}_t^{(i)} &= \hat{\mathbf{p}}_t^{(i-1)} + \mathbf{C}_t^{pY(i)} (\mathbf{C}_t^{Y(i)})^{-1} (\mathbf{Y}_t^{(i)} - \bar{\mathbf{Y}}_t^{(i)}) \\ \mathbf{C}_t^{p(i)} &= \mathbf{C}_t^{p(i-1)} - \mathbf{C}_t^{pY(i)} (\mathbf{C}_t^{Y(i)})^{-1} (\mathbf{C}_t^{pY(i)})^T \end{aligned} \right\} \quad (15)$$

where $\mathbf{Y}_t^{(i)}$ is the pseudo-measurement of $\mathbf{y}_t^{(i)}$. $\bar{\mathbf{Y}}_t^{(i)}$ and $\mathbf{C}_t^{Y(i)}$ are the predicted pseudo-measurement and its related error covariance, respectively, and $\mathbf{C}_t^{pY(i)}$ represents the correlation

between the pseudo-measurement and the motion estimate. These quantities are defined in [16].

V. CLASSIFICATION OF LiDAR MEASUREMENTS

A. Dataset

The performance of the proposed SLAMMOT approach is evaluated using the SemanticKITTI and KITTI tracking datasets [19][20]. The SemanticKITTI dataset contains sequences in which the ego vehicle travels along routes with many loops; however, the ground-truth positions and sizes of the target objects are not provided. Conversely, the KITTI tracking dataset presents ground-truth annotations for object positions and sizes but does not include loop-rich routes suitable for SLAM evaluation. For this reason, the SLAM and MOT are evaluated on the SemanticKITTI and KITTI tracking datasets, respectively.

In both datasets, the ego vehicle is equipped with a 64-layer LiDAR (Velodyne HDL-64E). The horizontal and vertical fields of view of the LiDAR are 360° and 26.8° , respectively. During each LiDAR sampling period (0.1 s), the 64 laser beams complete one full rotation in the horizontal direction, producing approximately 120,000 measurements.

B. Object Recognition Result

First, the object-recognition performance of the SLAMMOT approach is evaluated. Because the LiDAR-equipped ego vehicles in both datasets travel in similar environments, RangeNet++ is trained on the SemanticKITTI dataset alone. Table I shows the result of object recognition. As shown in the table, large objects—such as cars, roads, buildings, and vegetation—are accurately detected, whereas the classification accuracy is lower for small objects, including people and traffic signs.

For the six semantic voxel classes used in the loop-detection method (buildings, fences, poles, traffic signs, trunks, and vegetation), the average IoU, precision, and recall are computed as 54.8 %, 73.5 %, and 67.0 %, respectively. For the four semantic voxel classes used in MOT (cars, two-wheelers, other vehicles, and people), the average IoU, precision, and recall are 50.6 %, 70.8 %, and 60.1 %, respectively.

C. SLAM Result

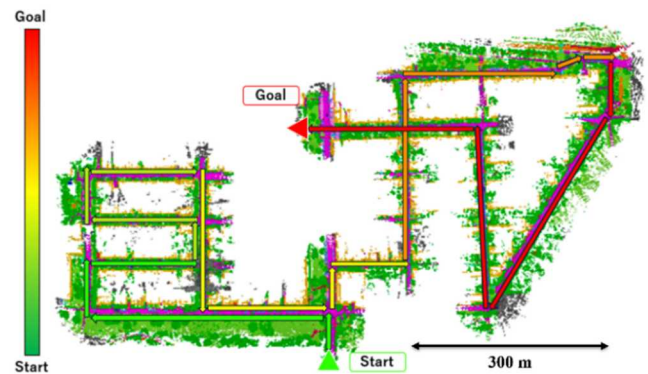
Among the 11 sequences in the SemanticKITTI dataset, sequences 00, 02, 05, and 08, which contain many loops, are selected for evaluating the SLAM performance. The SLAM evaluation is conducted under the following cases.

- Case 1 (proposed method): SLAM using only the stationary LiDAR measurements extracted by semantic segmentation. The moving and potentially moving LiDAR measurements are removed.
- Case 2 (our previous method [13]): SLAM using the stationary LiDAR measurements extracted by the occupancy-grid method. The moving LiDAR measurements are removed, but the potentially moving LiDAR measurements are retained as stationary.
- Case 3: SLAM using both the stationary and moving LiDAR measurements without removal.

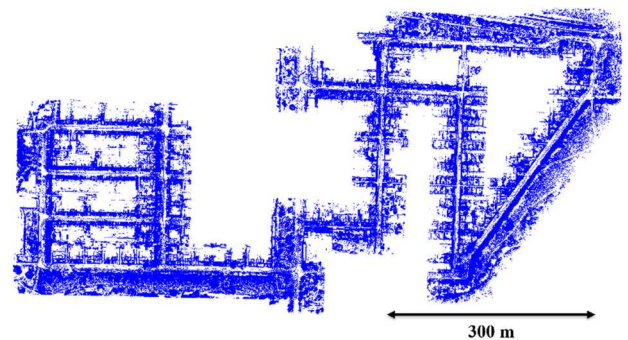
Figures 4 (a) and (b) show the true environmental map and corresponding mapping result, respectively, for sequence 08 in case 1. Figure 5 shows the SLAM-estimated trajectory of the ego vehicle in cases 1, 2, and 3. The ego-vehicle pose is

TABLE I. PERFORMANCE OF OBJECT CLASSIFICATION

Class	IoU [%]	Precision [%]	Recall [%]
Car	89.2	91.4	97.4
Two-wheeler	45.4	66.2	59.2
Other vehicles	46.6	70.5	57.8
Person	21.4	54.9	26.0
Road	93.1	97.0	96.9
Lawn	68.7	88.3	75.5
Building	80.9	88.4	90.5
Fence	46.7	67.5	60.2
Vegetation	81.4	86.8	93.0
Trunk	47.0	67.9	60.4
Pole	43.9	57.1	65.6
Traffic sign	28.9	73.3	32.2



(a) True map.



(b) Built map (case 1).

Figure 4. Environment mapping result for sequence 08 (top view).

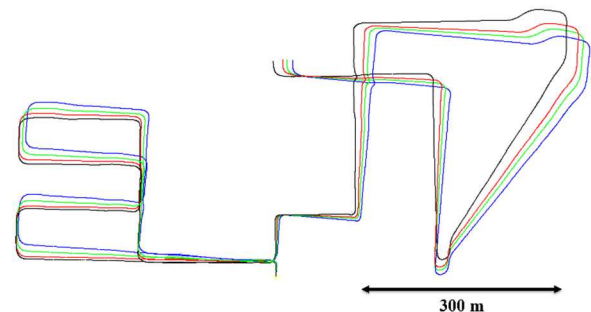


Figure 5. Estimated trajectory of the ego vehicle in sequence 08 (top view). The black line indicates the ground truth. The red, blue, and green lines indicate the results for cases 1, 2, and 3, respectively.

accurately estimated in case 1, whereas case 3 exhibits a large localization error.

The SLAM performance is quantitatively evaluated using Absolute Trajectory Error (ATE), defined as the error in the ego vehicle's trajectory. Table II shows the result. On all four sequences, case 1 is the lowest ATE among the three cases.

D. MOT Result

The MOT performance is evaluated on the KITTI tracking dataset. As an example, Figure 6 shows the MOT result of sequence 02. The black dotted line indicates the trajectory of the ego vehicle. The rectangles indicate the estimated bounding boxes of objects, and the lines indicate the heading direction of moving objects. The different colored dots indicate the moving LiDAR measurements extracted by semantic segmentation. The MOT performance is quantitatively evaluated using IoU metric, which measures the overlap between the true and estimated bounding boxes of moving objects. The KITTI tracking dataset provides ground-truth bounding boxes of objects visible to the camera mounted on the ego vehicle. For comparison, moving objects are also tracked using the conventional single model-based MEM-EKF.

Table III shows the results for five sequences. The proposed multimodel-based MEM-EKF outperforms the single model-based MEM-EKF for sequences 02, 03, and 06, which represent intersection environments where car motion frequently changes. For sequence 05, which represents a straight-road environment, both frameworks exhibit low MOT accuracy because most cars were located ahead of the ego vehicle, and the LiDAR measurements were acquired from the rear ends of the cars. Both frameworks achieve similar MOT accuracies for sequence 08, which represents a highway environment where most cars travel

TABLE II. ATE IN SLAM

Sequence	Case 1	Case 2	Case 3
00	10.36 m	10.74 m	19.02 m
02	55.46 m	61.42 m	108.99 m
05	8.25 m	8.69 m	10.54 m
08	12.04 m	15.10 m	27.44 m

The smallest ATEs are indicated in bold.

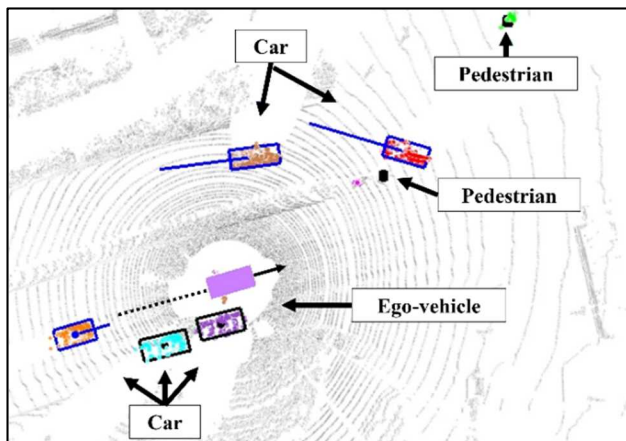


Figure 6. MOT result for sequence 02 of the KITTI tracking dataset (top view).

TABLE III. IoU PERFORMANCE OF MOT

Sequence	Multimodel-based EOT	Single model-based EOT
02	71.3 %	67.6 %
03	80.7 %	75.1 %
05	45.0 %	51.3 %
06	73.1 %	70.7 %
08	62.6 %	64.8 %

The highest IoUs are indicated in bold.

at nearly constant velocity. Overall, multimodel-based MEM-EKF achieves higher MOT accuracy than the single model-based MEM-EKF in scenes with frequent transitions among motion, such as those induced by intersections.

VI. CONCLUSION AND FUTURE WORK

This paper presented a LiDAR SLAMMOT approach based on the semantic segmentation and IMM-EOT methods. After discriminating stationary and moving objects using RangeNet++ based on onboard LiDAR measurements, the stationary LiDAR measurements were input to NDT-Graph SLAM, while the moving and potentially moving LiDAR measurements were used for MOT using the IMM-MEM-EKF. The performance of the proposed method was evaluated on the SemanticKITTI and KITTI tracking datasets.

Further evaluation and improvement of the proposed method are required, including MOT for small target objects, such as pedestrians and two-wheelers, as well as the development of a tightly coupled SLAM and MOT system. In this paper, SLAMMOT was executed offline on a laptop computer. As future work, we aim to reduce the computational time by optimizing the program code and employing embedded hardware for real-time operation.

ACKNOWLEDGMENT

This study was partially supported by the KAKENHI Grant #23K03781, the Japan Society for the Promotion of Science (JSPS).

REFERENCES

- [1] J. Zhao et al., "Autonomous Driving System: A Comprehensive Survey," *Expert Systems with Applications* 242, pp. 1–27, 2023.
- [2] Y. Fu, C. Li, F. R. Yu, T. H. Luan, and Y. Zhang, "A Survey of Driving Safety with Sensing, Vehicular Communications, and Artificial Intelligence-based Collision Avoidance," *IEEE Trans. Intelligent Transportation Systems*, vol. 23, pp. 6142–6163, 2022.
- [3] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous Localization and Mapping: A Survey of Current Trends in Autonomous Driving," *IEEE Trans. Intelligent Vehicles*, vol. 2, pp. 194–220, 2017.
- [4] E. Arnold et al., "A Survey on 3D Object Detection Methods for Autonomous Driving Applications," *IEEE Trans. Intelligent Transportation Systems*, vol. 20, pp. 3782–3795, 2019.
- [5] Y. Zhang, P. Shi, and J. Li, "3D LiDAR SLAM: A Survey," *The Photogrammetric Record*, vol. 39, pp. 457–517, 2024.
- [6] S. Steyer, G. Tanzmeister, and D. Wollherr, "Object Tracking Based on Evidential Dynamic Occupancy Grids in Urban Environments," *Proc. 2017 IEEE Intelligent Vehicles Symposiums*, pp. 1064–1070, 2017.
- [7] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, "A Tutorial on Graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, pp. 31–43, 2010.

- [8] R. B. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3D Registration," *Proc. IEEE Int. Conf. Robotics and Automation*, pp. 3212–3217, 2009.
- [9] F. Martín, R. Triebel, L. Moreno, and R. Siegwart, "Two Different Tools for Three-Dimensional Mapping: DE-based Scan Matching and Feature-based Loop Detection," *Robotica*, vol. 32, pp. 19–41, 2017.
- [10] L. Wang and Y. Huang, "A Survey of 3D Point Cloud and Deep Learning-based Approaches for Scene Understanding in Autonomous Driving," *IEEE Intelligent Transportation Systems Magazine*, vol. 14, pp. 135–154, 2022.
- [11] Y. Wang, Z. Wang, and X. Wu, "An Efficient and Accurate 3D SLAM Method for Dynamic Environment," *Proc. 4th Int. Conf. Robotics and Computer Vision*, pp. 148–153, 2022.
- [12] Z. Ying and H. Li, "IMM-SLAMMOT: Tightly-Coupled SLAM and IMM-based Multi-Object Tracking," *IEEE Trans. Intelligent Vehicles*, vol. 9, pp. 3964–3974, 2024.
- [13] S. Muramatsu, M. Hashimoto, and K. Takahashi, "Performance Improvement of Loop Closure Detection Using Semantic Segmentation in LiDAR SLAM," *Proc. the 10th Int. Conf. Advances in Sensors, Actuators, Metering and Sensing*, pp. 45–51, 2025.
- [14] Y. Tamura, R. Murabayashi, M. Hashimoto, and K. Takahashi, "Hierarchical Cooperative Tracking of Vehicles and People Using Laser Scanners Mounted on Multiple Mobile Robots," *Int. J. Advances in Intelligent Systems*, vol. 10, pp. 90–101, 2017.
- [15] K. Granstrom and M. Baum, "A Tutorial on Multiple Extended Object Tracking," *TechRxiv*, pp. 1–36, 2022.
- [16] S. Yang and M. Baum, "Tracking the Orientation and Axes Lengths of an Elliptical Extended Object," *IEEE Trans. Signal Process*, vol. 67, pp. 4720–4729, 2019.
- [17] S. Wang, C. Men, R. Li, and T. Yeo, "A Maneuvering Extended Target Tracking IMM Algorithm Based on Second-Order EKF," *IEEE Trans. Instrumentation and Measurements*, vol. 73, pp. 1–11, 2024.
- [18] Y. Nakatani, S. Uchiyama, M. Hashimoto, and K. Takahashi, "LiDAR-based Cooperative Tracking Using IMM Extended Object Tracking and Information Filter," *Proc. IEEE Region 10 Conf.*, pp. 51–55, 2025.
- [19] J. Behley et al., "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 9297–9307, 2019.
- [20] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.
- [21] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and Accurate LiDAR Semantic Segmentation," *Proc. 2019 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 4213–4220, 2019.
- [22] A. Jhaldiyal and N. Chaudhary, "Semantic Segmentation of 3D LiDAR Data Using Deep Learning: A Review of Projection-Based Methods," *Applied Intelligence*, vol. 53, pp. 6844–6855, 2023.
- [23] H. Zhao, L. Jiang, J. Jia, P. H. K. Torr, and V. Koltun, "Point Transformer," *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 16259–16268, 2021.
- [24] R. Nakamura, I. Inaga, M. Hashimoto, and K. Takahashi, "SLAM-based Mapping Using Micromobility-Mounted Solid-State LiDAR," *Proc. the 11th IIAE Int. Conf. Intelligent Systems and Image Processing 2024*, pp. 93–100, 2024.
- [25] R. Nakamura, T. Kambe, M. Hashimoto, and K. Takahashi, "SLAM-based Mapping in Truck-and-Robot System for Last-Mile Delivery Automation," *Proc. the 2023 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, pp. 31–37, 2023.
- [26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. the Second Int. Conf. Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [27] G. Toussaint, "Solving Geometric Problems with the Rotating Calipers," *Proc. IEEE Mediterranean Electrotechnical Conf.'83*, pp. 1–8, 1983.
- [28] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," *Proc. IEEE Int. Conf. Image Process*, pp. 3464–3468, 2016.

Majority Voting-Based Update of Point-Cloud Map Using LiDAR Information from Multiple Vehicles

Saki Fujisawa

Graduate School of Science and Engineering
Doshisha University
Kyotanabe, Kyoto, Japan
e-mail: ctwn0104@mail4.doshisha.ac.jp

Masafumi Hashimoto, Kazuhiko Takahashi

Faculty of Science and Engineering
Doshisha University
Kyotanabe, Kyoto, Japan
e-mail: {mhashimo, katakaha}@mail.doshisha.ac.jp

Abstract—This paper presents a crowd-sensing-based method of updating three-dimensional (3D) point-cloud maps in road environments. This method utilizes data from Light Detection and Ranging (LiDAR) mounted on multiple vehicles. Each vehicle maps its 3D LiDAR point-cloud data onto a two-dimensional (2D) local grid map. The state of each cell in the local grid map—unchanged (occupied or empty), changed (occupied or empty), or undetermined—is classified through majority voting after comparing the LiDAR point-cloud data with the original reference point-cloud map. The 2D local grid map containing the cell states is then sent to a central server, which collects the 2D local grid maps from all vehicles into a 2D global grid map and determines the final cell states through another majority-voting process. Finally, the 3D point-cloud map is updated based on the cell states in the 2D global grid map. This two-stage majority voting method enables simple yet robust map updates. Simulation experiments demonstrate the effectiveness of the map-update method.

Keywords—LiDAR; map update; point-cloud map; crowd-sensing; majority voting.

I. INTRODUCTION

Vehicles in Intelligent Transportation Systems (ITS) are increasingly equipped with autonomous driving capabilities and advanced driving-assistance systems [1]. In addition, last-mile delivery automation using delivery robots and drones is being actively pursued to support advanced logistics systems [2]. A key technology underlying these systems is environmental mapping, which assists self-localization and environment perception by vehicles and mobile robots [3].

Environmental mapping in ITS typically relies on Mobile Mapping Systems (MMSs) [4], which integrate high-precision sensors, such as Light Detection And Ranging (LiDAR), cameras, inertial navigation systems, and global navigation satellite systems. However, because MMSs are limited in number and cannot cover all road environments nationwide, most maps are built only for major areas, such as expressways, freeways, and primary urban roads. To compensate for the limitations of MMS-based mapping, a mapping method based on Simultaneous Localization and Mapping (SLAM) using cameras and LiDARs mounted on ordinary vehicles has been widely studied [5][6].

Typical SLAM technologies include Normal Distributions Transform (NDT) SLAM [7] and iterative closest point SLAM [8]. Previously, we proposed a method that builds three-dimensional (3D) environmental point-cloud maps based on

LiDAR NDT SLAM in dynamic road environments, along with a method that merges multiple environmental maps [9].

Accurate environmental maps are essential for vehicle self-localization and environment perception. When the environment changes—for example, due to new construction or the removal of existing buildings—the accuracy of self-localization and environment perception will deteriorate. Updating the environmental map in response to changes in environmental features, such as buildings and traffic signs, is an important challenge [10].

This paper presents a method that updates 3D environmental point-cloud maps built using LiDAR NDT SLAM. In most conventional map-update methods, the point cloud in the environmental map is matched to the current LiDAR point-cloud data, and points with large matching errors are identified as change points for map update [11][12]. However, because 3D mechanical LiDAR commonly used in automotive applications produces vertically sparse point-cloud data, it is difficult to detect changes in distant environmental features. In addition, in dynamic road environments, environmental features are often occluded by other vehicles or structures, making reliable map updates difficult when relying solely on LiDAR data from a single vehicle.

In real road environments, many vehicles travel in different lanes and at different times. To overcome the limitations of single-vehicle LiDAR-based map updates, researchers have proposed crowd-sensing-based methods that aggregate the sensor information from multiple vehicles (crowd-sourcing vehicles) on a central server to update the map [13][14][15]. For example, Kim et al. [15] updated a LiDAR point-cloud environmental map by mapping vehicle-mounted LiDAR point-cloud data onto a voxel map (3D grid map). The current LiDAR point-cloud data were compared with the environmental map and changes were detected based on evidence theory. The map changes collected from the crowd-sourcing vehicles were reproduced on a map-cloud server based on similarity-based clustering, and the environmental maps were updated. Although this method represents a theoretical advancement, its effectiveness seems difficult to demonstrate using vertically sparse point-cloud data from mechanical 3D LiDAR.

Inspired by the work of Kim et al. [15], this paper presents a crowd-sensing-based method for updating 3D environmental point-cloud maps. To address the vertical sparsity of mechanical LiDAR, the 3D LiDAR point-cloud data from each vehicle are mapped onto a two-dimensional (2D) grid map. By comparing

the current LiDAR point-cloud data with the environmental map, changes in the grid-cell states are detected through majority voting. The 2D grid map containing the cell states is sent to a central server, which collects and fuses the 2D grid maps from multiple vehicles and determines the final cell states through another majority-voting process. The 3D environmental point-cloud map is then updated based on the cell states. Compared with the method of Kim et al. [15], the proposed method employs a simpler voting-based framework using 2D grid maps instead of evidence-based theory using 3D voxel maps, thereby reducing computational complexity.

The remainder of this paper is structured as follows. Section II overviews the proposed map-update method. Sections III and IV describe the map-update processes on the vehicle side and the central server side, respectively. Section V presents simulation experiments for a three-vehicle system, and Section VI presents the main conclusions of this study.

II. OVERVIEW OF ENVIRONMENTAL MAP UPDATE

The 3D mechanical LiDAR used in this study is a 32-layer mechanical LiDAR (Velodyne HDL-32E) with horizontal and vertical Fields-Of-VIEWS (FOVs) of 360° and 41.34° , respectively, and angular resolutions of 0.16° and 1.33° , respectively. During each 0.1 s sampling period (or scan), the laser beams complete one full rotation in the horizontal direction. LiDAR acquires 70,000 point-cloud points per scan.

Our method updates the NDT-based environmental map. The 3D LiDAR point-cloud data are stored on a 3D grid map composed of cubic voxels with a side length of 0.3 m. For map-update process, the 3D LiDAR point-cloud data captured in each scan (referred to as the new LiDAR point-cloud data) and the previously built 3D point-cloud map (referred to as the environmental map) are projected onto a 2D grid map on a horizontal plane. Each cell in the 2D grid map is a square with a side length of 0.3 m.

Figure 1 illustrates the map-update flow. The environmental map is updated by comparing it with the new LiDAR point-cloud data obtained on each vehicle. The new LiDAR point-cloud data contain the data from stationary and moving objects, such as cars, two-wheelers, and pedestrians. Because the map-update process requires only stationary-object information, LiDAR data corresponding to moving objects are removed using the occupancy grid-based method [9].

The NDT map-matching method aligns the new LiDAR point-cloud data with the environmental map and estimates the vehicle's self-pose. Based on this estimated pose, both the environmental map and the new LiDAR point-cloud data are mapped onto a 2D local grid map. By comparing the new LiDAR point-cloud data with the environmental map in the 2D local grid map through majority voting, each cell is classified into one of five states: *unchanged (occupied or empty)*, *changed (occupied or empty)*, or *undetermined*. The 2D local grid map, embedded with these cell states, is sent to a central server, which collects the 2D local grid maps from multiple vehicles and determines the final cell states through majority voting. Finally, the 3D environmental point-cloud map is updated based on the cell states.

In this study, the 3D LiDAR point clouds are mapped onto a 2D grid map to mitigate the vertical sparsity inherent in mechanical 3D LiDAR, simplify the map-update process, and

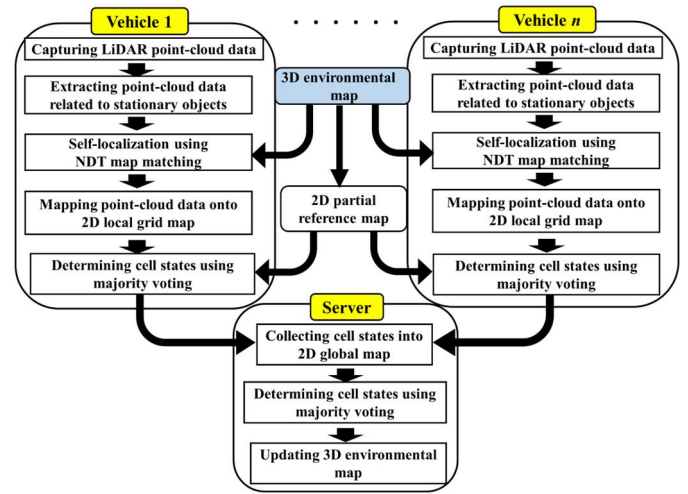


Figure 1. Overview of the map-update procedure.

reduce computational cost. However, because height information is discarded, the map-update accuracy may decrease for environmental features with significantly different vertical structures.

III. CELL-STATE DETERMINATION OF 2D LOCAL GRID MAP ON VEHICLE

A. Classification of Cell State

As shown in Figure 2 (a), a square area with a side length of 80 m, centered at the self-position (V_i) of the i -th vehicle ($i = 1, 2, \dots$), is extracted from the environmental map in the world coordinate frame. The extracted area is referred to as the partial reference map R_i . The new LiDAR point-cloud data obtained by the i -th vehicle at V_i is denoted as S_i .

As illustrated in Figure 2 (b), both the partial reference map R_i and the new LiDAR point-cloud data S_i are projected onto the same 2D local grid map, which consists of square cells with a side length of 0.3 m. The presence or absence of point-cloud data in R_i and S_i is compared on a cell-by-cell basis to detect environmental changes. Each cell is then assigned to one of five classes, as defined in Figure 3.

- Unchanged (occupied): The cell contains both partial reference-map data and new LiDAR point-cloud data
- Unchanged (empty): The cell contains neither partial reference-map data nor new LiDAR point-cloud data.
- Changed (occupied): The cell contains new LiDAR point-cloud data but no partial reference-map data.
- Changed (empty): The cell contains partial reference-map data but no new LiDAR point-cloud data.
- Undetermined: The new LiDAR point-cloud data in the cell are occluded and therefore unavailable.

Because of the limited vertical FOV of LiDAR, upper positions of tall environmental features may not be observed. In such cases, the states of those regions cannot be reliably determined. Therefore, cells in which the height of objects in

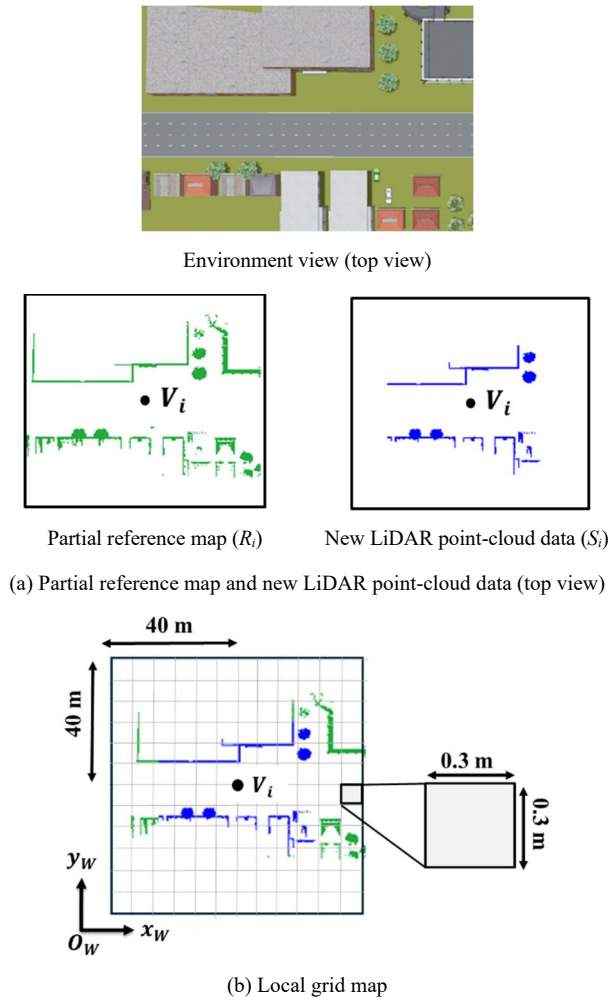


Figure 2. Mapping of a partial reference map and new LiDAR point-cloud data onto a 2D local grid map (top view).

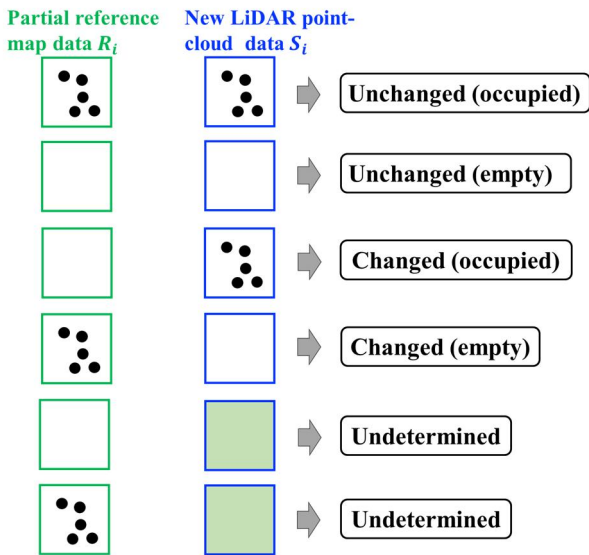


Figure 3. Classification of cell states. Cells with black circles (LiDAR measurement points) indicate occupied cells; the white and green cells indicate empty and occluded cells, respectively.

the environmental map exceeds that of the new LiDAR point-cloud data are treated as invalid cells and remain unclassified.

B. Determination of Cell State

The state of each cell in the 2D local grid map is classified at every LiDAR scan and determined through majority voting. For example, if the LiDAR point-cloud data are mapped onto a cell during ten scans and the cell state is classified as *unchanged (empty)* in four scans and *changed (empty)* in six scans, the final state is voted as *changed (empty)*. If the cell state is classified as *unchanged (empty)* in five scans and *changed (empty)* in five scans, the final state is voted as *undetermined*.

When many measurement points of new LiDAR point-cloud data are mapped onto one cell, that cell is more likely to contain an object than cells with fewer measurement points. Therefore, instead of assigning one vote per LiDAR scan for the five cell classes, occupied cells are assigned multiple votes per LiDAR scan. The number of votes per scan is determined by the following exponential function (see Figure 4):

$$w(n) = a + (b - a)(1 - e^{-k(n - n_{\min})}) \quad (1)$$

where n denotes the number of measurement points of the new LiDAR point-cloud data in an occupied cell, and $n_{\min}$ (set to 5 in this study) is the minimum number of measurement points required to classify a cell as occupied. Below this threshold, a cell cannot be classified as occupied. Parameters a and b represent the lower and upper bounds of the vote weight, respectively, while k controls the rate of increase. In the experiments (Section V), we set $a = 2$, $b = 3$, and $k = 0.5$ based on preliminary results.

In summary, majority voting assigns one vote per LiDAR scan for *unchanged (empty)*, *changed (empty)*, and *undetermined* cells, and multiple votes per LiDAR scan—computed by (1)—for *unchanged (occupied)* and *changed (occupied)* cells.

Table I presents exemplary cell states determined by the three vehicles. For vehicle 1, the cell is classified as *changed (occupied)* in 25 scans and *undetermined* in 20 scans. According to (1), the cell is voted as *changed (occupied)* in 33 votes and *undetermined* in 20 votes. Thus, the final state is *changed (occupied)*. Similarly, the final states for vehicles 2 and 3 are *changed (occupied)* and *undetermined*, respectively.

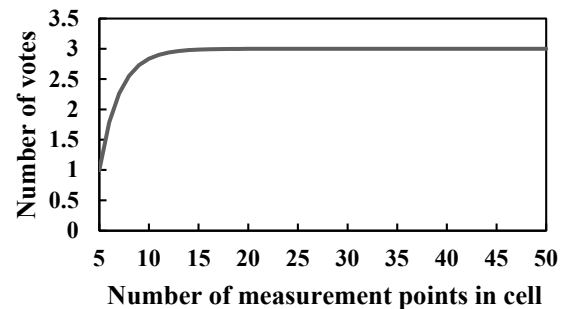


Figure 4. Number of votes versus number of measurement points in an occupied cell

TABLE I. EXAMPLE OF NUMBER OF CELL STATES AND DECISION BY THREE VEHICLES

	Unchanged (occupied)	Unchanged (empty)	Changed (occupied)	Changed (empty)	Undetermined	Decision
Vehicle 1	0	0	25	0	20	Changed (occupied)
Vehicle 2	0	0	25	0	20	Changed (occupied)
Vehicle 3	5	0	18	0	22	Undetermined

IV. CELL-STATE DETERMINATION OF 2D GLOBAL GRID MAP ON CENTRAL SERVER

Each vehicle generates a 2D local grid map containing its estimated cell states. The local grid maps from multiple vehicles are then collected into a 2D global grid map on the central server. Each cell on the global grid map is a $0.3 \text{ m} \times 0.3 \text{ m}$ square.

The state of each cell on the 2D global grid map is determined through majority voting based on the “one vote per vehicle” principle. For example, in Table I, two vehicles classify a cell as *changed (occupied)*, and one vehicle classifies it as *undetermined*. Consequently, the central server determines the final state as *changed (occupied)*.

For cells classified as *unchanged (occupied)* or *undetermined*, the point-cloud data are retained in the reference map. For cells classified as *changed (occupied)*, the reference map is updated using the new LiDAR point-cloud data. For cells classified as *changed (empty)*, the LiDAR data occupying the cell are deleted from the reference map. In *undetermined* cells where multiple classes receive equal votes, the point-cloud data are retained in the reference map.

This paper introduces a two-stage majority voting method, in which both the vehicles and the central server perform majority voting. Each vehicle decides its cell states in the 2D local grid map and passes them to the central server, which then examines the cell states reported by all vehicles to finalize the decision. A one-stage majority voting scheme performed solely by the central server is also conceivable. In this scheme, each vehicle would send the vote counts for five classes to the central server. The central server would determine the final cell states by aggregating these votes.

When vehicles travel in different lanes within the same environment, each vehicle decides different cell states in the 2D local grid map. The two-stage majority voting method reduces bias in the number of cell states determined by each vehicle. Furthermore, it reduces communication costs between vehicles and the central server compared to the one-stage majority voting method. For these reasons, this paper applies the two-stage majority voting method. The following section compares the performances of the one-stage and two-stage majority voting methods.

V. SIMULATION EXPERIMENTS

Simcenter Prescan (Siemens) [15] is used as the simulator to generate vehicle motion and the associated LiDAR point-cloud data in a dynamic road environment. The simulation environment includes many moving objects, such as cars and pedestrians. Figure 5 is an example of the LiDAR point-cloud data generated by the simulator.

Figure 6 (a) illustrates an urban road environment with two lanes in each side (four lanes in total), and Figure 6 (b) shows the built reference map. Three vehicles travel in different lanes

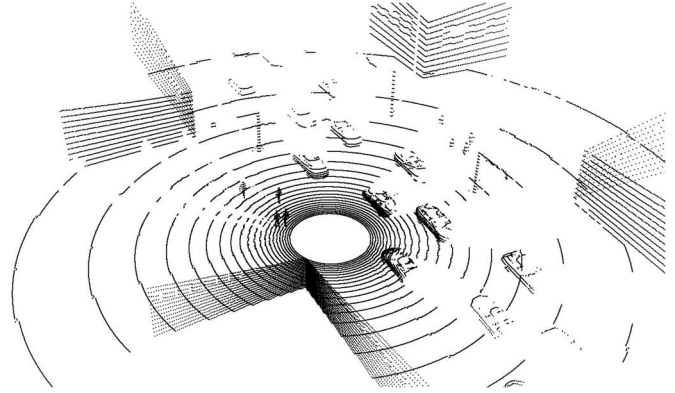


Figure 5. Example of LiDAR point-cloud data in an intersection environment (bird's-eye view).

and directions at different times; specifically, vehicles 1 and 2 travel from west to east at 40 km/h, and vehicle 3 travels from east to west at 40 km/h. Four areas have changed since the reference map was built. The changes include the construction of one building in area 1, the demolition of two buildings in area 2, the construction of two buildings in area 3, and the demolition of six buildings in area 4.

First, the effectiveness of the proposed map-update method is qualitatively demonstrated in areas 1 and 2 in Figure 6 (a). Figures 7 (a) and (b) show views of area 1 when the environmental map was built and at the time the three vehicles passed, respectively. Building 1, which does not exist in Figure 7 (a), appears in Figure 7 (b). The LiDAR point-cloud data of building 1, absent from the environmental map (Figure 7 (c)), are successfully added to the updated map (red dots in Figure 7 (d)).

In area 2, buildings 2 and 3 had been demolished by the time the three vehicles passed the site (Figure 8 (b)). Although both buildings exist in the environmental map (red dots in Figure 8 (c)), their LiDAR point-cloud data are removed in the updated map (Figure 8 (d)).

The map-update accuracy over the entire environment shown in Figure 6 is quantitatively evaluated in the following five cases.

- Case 1: Two-stage majority voting, applying the “multiple votes per LiDAR scan” principle on each vehicle and the “one vote per vehicle” principle on the central server.
- Case 2: Two-stage majority voting, applying the “one vote per LiDAR scan” principle on each vehicle and the “one vote per vehicle” principle on the central server.
- Case 3: One-stage majority voting on the central server under the “multiple votes per LiDAR scan” principle.



(a) Environment

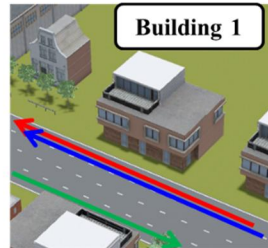


(b) Environmental map

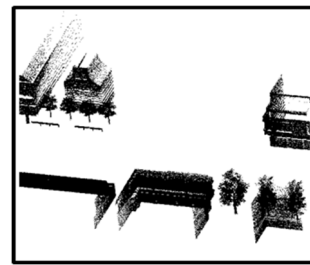
Figure 6. Simulation environment and environmental map (top view).



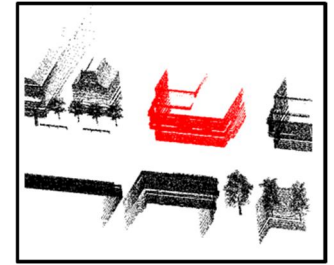
(a) Previous environment



(b) Current environment

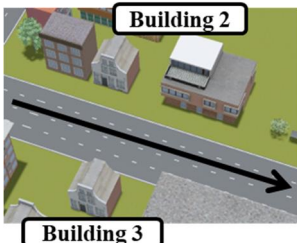


(c) Environmental map

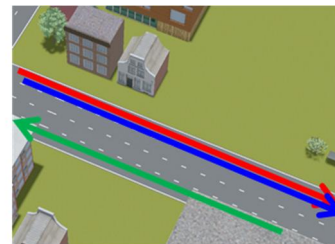


(d) Updated map

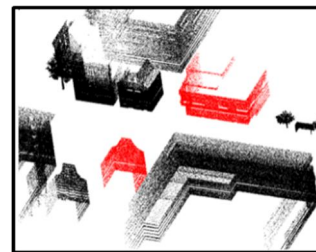
Figure 7. Simulation results of area 1 (bird's-eye view). In (a) and (b), the black, blue, and red lines indicate the movement paths of vehicles.



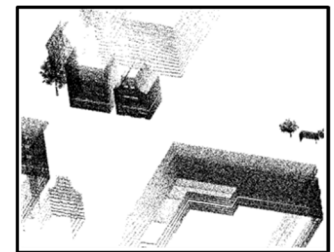
(a) Previous environment



(b) Current environment



(c) Environmental map



(d) Updated map

Figure 8. Simulation results in area 2 (bird's-eye view). In (a) and (b), the black, blue, and red lines indicate the movement paths of vehicles.

• Case 4: One-stage majority voting on the central server under the “one vote per LiDAR scan” principle.

• Case 5 (proposed method): Two-stage majority voting, applying the “multiple votes per LiDAR scan” principle on each vehicle and the “one vote per vehicle” principle on the central server. In addition, vehicles incorporate invalid cells caused by the vertical FOV limitation of the LiDAR (see the last sentence of Section III-A).

Tables II and III summarize the precision and recall results, respectively, of all 514,900 cells in the environment shown in Figure 6. Table IV presents the corresponding F1-score

(harmonic mean of precision and recall). Larger values of all three metrics indicate higher cell-state detection performance.

As shown by the F1-scores in Table IV, the proposed method (case 5) achieves the highest cell-state detection

TABLE II. PRECISION

	Unchanged	Changed (occupied)	Changed (empty)
Case 1	0.996	0.748	0.550
Case 2	0.996	0.787	0.536
Case 3	0.996	0.663	0.556
Case 4	0.995	0.768	0.499
Case 5	0.996	0.748	0.645

TABLE III. RECALL

	Unchanged	Changed (occupied)	Changed (empty)
Case 1	0.994	0.998	0.767
Case 2	0.994	0.954	0.767
Case 3	0.994	0.987	0.767
Case 4	0.993	0.882	0.767
Case 5	0.996	0.998	0.767

TABLE IV. F1-SCORE

Case 1	0.830
Case 2	0.829
Case 3	0.811
Case 4	0.807
Case 5	0.850

performance among all cases. In addition, the two-stage majority voting method (cases 1 and 2) outperforms the one-stage majority voting method (cases 3 and 4). Finally, majority voting based on the “multiple votes per LiDAR scan” principle (cases 1 and 3) yields better performance than majority voting based on the “one vote per LiDAR scan” principle (cases 2 and 4).

VI. CONCLUSIONS

This paper presented a crowd-sensing-based method for updating 3D point-cloud maps of road environments using LiDAR data collected from multiple vehicles. Each vehicle mapped its 3D LiDAR point-cloud data onto a 2D local grid map, and the state of each grid cell was classified using a majority voting method under the “multiple votes per LiDAR scan” principle. The 2D local grid map comprising the cell states was sent to a central server that aggregated the 2D local grid maps from all vehicles and determined the final cell states through majority voting under the “one vote per vehicle” principle. Simulation experiments demonstrated the effectiveness of the map-update method.

In this study, the 3D LiDAR point clouds were mapped onto a 2D grid map to mitigate the vertical sparsity inherent in mechanical LiDAR, simplify the map-update process, and reduce computational cost. However, because height information is discarded, the map-update accuracy may degrade for environmental features with substantial vertical variation. As future work, the impact of height-information loss on map-update accuracy will be further investigated. In addition, real-world experiments using a test vehicle will be conducted to validate the effectiveness and limitations of the proposed method.

ACKNOWLEDGMENT

This study was partially supported by the KAKENHI Grant #23K03781, the Japan Society for the Promotion of Science (JSPS).

REFERENCES

- [1] M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, “Autonomous Driving, Technical, Legal and Social Aspects,” Springer Open, 2016.
- [2] T. Hoffmann and G. Prause, “On the Regulatory Framework for Last-Mile Delivery Robots,” *Machines* 2018, 6, 33, pp. 1–16, 2018.
- [3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, “A Survey of Autonomous Driving: Common Practices and Emerging Technologies,” *IEEE Access*, vol. 8, pp. 58443–58469, 2020.
- [4] M. Elhashash, H. Albanwan, and R. Qin, “A Review of Mobile Mapping Systems: From Sensors to Applications,” *Sensors* 2022, 22, 4262, pp. 1–26, 2022.
- [5] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, and H. Voos, “Visual SLAM: What are the Current Trends and what to Expect?,” *Sensors* 2022, 22, 9297, pp. 1–29, 2022.
- [6] Y. Zhang, P. Shi, and J. Li, “3D LiDAR SLAM: A Survey,” *The Photogrammetric Record*, vol. 39, pp. 457–517, 2024.
- [7] P. Biber and W. Strasser, “The Normal Distributions Transform: A New Approach to Laser Scan Matching,” *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 2743–2748, 2003.
- [8] P. J. Besl and N. D. McKay, “A Method of Registration of 3-D Shapes,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, 1992.
- [9] S. Tanaka, C. Koshiro, M. Yamaji, M. Hashimoto, and K. Takahashi, “Point Cloud Mapping and Merging in GNSS-Denied and Dynamic Environments Using Onboard Scanning LiDAR,” *Int. J. Advances in Systems and Measurements*, vol. 13, pp. 275–288, 2020.
- [10] B. Wijaya, et al., “High Definition Map Mapping and Update: A General Overview and Future Directions,” *arXiv:2409.09726*, pp. 1–30, 2024.
- [11] I. Jinno, Y. Sasaki, and H. Mizoguchi, “3D Map Update in Human Environment Using Change Detection from LIDAR Equipped Mobile Robot,” *Proc. 2019 IEEE/SICE Int. Symp. System Integration*, pp. 330–335, 2019.
- [12] D. Fukutomi, K. Maeda, T. Azumi, S. Kato, and N. Nishio, “The Integration of Two Independent Point Cloud Maps Acquired at Different Times,” *Proc. Embedded System Symp., Information Proc. Society of Japan*, pp. 54–61, 2019 (in Japanese).
- [13] K. Kim, S. Cho, and W. Chung, “HD Map Update for Autonomous Driving with Crowdsourced Data,” *IEEE Robotics and Automation Letters*, vol. 6, pp. 1895–1901, 2021.
- [14] C. Kim, et al., “Updating Point Cloud Layer of High Definition (HD) Map Based on Crowd-Sourcing of Multiple Vehicles Installed LiDAR,” *IEEE Access*, vol. 9, pp. 8028–8046, 2021.
- [15] C. Kim, et al., “Cloud Update of Geodetic Normal Distribution Map Based on Crowd-Sourcing Detection Against Road Environment Changes,” *J. Advanced Transportation*, vol. 2022, pp. 1–26, 2022.
- [16] Simcenter Prescan, <https://resources.sw.siemens.com/en-US/article-introducing-simcenter-prescan360/> [retrieved: Dec, 2025].

Evaluating LLM-based Pseudonymization for Cross-Platform Identifiers

Yeong Su Lee , Hendrik Bothe  and Michaela Geierhos 

Research Institute CODE, University of the Bundeswehr Munich,
Neubiberg, Germany

e-mail: {yeongsu.lee | hendrik.bothe | michaela.geierhos}@unibw.de

Abstract—This paper addresses the impact of Large Language Model (LLM)-based pseudonymization on cross-platform identifier similarity and re-identification risk. Cross-platform profile linking enables organizations to analyze user behavior across multiple social networking platforms. However, this practice raises significant privacy and security concerns. Identifiers such as usernames and real names can be used to reconstruct user identities when aggregated across different online platforms. Therefore, addressing these challenges is highly relevant to current research on privacy-preserving data analytics, digital identity protection, and secure information management. Using Linktree profile connection datasets from Facebook, Instagram, LinkedIn, and Twitter/X, we analyze how pseudonymization affects identifier similarity and the potential recovery of original identifiers. Our findings demonstrate that pseudonymization significantly reduces identifier similarity, resulting in low re-identification success rates. The results highlight the potential of LLM-driven privacy-enhancing techniques for mitigating identity disclosure risks while preserving limited analytical utility. Consequently, this work contributes to ongoing research in privacy-aware data processing, secure cross-platform analytics, and trustworthy LLMs for online social networks.

Keywords—pseudonymization; privacy-preserving data transformation; privacy-utility trade-off.

I. INTRODUCTION

Social networking platforms host vast amounts of user-generated data and have become important sources for digital analytics and sociotechnical research. Since individuals often have accounts on multiple platforms [1], researchers and organizations attempt to link user identities across different Social Networking Services (SNS) to better understand user behavior and digital presence [2]–[4]. However, cross-platform profile linking raises significant privacy concerns because identifiers, such as names and usernames, can be used to reconstruct user identities across platforms [5][6]. Recent research highlights how seemingly fragmented pieces of personal data can be combined to reconstruct digital identities. Studies on digital identity reconstruction demonstrate that data fragments from different platforms may collectively reveal sensitive personal information when linked together [5][7]. Similarly, prior work has shown that landing pages, such as Linktree, often act as central hubs that link multiple social media accounts. This facilitates cross-platform profile discovery and matching [8]. While these linkages enhance analytical capabilities, they also amplify privacy risks [9][10]. Consequently, there is a growing need for privacy-preserving mechanisms that enable social media analytics while protecting personal identifiers.

Previous studies have examined cross-platform profile linking [3][11] and the privacy risks associated with aggregating

social media data [9][10]. However, existing studies have primarily focused either on improving matching accuracy or on anonymization techniques that remove identifiers entirely [4][12]–[14]. Traditional privacy protection approaches include pseudonymization techniques that aim to reduce the exposure of personal identifiers while maintaining data usability for analytical purposes [15]–[17]. Unlike anonymization, pseudonymization transforms identifiers while preserving the structural properties necessary for analytical tasks [18][19]. However, traditional pseudonymization methods, such as rule-based transformations or manual approaches, fail to optimize the trade-off between privacy and data utility, often damaging semantic and pragmatic meaning of the original texts [18]. Recent advances in LLMs introduce new opportunities for semantic pseudonymization, which may preserve the structural properties of identifiers while reducing their direct identifiability [19]–[21]. Researchers have explored mixed-methods approaches to pseudonymizing social media profile data, highlighting both the opportunities and challenges of automated pseudonymization techniques [19]. However, little systematic research has evaluated the impact of LLM-based pseudonymization on cross-platform identity matching tasks or analyzed the resulting trade-off between privacy protection and analytical utility. To address this gap, this study explores the following research question: How does LLM-based pseudonymization affect the privacy-utility trade-off in cross-platform social media profile matching? To answer this question, we use the datasets of linked social media profiles introduced in [5][7][8]. These datasets exploit publicly accessible Linktree landing pages that connect user accounts across multiple platforms. We extract person names and usernames from the linked profiles and evaluate their similarity across platforms before and after LLM-based pseudonymization.

This paper makes three main contributions. First, it empirically evaluates the use of LLM-based pseudonymization to transform social media identifiers in cross-platform analysis scenarios. Second, it analyzes the effect of pseudonymization on identifier similarity across platforms using real-world Linktree-based profile connection datasets. Third, it assesses the privacy implications of pseudonymization by performing simulated re-identification attacks and employing ranking-based metrics to quantify the privacy-utility trade-off.

The remainder of this paper is structured as follows: Section II reviews related work on cross-platform identity linkage and privacy-preserving identifier transformation. Section III describes the research design and experimental setup, followed by the evaluation results in Section IV. Finally, Section V

concludes the paper and outlines directions for future work.

II. RELATED WORK

Cross-platform profile linking, also known as user identity linkage, is the process of identifying accounts belonging to the same individual across multiple social networking platforms. Existing approaches can be broadly categorized into three groups: attribute-based, graph-based, and anchor-based methods [4][8][12][22]. Attribute-based and graph-based methods operate directly on data from social networking services. Attribute-based approaches compare profile attributes, such as usernames or display names, across platforms. Graph-based methods exploit structural information, such as friendship relations or interaction networks. In contrast, anchor-based approaches rely on external references that explicitly link accounts across platforms. Landing pages that aggregate links to multiple social media profiles, such as Linktree, serve as hubs that connect the accounts of the same individual. These pages also provide valuable ground-truth datasets for cross-platform profile matching. However, cross-platform profile linking also raises significant privacy concerns, as discussed extensively in [5]. Consequently, privacy-preserving data processing has become an important research topic in data services and applications. Privacy-preserving data processing commonly relies on two main approaches: anonymization and pseudonymization. Anonymization aims to remove or generalize identifying information so that individuals cannot be re-identified [23]. Typical anonymization techniques include data suppression, generalization, or aggregation [12]. Although anonymization can provide strong privacy protection, it often removes information necessary for analytical tasks.

In contrast, pseudonymization replaces identifying attributes with alternative identifiers while preserving the data's structural and analytical properties [16][17]. Since pseudonymized data can support analytical processing, this approach is often used in systems requiring both privacy protection and data usability. Traditional pseudonymization methods include rule-based transformations, hashing techniques, and cryptographic transformations that replace identifiers with encoded representations [16][17]. While these methods can effectively hide original identifiers, they often disrupt the semantic relationships necessary for similarity-based analysis [18]. Recent advances in LLMs provide new opportunities for semantic pseudonymization of textual identifiers. LLM-based methods can generate alternative identifiers that preserve structural characteristics, such as length and linguistic patterns, while reducing direct identifiability [19]–[21][24]–[26]. However, empirical studies evaluating how such transformations affect analytical tasks such as cross-platform profile matching remain limited. In the context of social networking platforms, direct identifiers such as display names and usernames, represent key attributes for cross-platform identity linkage. Therefore, protecting these identifiers while preserving their analytical usefulness remains a central challenge in privacy-preserving social media analytics. Recent work has explored pseudonymization approaches that transform such identifiers while retaining structural properties

necessary for subsequent analysis, e.g., by combining LLM-based pseudonymization with lightweight encryption [19].

III. RESEARCH DESIGN AND EXPERIMENTAL SETUP

This study relies on publicly available datasets that were introduced in prior research on cross-platform profile linking and privacy risks in social media environments. Specifically, we use datasets derived from Linktree landing pages, as well as datasets employed in studies on digital identity reconstruction [5][7][8]. These datasets are available upon request. Linktree landing pages often serve as hubs that connect various social media accounts belonging to the same individual. Thus, these pages provide explicit cross-platform references between user profiles and can be used as the ground truth for linking identities across platforms. The overall processing pipeline is shown in Figure 1.

A. Dataset

The dataset is derived from landing pages and contains links to user profiles on various social networking platforms, such as Facebook, Instagram, LinkedIn, and Twitter/X. Two identifiers were extracted from these linked profiles in these datasets for further analysis: the person name (display name) and the username (platform handle). These identifiers are commonly used in cross-platform profile matching because they are publicly visible and often contain information that links user accounts across platforms [11]. Prior to analysis, the identifiers were normalized by converting them to lowercase and removing platform-specific formatting elements to ensure consistent similarity comparisons across platforms. Table I (a) summarizes the number of collected profiles per platform, and Table I (b) reports the overlap of profiles across platform combinations based on shared Linktree identifiers.

TABLE I. DATASET STATISTICS.

(a) Profiles per platform		(b) Cross-platform profiles	
Platform	Profiles	Platform Combination	Shared Profiles
Facebook (FB)	4,907	FB, IG, IN, TW	2,797
Twitter/X (TW)	4,953	FB, IG, TW	3,530
Instagram (IG)	4,566	FB, IN, TW	3,096
LinkedIn (IN)	3,735	FB, IG, IN	2,874
		IG, IN, TW	3,145
		FB-IG	3,648
		FB-IN	3,205
		FB-TW	3,992
		IG-IN	3,236
		IG-TW	4,006
		IN-TW	3,552

B. LLM-based Pseudonymization

To protect personal identifiers while maintaining analytical usability, the extracted identifiers were pseudonymized using LLMs. Specifically, the Llama language model [27] was used

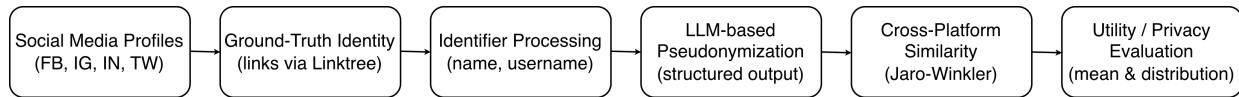


Figure 1. Overview of the proposed LLM-based pseudonymization and evaluation process. Social media profiles from Facebook, Instagram, LinkedIn, and Twitter/X are linked through Linktree-based ground-truth identities. User identifiers (names and usernames) are extracted and processed using an LLM-based pseudonymization approach with structured outputs. The resulting pseudonyms are then evaluated using cross-platform similarity analysis based on the Jaro-Winkler metric, followed by a utility and privacy assessment using similarity distributions and mean scores.

to generate alternative identifiers for both person names and usernames. The model was prompted to generate pseudonyms that removed direct personal references while maintaining the structural characteristics of the original identifiers. This approach reduces direct identifiability while preserving partial structural similarity, which may support analytical tasks such as cross-platform similarity analysis.

For prompting, Llama 3.3 with 70 billion parameters is employed via the Ollama Python library [28]. Table II shows an example of a query sent to the local model to process a dictionary-based input, along with the generated response. The `system` role specifies the model's behavior and provides instructions that guide the generation of responses to user requests.

TABLE II. AN EXAMPLE OF A PROMPT REQUEST AND RESPONSE.

Request	
role:	system
content:	Your task is to replace the name and username with a pseudonym in the given dict data. 1. Get the dict data! 2. Find out the items dict_data['name'] and dict_data['username'] from dict data and replace them with a pseudonym. 3. Keep the language and format of name and usernames for the pseudonyms! 4. Do not use the same pseudonym for different names and usernames! 5. Give the result back as dict format! 6. Do not include any explanation, example, transformation, or comments! Are you ready?
messages:	{'role': 'user', 'content': {'name': 'Renata Schmitt', 'username': 'renata.schmitt'}} {'role': 'assistant', 'content': {'name': 'Laris Lavati', 'username': 'laris.lavati'}}
parameter:	{temperature: 0.0, top_p: 0.9}
role:	user
content:	{ "name": "Vlad Plavlkov", "username": "vld-plalkv" }
Response	
role:	assistant
content:	{ "name": "Bryan Petrov", "username": "bryn-petrov" }

C. Evaluation Strategy

Before analyzing the similarity distributions, the consistency of the LLM-based pseudonymization process was verified. We

implemented pseudonymization using structured, dictionary-based inputs and outputs. This ensured that the generated pseudonyms would follow a predefined format and that both identifiers (person name and username) would be produced consistently. This approach guaranteed that all identifiers were successfully transformed without formatting errors or incomplete outputs. Consequently, no entries were removed during preprocessing, and the number of identifiers in the pseudonymized dataset is the same as in the original dataset.

To evaluate the impact of pseudonymization, the Jaro-Winkler similarity metric was computed across platform pairs to measure identifier similarity [29]. We analyzed similarity distributions for both the original and the pseudonymized identifiers to assess the transformation's effect on cross-platform similarity patterns. The Jaro-Winkler metric was chosen because it effectively measures short textual identifiers, such as personal names. It emphasizes prefix similarity and tolerates minor spelling variations [30][31]. We computed the average similarity between identifiers across platform pairs to quantify the overall reduction in identifier similarity caused by pseudonymization. To evaluate the remaining privacy risk, a simulated re-identification attack was performed, comparing the pseudonymized identifiers against candidate pools of original identifiers and ranking them according to the Jaro-Winkler metric. These analyses allow us to evaluate the privacy-utility trade-off between reduced identifier similarity and the risk of identifier recovery.

These experiments compare the similarity distributions and average similarity scores of original and pseudonymized identifiers to quantify the effect of pseudonymization on cross-platform identifier similarity. Additionally, a re-identification experiment is conducted to evaluate the privacy risks associated with pseudonymized identifiers. In this experiment, the identifiers are compared against candidate pools of original identifiers from the same platform, and the candidates are ranked according to their similarity scores. The effectiveness of this attack is evaluated using ranking-based metrics, such as Mean Reciprocal Rank (MRR), Mean Rank (MeanR) and Median Rank (MedR), and top- k accuracy.

IV. RESULTS AND EVALUATION

This section presents the results and evaluation metrics, as described in Section III.

A. Average Similarity Scores

Table III shows the average Jaro-Winkler similarity between identifiers across platform pairs for both the original and the pseudonymized data. Similarity values for original usernames

range from 0.747 to 0.845, indicating the frequent reuse of identical or highly similar usernames across platforms. The highest similarity is observed between Instagram and Twitter (0.845). After pseudonymization, however, similarity decreases across all platform pairs. The average similarity for usernames drops by roughly 0.15–0.20, depending on the platform combination. A smaller reduction is observed for person names, where the average similarity decreases from 0.79–0.81 to 0.69–0.71. Overall, LLM-based pseudonymization weakens direct identifier similarity while preserving partial structural characteristics of the original identifiers. This reflects a privacy-utility trade-off: reduced similarity improves privacy protection, yet still allows for limited, similarity-based analysis.

TABLE III. AVERAGE JARO-WINKLER SIMILARITY BETWEEN PLATFORM PAIRS. UN:USERNAME, O:ORIGINAL, AND P:PSEUDONYM.

Identifier	FB-TW	FB-IG	FB-IN	IG-IN	IG-TW	TW-IN
UN (O)	0.792	0.822	0.747	0.763	0.845	0.751
UN (P)	0.611	0.633	0.595	0.616	0.645	0.597
Name (O)	0.811	0.805	0.785	0.812	0.809	0.797
Name (P)	0.693	0.706	0.691	0.711	0.702	0.687

B. Similarity Distribution

Figures 2 and 3 show the distribution of Jaro-Winkler similarity scores across different platform pairs, both using original and pseudonymized identifiers. A large proportion of username pairs exhibit very high similarity values in the original data. In particular, most username pairs fall into the highest similarity interval (≥ 0.9). For instance, 61.8 % of username pairs on Instagram and Twitter, as well as 56.9 % on Facebook and Instagram, fall into this interval. These results confirm that many users reuse identical or highly similar usernames across platforms. However, compared to the original data, the proportion of very high similarity values (≥ 0.9) decreases substantially across all platform pairs for the pseudonymized data. Instead, most pairs shift toward lower similarity intervals, particularly the <0.5 and $0.5\text{--}0.69$ ranges. For instance, for the Facebook-Twitter pair the proportion of similarity values below 0.7 increases to more than 70 %. This indicates that LLM-based pseudonymization effectively reduces direct structural similarity between usernames.

Similar to usernames, original person names exhibit a strong concentration in the highest similarity interval (≥ 0.9). Across all platform pairs, approximately half of the name pairs fall into this range. After applying LLM-based pseudonymization, the similarity distribution shifts toward lower similarity intervals. The proportion of highly similar names decreases significantly, while the proportion of medium similarity values ($0.5\text{--}0.69$) increases across most platform pairs. This indicates that the pseudonymization process removes direct name correspondences while preserving partial structural similarity.

These results indicate a clear trade-off between privacy and utility introduced by LLM-based pseudonymization. While original identifiers provide strong signals for cross-platform matching, pseudonymization substantially reduces the similarity

of identifiers across platforms. For usernames, the proportion of highly similar identifiers (similarity ≥ 0.9) drops from about 47 % to 17 %, and the matching utility (recall with a threshold of 0.7) decreases from approximately 64 % to 24 %. A similar, albeit less pronounced, effect is observed for person names: highly similar identifiers decrease from approximately 51 % to 27 %, and recall decreases from about 68 % to 40 %. Overall, the results show that pseudonymization weakens direct identifier correspondence across platforms while preserving limited structural characteristics. While this reduction lowers matching performance, it significantly decreases the risk of direct identity reconstruction across social networking services.

C. Re-Identification Attack

To evaluate the privacy risks associated with pseudonymized identifiers, we conducted a re-identification (re-ID) attack simulation. In this scenario, an attacker observes a pseudonymized identifier and attempts to find its corresponding original identifier by searching a pool of identifiers from the same platform. Similarity scores are computed for each pseudonymized identifier (p_i) against all original identifiers (o_j) using the Jaro-Winkler metric. The original identifiers are then ranked according to their similarity scores. The position of the correct identifier in this list determines the success of the attack. Table IV summarizes the results of the simulated re-identification attack.

We evaluate the attack using ranking-based metrics commonly used in information retrieval [32]–[34]. The MRR metric measures how early the correct identifier appears in the ranked candidate list. Additionally, MeanR and MedR describe the average and median rank positions of the correct identifier, respectively. Top- k accuracy measures whether the correct identifier appears among the top k candidates ($k \in \{1, 3, 5, 10\}$). The candidate pool size differs from the total number of profiles because some Linktree landing pages reference multiple accounts from the same platform. In such cases, several social media profiles share the same Linktree identifier. These profiles are consolidated when constructing the candidate pool in order to avoid duplicate ground-truth mappings.

Overall, success rates remain relatively low across all platforms. For usernames, top-1 accuracy ranges from about 1.4 % to 2.6 %, indicating that directly reconstructing original usernames from pseudonymized identifiers is difficult. Person names have slightly higher re-identification rates, reaching 5.8 % on Twitter. Ranking-based metrics further confirm this observation. The relatively high mean and median ranks suggest that correct identifiers usually appear toward the end of the ranked candidate list. Although top- k accuracy increases with larger k , the overall success rates are low compared to the size of the candidate pools. These results suggest that LLM-based pseudonymization substantially reduces the risk of identifier reconstruction, though it does not eliminate the possibility of similarity-based re-identification.

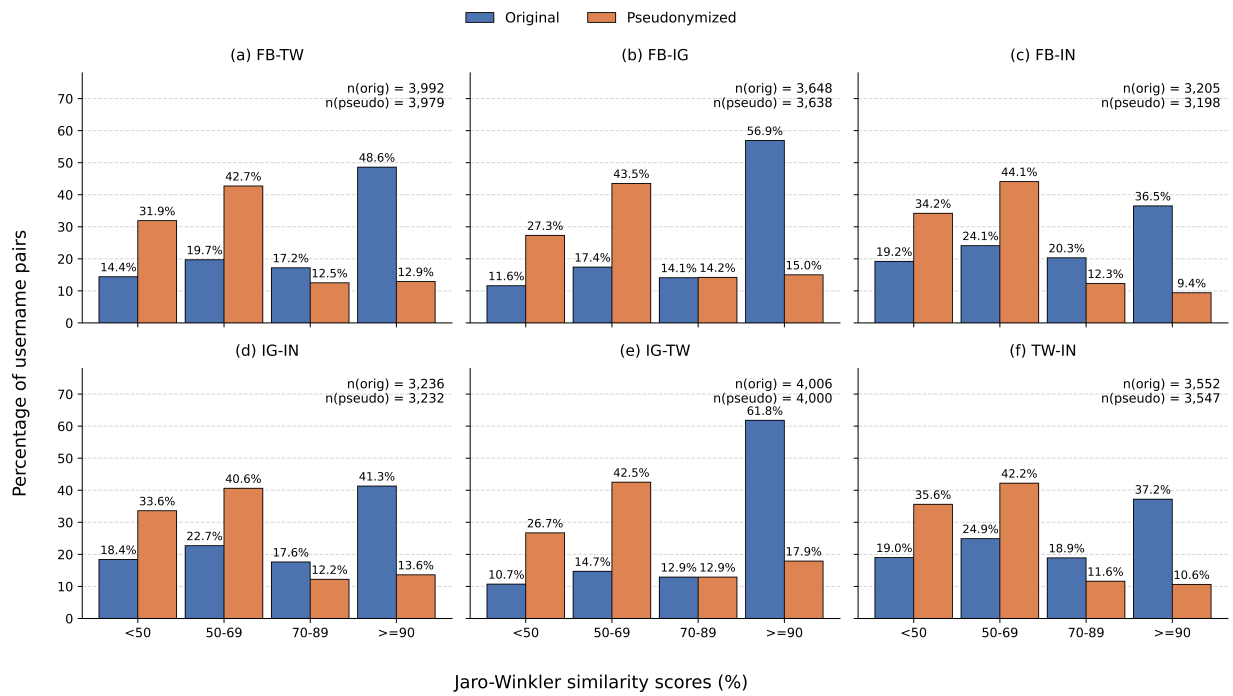


Figure 2. Similarity distribution of usernames. The slight reduction of numbers of pseudonymized usernames indicates that some pseudonymized usernames are duplicated.

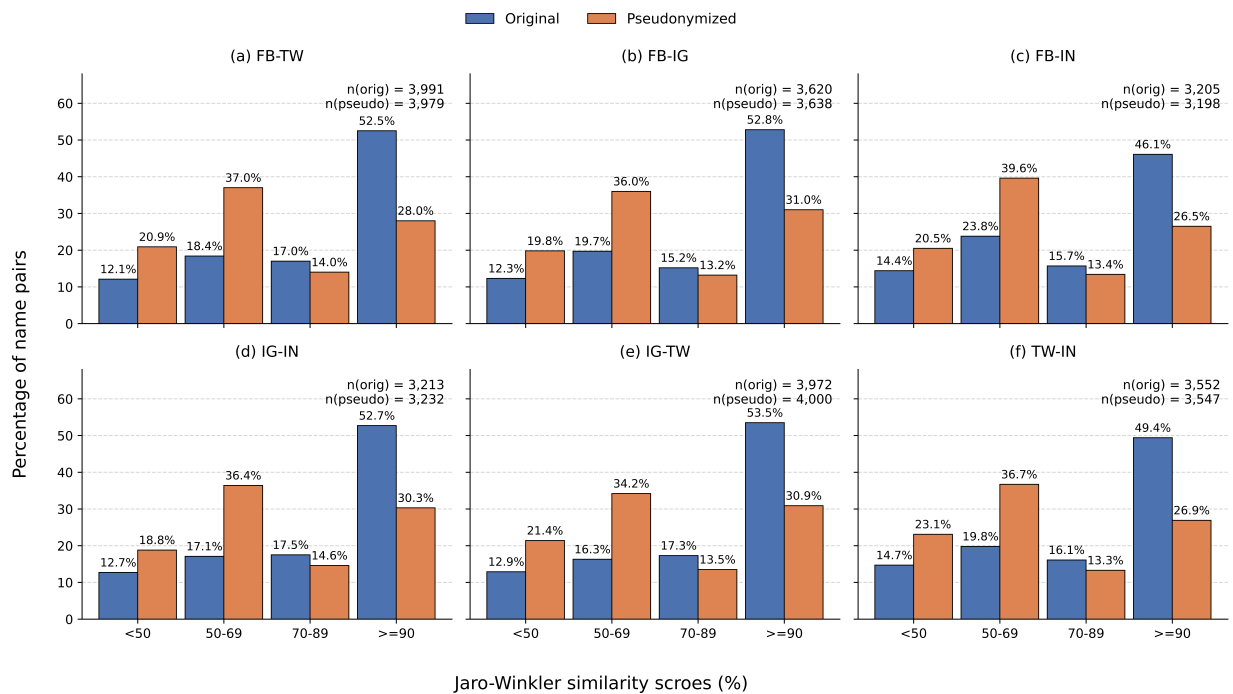


Figure 3. Similarity distribution of names. The slight difference of numbers of names from those of Table I indicates that some names in original and pseudonymized names are duplicated.

V. CONCLUSION AND FUTURE WORK

This study examined the impact of LLM-based pseudonymization on the similarity of social media identifiers across platforms. Using Linktree-based profile connection

datasets, we analyzed the impact of pseudonymization on two frequently used identifiers for cross-platform identity linkage: usernames and personal names. Our findings reveal that LLM-based pseudonymization significantly reduces the similarity of identifiers across platforms. This reduction

TABLE IV. RE-IDENTIFICATION ATTACK RESULTS

Platform	Identifiers	Pool	MRR	MeanR	MedR	Top-1	Top-3	Top-5	Top-10
Facebook	Username	4,169	0.035	1,329.96	965.0	2.10 %	3.39 %	4.31 %	6.06 %
	Name	4,168	0.059	1,145.24	762.5	3.84 %	6.05 %	7.50 %	9.21 %
Twitter	Username	4,646	0.035	1,525.25	1,193.0	2.12 %	3.60 %	4.24 %	5.68 %
	Name	4,646	0.086	1,085.69	572.0	5.76 %	8.93 %	10.71 %	13.58 %
Instagram	Username	4,149	0.041	1,326.86	996.0	2.57 %	3.91 %	5.31 %	6.43 %
	Name	4,115	0.078	1,066.99	639.5	5.19 %	8.29 %	9.90 %	12.29 %
LinkedIn	Username	3,697	0.026	1,115.44	837.0	1.45 %	2.44 %	3.32 %	4.29 %
	Name	3,697	0.035	1,137.68	848.0	1.98 %	3.56 %	4.61 %	5.95 %

is particularly pronounced for usernames, which often contain unique identifiers that are reused across multiple SNS. Person names exhibit a smaller decrease in similarity, suggesting that linguistic structure is partially preserved during pseudonym generation. These results highlight a privacy-utility trade-off. While pseudonymization weakens direct identifier correspondence and reduces the risk of identity reconstruction across platforms, partial structural similarity remains, which may support analytical tasks based on similarity. From an information systems perspective, this behavior is relevant to social media analytics scenarios in which identifiers must be transformed before analysis or data sharing in order to comply with privacy requirements.

Future research could expand upon this work in several ways. First, including additional profile attributes, such as biographies, profile descriptions, and user-generated content, would allow for a better evaluation of pseudonymization in richer social media contexts. Second, alternative baselines, prompting strategies, and model configurations could be explored to assess the impact of structural similarity preservation, reproducibility, and output consistency. Third, future studies could investigate the privacy-utility trade-off further by analyzing downstream analytical tasks and evaluating more advanced cross-platform linkage and re-identification attacks under stronger adversarial settings. Finally, future studies should address ethical limitations and residual risks, emphasizing that pseudonymization does not guarantee anonymization.

ACKNOWLEDGMENT

This research within the MuQuaNet project is partly funded by dtcc.bw – Digitalization and Technology Research Center of the Bundeswehr. dtcc.bw is funded by the European Union – NextGenerationEU.

REFERENCES

- [1] DataReportal, *Digital 2026: Global overview report*, Feb. 2026.
- [2] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics – challenges in topic discovery, data collection, and data preparation”, *International Journal of Information Management*, vol. 39, pp. 156–168, 2018, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2017.12.002.
- [3] M. Wang, W. Chen, J. Xu, P. Zhao, and L. Zhao, “User profile linkage across multiple social platforms”, in *Web Information Systems Engineering – WISE 2020*, Z. Huang, W. Beek, H. Wang, R. Zhou, and Y. Zhang, Eds., Cham: Springer International Publishing, 2020, pp. 125–140, ISBN: 978-3-030-62005-9.
- [4] C. Senette, M. Siino, and M. Tesconi, “User identity linkage on social networks: A review of modern techniques and applications”, *IEEE Access*, vol. 12, pp. 171 241–171 268, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3500374.
- [5] F. S. Bäumer, S. Schultenkämper, M. Geierhos, and Y. S. Lee, “Mirroring Privacy Risks with Digital Twins: When Pieces of Personal Data Suddenly Fit Together”, *SN COMPUT. SCI.*, vol. 5, 2024. DOI: 10.1007/s42979-024-03413-z.
- [6] S. Min, Y. Luo, K. Liu, Q. Gong, and Y. Chen, “From identification to obfuscation: A survey of cross-network mapping and anti-mapping methods”, *Computers, Materials and Continua*, vol. 86, no. 2, pp. 1–23, 2025, ISSN: 1546-2218. DOI: 10.32604/cmc.2025.073175.
- [7] S. Schultenkämper, F. S. Bäumer, B. Bellgrau, Y. S. Lee, and M. Geierhos, “From Digital Tracks to Digital Twins: On the Path to Cross-Platform Profile Linking”, in *Enterprise Design, Operations, and Computing. EDOC 2023 Workshops*, T. P. Sales et al., Eds., Cham: Springer Nature Switzerland, 2024, pp. 158–171, ISBN: 978-3-031-54712-6. DOI: 10.1007/978-3-031-54712-6_10.
- [8] S. Denisov and F. S. Bäumer, “The Only Link You’ll Ever Need: How Social Media Reference Landing Pages Speed Up Profile Matching”, in *Information and Software Technologies*, A. Lopata, D. Gudonienė, and R. Butkienė, Eds., Cham: Springer International Publishing, 2022, pp. 136–147, ISBN: 978-3-031-16302-9. DOI: 10.1007/978-3-031-16302-9_10.
- [9] S. J. De and A. Imine, “Privacy Risk Analysis of Online Social Networks”, *Synthesis Lectures on Information Security, Privacy, and Trust*, vol. 10, no. 1, 2020, ISSN: 1945-9742. DOI: 10.2200/s01056ed1v01y202009spt024.
- [10] W. Wong, Z. Alomari, and Y. Liu, “Linkage Deanonimization Risks, Data-Matching and Privacy: A Case Study”, in *8th International Conference on Cryptography, Security and Privacy (CSP)*, 2024. DOI: 10.1109/CSP62567.2024.00010.
- [11] Y. Li, Y. Peng, Z. Zhang, H. Yin, and Q. Xu, “Matching user accounts across social networks based on username and display name”, *World Wide Web*, vol. 22, no. 3, 2019, ISSN: 15731413. DOI: 10.1007/s11280-018-0571-4.
- [12] A. Majeed and S. Lee, “Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey”, *IEEE Access*, vol. 9, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3045700.
- [13] J. Zhao, B. Su, X. Rao, and Z. Chen, “A Cross-Platform Personalized Recommender System for Connecting E-Commerce

- and Social Network”, *Future Internet*, vol. 15, no. 1, 2023, ISSN: 1999-5903. DOI: 10.3390/fi15010013.
- [14] M. G. Ayadi, H. Mezni, H. Elmannai, and R. I. Alkanhel, “Privacy-preserving cross-network service recommendation via federated learning of unified user representations”, *Data & Knowledge Engineering*, vol. 158, p. 102422, 2025, ISSN: 0169-023X. DOI: 10.1016/j.datak.2025.102422.
- [15] European Commission, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016.
- [16] European Union Agency for Cybersecurity (ENISA), “Data pseudonymisation: Advanced techniques and use cases”, European Union Agency for Cybersecurity (ENISA), Technical Report, Jan. 2021. DOI: 10.2824/860099.
- [17] European Union Agency for Cybersecurity (ENISA), “Deploying pseudonymisation techniques: The case of the health sector”, European Union Agency for Cybersecurity (ENISA), Technical Report, Mar. 2022. DOI: 10.2824/092874.
- [18] E. Volodina, S. Dobnik, T. L. M. Tiedemann, and X. S. Vu, “Grandma Karl is 27 years old - research agenda for pseudonymization of research data”, in *Proceedings - IEEE 9th International Conference on Big Data Computing Service and Applications, BigDataService 2023*, 2023. DOI: 10.1109/BigDataService58306.2023.00047.
- [19] Y. S. Lee, H. Bothe, and M. Geierhos, “A mixed-methods approach to pseudonymizing users’ social media profile data”, in *2025 12th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Nov. 2025, pp. 50–57. DOI: 10.1109/SNAMS67467.2025.11390950.
- [20] O. Yermilov, V. Raheja, and A. Chernodub, “Privacy- and Utility-Preserving NLP with Anonymized Data: A case study of Pseudonymization”, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. DOI: 10.18653/v1/2023.trustnlp-1.20.
- [21] T. Yang, X. Zhu, and I. Gurevych, “Robust utility-preserving text anonymization based on large language models”, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 28922–28941, ISBN: 979-8-89176-251-0.
- [22] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, “User identity linkage across online social networks: A review”, *SIGKDD Explor. Newsl.*, vol. 18, no. 2, pp. 5–17, Mar. 2017, ISSN: 1931-0145. DOI: 10.1145/3068777.3068781.
- [23] A. Rossi, M. P. Arenas, E. Kocyigit, and M. Hani, “Challenges of protecting confidentiality in social media data and their ethical import”, in *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022, pp. 554–561. DOI: 10.1109/EuroSPW55150.2022.00066.
- [24] D. Pissarra et al., “Unlocking the potential of large language models for clinical text anonymization: A comparative study”, in *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, I. Habernal et al., Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 74–84.
- [25] S. Sousa, M. Jantscher, M. Kröll, and R. Kern, “Large Language Models for Electronic Health Record De-Identification in English and German”, *English, Information*, vol. 16, no. 2, Feb. 2025, ISSN: 2078-2489. DOI: 10.3390/info16020112.
- [26] O. Dorémus et al., “Harnessing Moderate-Sized Language Models for Reliable Patient Data Deidentification in Emergency Department Records: Algorithm Development, Validation, and Implementation Study”, *JMIR AI*, vol. 4, e57828, Apr. 2025, ISSN: 2817-1705. DOI: 10.2196/57828.
- [27] A. Grattafiori et al., *The Llama 3 Herd of Models*, 2024. arXiv: 2407.21783 [cs.AI].
- [28] Ollama, “Ollama”, Accessed: May 29, 2026. [Online]. Available: <https://ollama.com>.
- [29] RapidFuzz, “JaroWinkler GitHub Repository”, Accessed: May 29, 2026. [Online]. Available: <https://github.com/rapidfuzz/JaroWinkler>.
- [30] W. Winkler, “String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage”, in *Proceedings of the Survey Research Methods*, 1990, pp. 354–359.
- [31] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, “A comparison of string distance metrics for name-matching tasks”, in *Proceedings of the 2003 International Conference on Information Integration on the Web*, ser. IIWEB’03, Acapulco, Mexico: AAAI Press, 2003, pp. 73–78. DOI: 10.5555/3104278.3104293.
- [32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, Web publication at informationretrieval.org.
- [33] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data”, in *Advances in neural information processing systems*, 2013, pp. 2787–2795.
- [34] N. Craswell, “Mean reciprocal rank”, in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 1703–1703, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_488.

Agreement-based SVM One-Class Classification: an Ensemble Method for Software Defect Prediction

Luigi Lavazza

Università degli Studi dell'Insubria
Varese, Italy
email:luigi.lavazza@uninsubria.it

Sandro Morasca

Università degli Studi dell'Insubria
Varese, Italy
email:sandro.morasca@uninsubria.it

Gabriele Rotoloni

Università degli Studi dell'Insubria
Varese, Italy
email:gabriele.rotoloni@uninsubria.it

Abstract—One-class Support Vector Machine (SVM) classification is well-suited to train classifiers when the training dataset is very imbalanced. So, it is increasingly used for software defect prediction, since faulty modules are usually relatively rare, compared to non-faulty ones. Though a minority, faulty modules often are a non-negligible fraction of the total; hence, completely ignoring them when training a one-class model can lead to discarding potentially useful information: we investigate how to exploit the information embedded in the minority class when using one-class SVM classification. To this end, we devised an ensemble software defect prediction method that takes advantage of one-class SVM classification without ignoring the minority class modules. Two one-class SVM defectiveness classification models are built: one based on the majority class modules alone and the other on the minority class modules alone. If both models agree on the defectiveness of a module, it is classified accordingly; otherwise, as “uncertain.” Uncertainty can then be treated in many different ways. Here, we adopt a “prudential” approach: a module is classified non-faulty only if it is estimated non-faulty by both one-class classifiers; otherwise, faulty. The proposed approach was tested by comparing its results with the ones obtained via two-class and one-class SVM classifiers. The results show that our technique is a valid alternative to both two-class and one-class classifications. Specifically, our approach achieves lower global cost when the cost of a false negative (i.e., a faulty module incorrectly believed non-faulty) is sufficiently higher than the cost of a false positive (a non-faulty module incorrectly believed faulty).

Keywords—Software defect prediction; ensemble methods; machine Learning; Support Vector Machine; One-class classification; One-class SVM.

I. INTRODUCTION

Traditional Support Vector Machine (SVM) classification distinguishes two or more classes by learning from a training set with instances from all the classes. Instead, One-Class Classification (OCC) identifies instances of a specific class amongst all instances, by learning from a training set with only instances of that class.

OCC is particularly useful when instances of one class are rare, so that the available training sets are severely imbalanced. This is often the case with software defect prediction, since usually a small subset of the modules (e.g., classes) in a software system is faulty, so most datasets are imbalanced in favor of non-faulty modules. For datasets with medium to high imbalance, OCC methods may perform better than traditional methods [1].

However, completely ignoring information on the minority class instances may not be a good idea, especially because

faulty software modules may still account for 10–20% of the total (see Section III-A). Faulty modules are definitely fewer than non-faulty ones, but they are not really extremely rare, in absolute terms.

In this paper, we investigate whether useful information can be obtained by considering both the majority and minority classes, while using OCC SVM. We define an ensemble method that uses the results of two OCC SVM models, one based on the majority class and the other on the minority class. We call this method ABOCC (Agreement-Based One-Class Classification).

When the two OCC models agree that a module is non-faulty, ABOCC classifies it as non-faulty. Likewise, when both models classify it as faulty, the module is classified as faulty by ABOCC. A module that receives conflicting classifications from the two OCC models is classified as “uncertain.” There are several ways to deal with this uncertainty. We here investigate a “prudential” approach, in which uncertain modules are classified as faulty.

We carried out an empirical study to evaluate ABOCC and compare it with both traditional SVM and OCC SVM.

We find that “prudential” ABOCC is the best at minimizing false negatives, but it yields the most false positives. Therefore, when releasing a faulty module has severe consequences, ABOCC is economically more beneficial than traditional classification. These results show that considering both the minority class and majority class can be important in some contexts.

The paper is organized as follows. Section II describes the proposed ensemble method. Section III describes the empirical study. The execution and results of the study are discussed in Section IV, along with hints on how to use the proposed method. Section V discusses the threats to the validity of the empirical study. Section VI accounts for related work. Section VII draws some conclusions and outlines future work.

II. AN ENSEMBLE CLASSIFICATION APPROACH BASED ON ONE-CLASS CLASSIFICATIONS

Building prediction models using the data from just the majority class of instances has the obvious drawback of not considering the knowledge embedded in minority class instances. To avoid this issue, ABOCC uses data from both classes.

The idea is to train two OCC classifiers, one only on positive instances and one only on negative instances, and then combine

them via simple voting. If both classifiers agree, there is reasonable confidence that the classification is correct (more than with a classification by a single classifier, whether obtained on the entire dataset or on one class only). If, instead, the two classifications are conflicting, leaving a module's classification uncertain seems reasonable, at least at first.

A. Classification with Uncertainty

The ABOCC classifier is defined as follows.

- 1) A binary classifier is built via one-class classification, using only positive instances. We denote it as bc_p .
- 2) A binary classifier is built via one-class classification, using only negative instances. We denote it as bc_n .
- 3) An instance i is then classified using bc_p and bc_n : let $c_p(i)$ and $c_n(i)$ be the obtained classifications, respectively. The classification $c_e(i)$ by ABOCC is

$$c_e(i) = \text{positive} \Leftrightarrow c_p(i) = \text{positive} \wedge c_n(i) = \text{positive}$$

$$c_e(i) = \text{negative} \Leftrightarrow c_p(i) = \text{negative} \wedge c_n(i) = \text{negative}$$

$$c_e(i) = \text{uncertain} \Leftrightarrow c_p(i) \neq c_n(i)$$

The one-class classification method used in steps (1) and (2) above is one-class SVM, a kernel-based method [2] of SVM, proposed by Schölkopf and Müller in 2001 [3][4]. The main idea of one-class SVM is that training data in input space are mapped into feature space via a kernel to find a hyperplane with a maximum margin in feature space to separate the transformed data from the origin.

Since the two OCC models are trained on two separate datasets, ABOCC mitigates a problem of vote-based ensemble models, i.e., the possible correlation among the training set features. Take two classifiers, one using the lines of code and the other the number of statements, trained on a dataset. The two metrics are highly correlated, so the two classifications will be very similar. If these models, along with a third model, are used in a vote-based ensemble model, their classification always prevails on the third model.

B. The Prudential Classification Approach

Since uncertain instances must be classified one way or the other, one possibility is to consider all uncertain instances as positive. This decision corresponds to a prudential attitude, which in software engineering is justified by the cost of not properly treating defective modules, which can cause severe consequences if released to end-users. The corresponding ensemble classifier, ABOCC_p (where 'p' is for prudential or pessimistic), is defined as follows.

Given bc_p and bc_n described in Section II-A, the classification by the binary prudential ensemble method $c_{bp}(i)$ is

$$c_{bp}(i) = \text{negative} \Leftrightarrow c_p(i) = \text{negative} \wedge c_n(i) = \text{negative};$$

$$c_{bp}(i) = \text{positive} \Leftrightarrow c_p(i) = \text{positive} \vee c_n(i) = \text{positive}.$$
 Unlike ABOCC, which returns three classes, i.e., positive, negative, and uncertain, ABOCC_p is a proper binary classifier.

III. EMPIRICAL EVALUATION: GOALS AND METHOD

The proposed method ABOCC_p was evaluated via an empirical study. We here describe the goals and organization of the study; its execution and results are described in Section IV.

To investigate whether ignoring one of the classes has an ill effect on classification, we evaluate ABOCC_p compared to traditional classification models with the following research question.

Research Question: How well does ABOCC_p perform compared to “regular” SVM, one-class SVM, and random classification?

To answer the research question, we consider the performance in predicting (1) actually faulty and actually non-faulty modules separately, and (2) the entire test set, containing both faulty and not-faulty modules. To this end, we use widely used performance metrics.

Though performance metrics provide quantitative indications of how good a model is at predicting faultiness, software developers are not all that interested in optimizing this kind of abstract evaluations. Instead, they want to minimize total software costs, which include the cost of development activities and the costs due to software failures after the release to end-users. If we assume that modules estimated faulty undergo some quality improvement activity, and modules estimated non-faulty are released as-is, costs are clearly related to the performance of faultiness predictors. Thus, we could evaluate predictors based on the costs they imply. These considerations are out of the scope of this paper and will be the object of future work.

A. Dataset

We perform cross-version prediction: the data from a version are the training set, and the data from the following version are the test set. For example, when the versions available in the collection are `project-1.1`, `project-1.2`, and `project-2.0`, first we use `project-1.1` for training and `project-1.2` for testing, then we use `project-1.2` for training and `project-2.0` for testing.

We used the dataset collection by Jureczko and Madeyski (J&M in what follows) [5], available from Zenodo [6]. Specifically, we used the datasets providing data from multiple versions of the same project, namely those concerning projects `ant`, `ivy`, `jedit`, `lucene`, `poi`, `synapse`, `velocity`, `xalan`, `xerces`. Overall, 24 training-testing dataset pairs were used. Descriptive statistics for the dataset collections are in Table I.

TABLE I. DESCRIPTIVE STATISTICS.

	mean	stdev	median	min	max
Modules	423	212	351	178	909
Prevalence	0.340	0.248	0.258	0.022	0.988

B. Model Building

We built software defect prediction models via “traditional” binary SVM, one-class SVM, and ABOCC_p. We also built prediction models based on random classification: these models were used as baselines, since one expects that a practically useful prediction model performs better than random classification.

As for hyperparameter tuning, SVM was tuned on cost, epsilon, and gamma. We used the R function `tune.svm`, from library `e1071`, which performs a Grid Search with cross validation; the process was repeated four times and we selected the best performing model.

For the one-class models, the most important parameter to tune is ν , the expected proportion of outliers in the training set. Since there is no real consensus on its possible values, we considered both low values, like 0.05, as in the literature [7] and higher values for different possible models, specifically 0.1, 0.2, 0.5 and 0.8. Since `tune.svm` does not support the tuning of one-class models, we wrote a function to perform the Grid Search with cross validation that chooses the best performing model. Like before, the process is repeated four times and the best performing model is selected.

We built models by using the training set “as-is,” without balancing, and by using training sets balanced via SMOTE [8]. We used the R library `performanceEstimation` and its `smote` function, which performs both over-sampling and under-sampling.

Over-sampling generates elements of the minority class using the K-nearest neighbors. We set $K=5$, like Chawla et al. did in their original paper. The number of elements generated depends on parameter `perc.over`. For example, with `perc.over=2`, the algorithm generates twice the amount of minority class elements.

Under-sampling randomly selects elements of the majority class. Parameter `perc.under` indicates the amount of elements to pick in relation to the amount of elements generated during the over-sampling. For example, `perc.under=3` means that the algorithm will pick three times the amount of elements generated by the over-sampling. Thus, the size of the resulting dataset depends on the number of elements of the minority class, meaning that the final result can be smaller than the starting dataset, especially when the ratio between minority and majority is very small. The larger the values of `perc.over` and `perc.under`, the larger and less balanced the resulting dataset. To achieve a good trade-off between size and balance, we set `perc.over=perc.under=2`, which makes prevalence (the proportion of positive instances) equal to $3/7$, since the positives are the minority class. The result is a relatively large datasets with total amount of elements seven times the original number of minority class elements.

Random classification was performed by estimating every module faulty with probability equal to ρ , the proportion of faulty modules in the test set. The performance of random classification is computed on the mean values of TP, FN, TN and FP.

C. Model Evaluation and Comparison

For each training set, each model yields a confusion matrix.

Given a dataset containing n modules, of which AP actually positive (i.e., faulty) and AN actually negative (i.e., non-faulty), each model estimates EP modules positive and the remaining EN modules negative. Of the EP estimated positives, TP are true positives, i.e., they are correctly estimated; the remaining

TABLE II. A CONFUSION MATRIX.

	Actual Neg.	Actual Pos.	
Est. Negative	TN	FN	EN=TN+FN
Est. Positive	FP	TP	EP=FP+TP
	AN=TN+FP	AP=FN+TP	$n=AN+AP=EN+EP$

FN are false negatives, i.e., they are positive, but are incorrectly estimated negative. Similarly, of the EN estimated negatives, TN are true negatives, i.e., they are correctly estimated; the remaining FP are false positives, i.e., they are negative, but are incorrectly estimated positive.

AP and AN, hence $n=AP+AN$, are properties of the dataset and so is prevalence $\rho = \frac{AP}{n}$, the proportion of faulty instances. A perfectly balanced dataset has $\rho = \frac{1}{2}$; the closer ρ to zero or one, the more imbalanced is the dataset.

Usually, the performance of a given classifier is expressed via performance metrics, defined as functions of the confusion matrix cells. We evaluate classifiers via the following performance metrics.

- $TPR = \frac{TP}{AP}$. The True Positive Rate is the proportion of actually faulty modules that were correctly classified. TPR is a quite popular metric, also known as recall or sensitivity.
- $TNR = \frac{TN}{AN}$. The True Negative Rate is the proportion of actually non-faulty modules that were correctly classified. It is also known as specificity. Note that $TNR=1-FPR$, where FPR is the False Positive Ratio, used—along with TPR—to define Receiver Operating Characteristic (ROC) curves.
- $Gmean = \sqrt{TPR \cdot TNR}$, the geometric mean of TPR and TNR.
- $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{AP \cdot AN \cdot EP \cdot EN}}$, the Matthews Correlation Coefficient [9].

TPR and TNR are used to represent a model’s performance in correctly classifying faulty and, respectively, non-faulty modules. Gmean and MCC are used to provide “global” evaluations of the model, thus summarizing the contents of the confusion matrix. We use both Gmean and MCC as “overall” performance indicators because Gmean, being the geometric mean if TPR and TNR, does not depend on the (im)balance of the test set, while MCC takes into account the proportion of faulty modules, along with TPR and TNR.

For the random model, the mean performance metrics are $TPR_{rnd} = \rho$, $TNR_{rnd} = 1 - \rho$, and $MCC_{rnd} = 0$ [10]. We are not aware of a formula for computing $Gmean_{rnd}(\rho)$, hence we use values of $Gmean_{rnd}(\rho)$ obtained via simulation.

IV. EXPERIMENTAL EVALUATION: RESULTS

We report the results obtained for five types of models:

- SVM. Since the literature reports that better results are achieved with balanced datasets, we used SMOTE to obtain fairly balanced datasets to train SVM models.
- One-class SVM. Since this method is deemed useful for unbalanced datasets, we used it without balancing the datasets.

- Random classification is carried out estimating each module with a probability (the proportion of faulty modules) that depends on the testing set: there is no training set, and using a balancing technique like SMOTE is just not applicable.
- Concerning $ABOCC_p$, we investigated whether it could benefit from training set balancing. Hence, we created $ABOCC_p$ models both with and without SMOTE.

A summary of the performance metrics obtained with the prediction methods is in Table III. In the table and afterwards, the methods are labeled as follows: OCSVM indicates one-class SVM, SVM_SMOTE indicates traditional two-class SVM with SMOTE, ABOCC_SMOTE and ABOCC indicate $ABOCC_p$ with and without SMOTE.

TABLE III. SUMMARY OF PERFORMANCE METRICS PER METHOD.

		mean	sd	median	min	max
OCSVM	TPR	0.596	0.206	0.584	0.081	0.944
	TNR	0.553	0.197	0.589	0.081	1.000
	Gmean	0.543	0.129	0.581	0.218	0.708
	MCC	0.114	0.167	0.14	-0.234	0.376
SVM_SMOTE	TPR	0.497	0.248	0.550	0.178	0.958
	TNR	0.686	0.212	0.743	0.192	1.000
	Gmean	0.539	0.123	0.540	0.383	0.765
	MCC	0.147	0.136	0.121	-0.140	0.419
ABOCC	TPR	0.801	0.145	0.808	0.599	1.000
	TNR	0.341	0.192	0.377	0.007	0.606
	Gmean	0.476	0.174	0.525	0.081	0.697
	MCC	0.127	0.110	0.120	-0.075	0.360
ABOCC_SMOTE	TPR	0.868	0.117	0.886	0.605	1.000
	TNR	0.264	0.168	0.291	0.000	0.565
	Gmean	0.422	0.194	0.487	0.000	0.618
	MCC	0.121	0.082	0.134	-0.004	0.255
random	TPR	0.340	0.248	0.257	0.023	0.988
	TNR	0.660	0.248	0.742	0.013	0.978
	Gmean	0.393	0.107	0.433	0.114	0.500
	MCC	0.000	0.000	0.000	0.000	0.000

We illustrate the comparisons among predictors' performances via boxplots, whose indications are confirmed by the Wilcoxon Sign Rank test. The results of the Wilcoxon Sign Rank tests are not reported here because of space reasons; however, all of them were significant at the usual $\alpha = 0.05$ level.

Figure 1 shows the boxplots of the distributions of TPR and TNR obtained by faultiness predictors; the blue triangles represent the mean values. TPR quantifies the performance of the predictors applied to actually faulty modules, while TNR represents the performance of the predictors applied to actually non-faulty modules. Figure 1 shows that no model performs best for both TPR and TNR:

- $ABOCC_p$, both with and without SMOTE, has the best TPR and the worst TNR.
- SVM with SMOTE and random classification perform best in terms of TNR, but have low TPR, i.e., they are quite inaccurate when predicting actually faulty modules.
- One-class SVM performs decently with both faulty and non-faulty modules. Nonetheless, it is outperformed by $ABOCC_p$ at predicting faulty modules and by SVM and random classification at predicting non-faulty modules.

$ABOCC_p$ achieves more extreme performances (i.e., better TPR and worse TNR) when trained with datasets balanced via SMOTE.

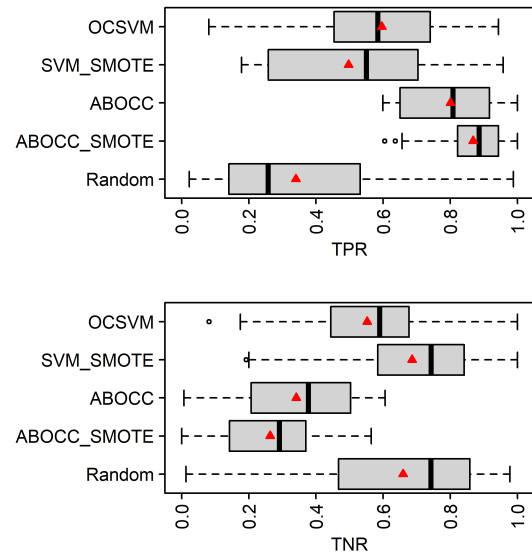


Figure 1. Boxplots of TPR and TNR.

Random classification is quite good at predicting non-faulty modules. This is not surprising, since non-faulty modules are the vast majority in most datasets (the median prevalence is 0.258, according to Table I). Hence, random classification predicts modules non-faulty with high probability, which leads to having high values of EN and TN. Symmetrically, random classification gives low EP and TP, hence it achieves low TPR.

As no model has both the best TPR and TNR, we check which method has the best “overall” performance, in terms of Gmean and MCC (see Section III-C). Figure 2 shows Gmean and MCC boxplots.

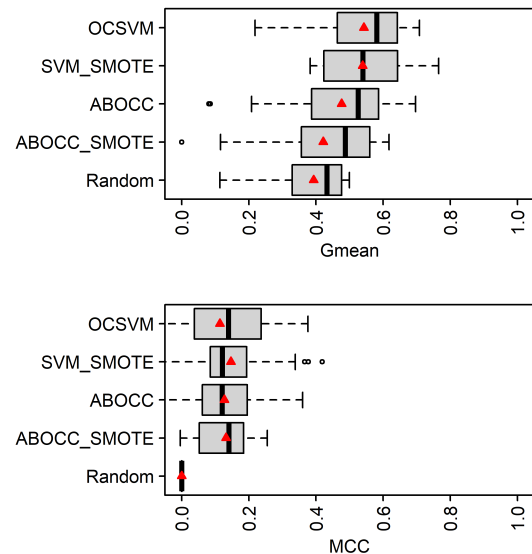


Figure 2. Boxplots of Gmean and MCC.

Coherently with Figure 2, the Wilcoxon Sign Rank tests indicate that i) OCSVM is better than all other methods; ii) SVM and $ABOCC_p$ without SMOTE are equivalent to each

other and both better than $ABOCC_p$ with SMOTE; iii) random classification is worse than all other methods.

It is now interesting to look at the evaluations based on MCC: since MCC computes the effect size of a model with respect to random classification, it indirectly shows how much a method is better or worse than another. In this respect, Figure 2 shows that all models achieve similar improvements over random estimation.

All of the models have relatively low MCC values (means and medians are just above 0.15), which should not be interpreted as indicating bad performance. MCC evaluates models with respect to the random classification, whose performance is not really bad, as Figure 2 shows.

The above results show that none of the methods appears better than the others across the board. Specifically, $ABOCC_p$ appears to be the best of the lot at estimating positive (i.e., defective) modules and the worst at estimating negative (i.e., non-defective) modules. This is a consequence of the “prudential” approach that classifies as faulty the modules that are deemed uncertain by the basic $ABOCC$.

As for overall performance, $ABOCC_p$ achieves Gmean and MCC values that are quite close to those of the other methods.

Thus, we can conclude that $ABOCC_p$ is competitive with respect to other faultiness prediction methods, but choosing the “best” model depends on what is deemed more important by the users of the predictive models: maximizing TPR (hence minimizing false negatives) or maximizing TNR (hence minimizing false positives).

V. THREATS TO VALIDITY

Internal Validity. A possibly relevant threat concerns the correctness of the used data. In this respect, the fact that the J&M dataset has been used by many researchers, shows that it is a trustable (or at least widely trusted) dataset.

Another potential internal validity issue concerns the methods through which the machine learning models were built. This threat was mitigated by using well-established SVM and one-class SVM computation programs, as specified in Section III-B. However, we cannot exclude that different parameterizations of the machine learning (ML) models could lead to different results. Similarly, different balancing techniques and parameters could have been used, so that—depending on the nature of the data and the balancedness of the datasets—different results could have been obtained.

Construct Validity. A potential construct validity issue concerns the way faultiness prediction errors were measured and evaluated. To this end, note that, when dealing with software defect prediction, it is hardly possible to identify a performance metric that satisfies completely all the evaluation needs and that fits equally well in all development processes. We addressed this threat by using two basic indicators, TPR and TNR, which provide a clear view of the performance of each model in identifying faulty and, respectively, non-faulty modules. TPR and TNR do not depend on prevalence ρ either, thus making the provided indication less context-dependent. As overall performance indicators, we used Gmean, the geometric

mean of TPR and TNR, which is also independent of prevalence ρ , and MCC, which depends on ρ .

External Validity. As for the generalizability of our results, the datasets we used collect modules with a wide range of code features (size, complexity, cohesion, coupling, etc.); nonetheless, it is still possible that some software development environments do not match the situation represented by the datasets we used.

VI. RELATED WORK

One-class classification has raised the interest of researchers in several areas, as witnessed by multiple reviews [11]–[13]. Noticeably, it has been studied for anomaly and defect detection in several industrial domains, not just software development.

The One-class SVM has been tested and compared with other ML techniques in multiple occasions and in different contexts. Shin et al. [14] compared it with the multilayer perceptron for fault detection and classification in electro-mechanical machinery. Mahadevan and Shah [15] compared it with Principal Component Analysis (PCA) and Dynamic Principal Component Analysis (DPCA) for fault detection and diagnosis in process data. Hejazi and Singh [16] compared it with the traditional SVM for the detection of fraudulent credit card transactions. In these studies, the One-class model showed better performance than the more traditional approaches. However, in the research by Shin et al., one-class SVM appeared to be sensitive to the parameters and the choice of kernel.

Moussa et al. [17] conducted a study by comparing the one-class SVM with other ML techniques, including SVM, for cross-version and cross-project defect detection. The one-class model outperformed other traditional techniques for cross-version and cross-project detection, but had poor performance for within-project detection. Ciubotariu et al. [18] compared the performance of traditional SVM and one-class SVM. The one-class models had worse FPR than the traditional SVM, but better TPR. Even though the experiment was not conclusive, they suggest that the joint usage of both approaches, in order to compensate each other weaknesses, should be investigated further. Finally, Shehab et al. [1] compared different one-class SVM, Isolation Forest, and k-NN to their traditional counterparts. Their tests showed that, for datasets with medium to high imbalanced ratio, the one-class variants performed better than the traditional models, with and without balancing, using cross and time-sensitive validation approaches.

Ahmad and Bezawada studied the performance of One-Class Classification by Ensembles of Regression model (OC-CER) [19]. This model, originally proposed by Noto et al. [20], consists of multiple regression models, one for each feature of the one-class training set. The test data are labeled as outliers depending on the Outlier Score computed by every regression model. Their test showed that OCCER performs well in relation to other one-class solutions.

VII. CONCLUSIONS AND FUTURE WORK

Ignoring one of the classes during one-class classification leads to a loss of potentially useful information. To exploit

the information embedded in the minority class, though still using one-class estimation, we defined ABOCC, an ensemble method that performs one-class SVM classification twice, first using the majority class, then using the minority class, and finally estimating module defectiveness using both the resulting models. The “prudential” approach ABOCC_p classifies as non-faulty only the modules that are so classified by both models.

ABOCC_p has been tested via an empirical study, in which it has been compared with regular SVM (i.e., with SVM trained using both classes) and one-class SVM. The performances of the considered models were evaluated both globally and considering the ability to detect actual positive and actual negative modules separately. This is important, because in some contexts minimizing false negatives could be more important than minimizing false positives, while in other contexts it could be the opposite. Therefore, in some situations, using the ensemble method ABOCC based on one-class classification can be more advantageous with respect to traditional one class and two-class classification methods.

ACKNOWLEDGMENT

The work reported here was partly supported by Fondo per la Ricerca di Ateneo, Università degli Studi dell’Insubria.

REFERENCES

- [1] M. A. Shehab, W. Khreich, A. Hamou-Lhadj, and I. Sedki, “Commit-time defect prediction using one-class classification,” *Journal of Systems and Software*, vol. 208, p. 111914, 2024.
- [2] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” *ACM trans. on intelligent systems and technology*, vol. 2, no. 3, pp. 1–27, 2011.
- [3] K.-R. Müller, S. Mika, K. Tsuda, and K. Schölkopf, “An introduction to kernel-based learning algorithms,” in *Handbook of neural network signal processing*, CRC Press, 2018, pp. 4–1.
- [4] C.-Y. Yeh and S.-J. Lee, “A kernel-based two-stage one-class support vector machines algorithm,” in *Int. Symp. on Neural Networks*, Springer, 2007, pp. 515–524.
- [5] M. Jureczko and L. Madeyski, “Towards identifying software project clusters with regard to defect prediction,” in *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*, 2010, pp. 1–10.
- [6] Zenodo, <https://zenodo.org/records/6342328>.
- [7] M. Zhang, B. Xu, and J. Gong, “An anomaly detection model based on one-class svm to detect network intrusions,” in *2015 11th Int. Conf. on Mobile Ad-hoc and Sensor Networks (MSN)*, 2015, pp. 102–107.
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [9] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
- [10] L. Lavazza and S. Morasca, “Common problems with the usage of f-measure and accuracy metrics in medical research,” *IEEE Access*, 2023.
- [11] S. S. Khan and M. G. Madden, “One-class classification: Taxonomy of study and review of techniques,” *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 2014.
- [12] S. Alam, S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, “One-class support vector classifiers: A survey,” *Knowledge-Based Systems*, vol. 196, p. 105754, 2020.
- [13] N. Seliya, A. Abdollah Zadeh, and T. M. Khoshgoftaar, “A literature review on one-class classification and its potential applications in big data,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–31, 2021.
- [14] H. J. Shin, D.-H. Eom, and S.-S. Kim, “One-class support vector machines—an application in machine fault detection and classification,” *Computers & Industrial Engineering*, vol. 48, no. 2, pp. 395–408, 2005.
- [15] S. Mahadevan and S. L. Shah, “Fault detection and diagnosis in process data using one-class support vector machines,” *Journal of process control*, vol. 19, no. 10, pp. 1627–1639, 2009.
- [16] M. Hejazi and Y. P. Singh, “One-class support vector machines approach to anomaly detection,” *Applied Artificial Intelligence*, vol. 27, no. 5, pp. 351–366, 2013.
- [17] R. Moussa, D. Azar, and F. Sarro, “Investigating the use of one-class support vector machine for software defect prediction,” *arXiv preprint arXiv:2202.12074*, 2022.
- [18] G. Ciubotariu, G. Czibula, I. G. Czibula, and I.-G. Chelaru, “Uncovering behavioural patterns of one: And binary-class svm-based software defect predictors,” in *Proceedings of the 18th International Conference on Software Technologies (ICSOFTE 2023)*, Rome, Italy: SciTePress, 2023, pp. 5881–588.
- [19] A. Ahmad and S. Bezawada, “One-class classification by ensembles of regression models—a detailed study,” *arXiv:1912.11475*, 2019.
- [20] K. Noto, C. Brodley, and D. Slonim, “Anomaly detection using an ensemble of feature models,” in *2010 IEEE International Conference on Data Mining*, 2010, pp. 953–958.

Process Simulation and AI: A Perfect Match?!

Carlo Simon and Merlin Tankred Philipp Hladik

Hochschule Worms

Erenburgerstr. 19, 67549 Worms, Germany

e-mail: {simon,hladik}@hs-worms.de

Abstract—Since the dramatic improvement of Large Language Models (LLM), academics feel that students increasingly avoid learning programming, and practitioners discuss Artificial Intelligence (AI) as an alternative to process simulation. Both groups overlook the duality of simulation and programming, which are based on causality, in contrast to AI, which is based on probability. This paper develops requirements to successfully integrate both approaches. Though integration efforts have been made in recent years — as a literature analysis of simulation conferences and journals from the last 15 years with the aid of a novel, LLM-based tool indicates — a systematic approach is still largely missing. This paper is the second of three planned papers to overcome this situation. The first paper considered the simulation life-cycle to reveal the expectations of academia and practice, and how AI might help to meet them. This paper uses a survey to understand the potentials and barriers to the application of simulation and AI in practice. The third paper will discuss an AI-integration framework for simulation. Besides the results of the literature analysis and a summary of the survey, the main output of this paper is a research agenda building on these outcomes to clarify and discuss the next steps for integrating process simulation and AI.

Keywords—Process Simulation; AI Integration; Literature Review; Survey on Simulation; Research Agenda.

I. INTRODUCTION

Like any other research discipline, the process simulation community benefits from progress, new methods, and further development. For a long time, conferences and journals on this topic focused on elegant new mathematical methods and challenging applications. Since the dramatic improvement of Large Language Models (LLM) in recent years, Artificial Intelligence (AI) methods seem to receive greater attention. In this paper, AI encompasses two relevant method clusters: (1) LLM-based methods, which process and generate natural language, and (2) data-driven Machine Learning (ML) and Reinforcement Learning (RL) methods, which learn from structured data.

To investigate this observation, this paper starts with a literature analysis of a total of 1916 papers from simulation-related conferences and journals:

- 14 conference volumes of the International Conference on Simulation (IARIA, SIMUL, 2011-2024) [1],
- 30 volumes of the International Journal on Advances in Systems and Measurements (IARIA, IJSysMea, 2010-2025) [2], and
- 16 conference volumes of the ECMS International Conference on Modeling and Simulation (ECMSConf, 2010-2025) [3].

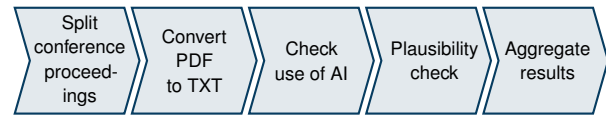


Figure 1. Phases of the literature analysis.

To process this large number of papers, a novel, LLM-based tool was developed. Figure 1 summarizes the analysis pipeline. The proceedings were converted to plain text and each paper was processed (gpt-oss-20b, 20 billion parameters, served via LM Studio with a temperature of 0.1, single workstation). A system prompt instructed the model to decide whether a paper uses AI methods and to return a structured result in JSON format. The prompt can be requested from the authors. The intermediate results were aggregated per venue and on a random basis manually checked for plausibility. A full validation against manual coding remains future work. The resulting counts of AI-related papers are shown in Figures 2 and 3.

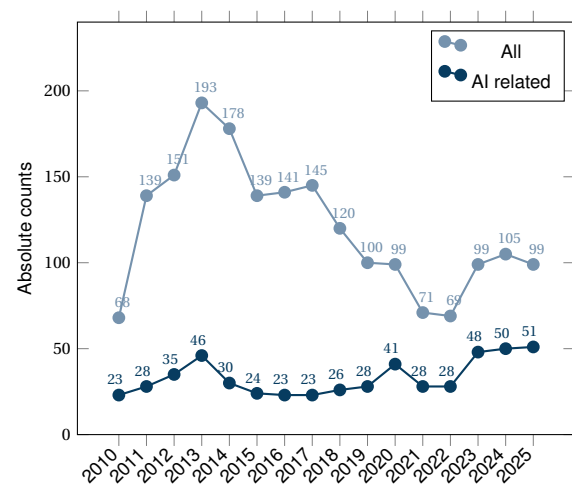


Figure 2. Total number and number of AI-related papers in the mentioned proceedings.

Although this review addresses only a limited number of contributions, at least a tendency can be observed as follows: with the exception of an outlier in 2010, a notable relative increase in AI-related papers can be observed from 2020 onward.

A closer look at these publications shows that specific applications of AI methods are addressed in various phases

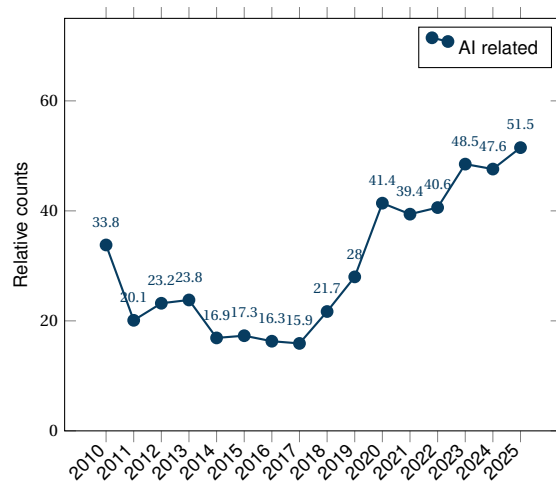


Figure 3. Relative number of AI-related papers in the mentioned proceedings.

of simulation which replicates and analyses an administrative, production, or logistic system in a controlled virtual environment without affecting the real system [4]. What was found is work on optimization algorithms, process mining, AI-assisted modeling, and AI-supported decision-making [5], [6], [7]. A systematic approach that maps AI methods to all phases of a process simulation life-cycle, however, could not be found, although it would help academia and practice to focus on suitable methods.

Before suggestions can be made, it is important to answer the following research questions:

- 1) For what purposes do process manager in administration, production, or logistics use process simulation and AI?
- 2) Which restrictions limit the use of process simulation by these persons, and could AI overcome these restrictions?
- 3) Based on this, what should a future research agenda encompass that integrates process simulation and AI?

This paper presents the results of a survey that delivers answers to the first two questions and uses these answers to further develop a research agenda on the combination of simulation and AI, thereby also answering the third question.

The survey is limited by the small number of participants, who were mostly practitioners, and by its regional focus on Germany.

The remaining of the paper is structured as follows: Section II explains the survey design, and Section III summarizes the main results of the survey. Section IV explains the research agenda, while Section V concludes the paper.

II. SURVEY DESIGN

A survey investigating the opportunities and challenges of applying simulation and AI in practice was conducted in December 2025 and January 2026 for a master's thesis which contains a complete presentation of the results [8].

The survey contains 20 questions that, besides demographics, are organized into the seven thematic blocks shown

TABLE I. STRUCTURE OF THE SURVEY.

Thematic block	Question type
Demographics	single choice
Current use of simulation and AI	single choice
Deployment areas	multiple choice
Purposes and expectations	multiple choice
Barriers and difficulties	multiple choice
AI support in simulation	single / multiple choice
Result preparation and AI functions for understanding results	5-point Likert
Willingness to invest	multiple choice

in Table I. Most questions are single- or multiple-choice, except the perceived usefulness of AI functions to improve the comprehension of simulation results which was rated on a five-point Likert scale.

The mandatory questions on the current use of simulation and AI directed users and non-users to separate follow-up questions concerning their experiences and perceived barriers, ensuring that both groups provided assessments across the same value dimensions. As these questions were optional, the number of responses varied across questions. The reported percentages are based on the number of responses to the respective question.

Participants were recruited through professional contacts on LinkedIn and through a personal network of professionals working in logistics. A total of 70 individuals accessed the survey, of whom 41 completed it and 16 provided at least partial responses. These 57 at-least-partial responses were included in the evaluation.

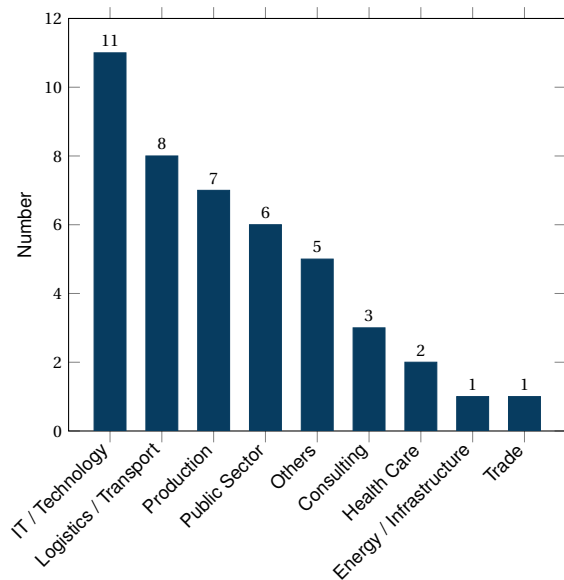


Figure 4. Industry distribution.

The respondents work primarily as process managers, most frequently in project management, analysis and plan-

ning, and consulting, and span nine industries led by IT, logistics, and manufacturing (Figure 4). All responses were collected and analyzed anonymously, but participants could request the aggregated results by email.

Despite the small number of participants and the focus on Germany, the study provides exploratory insights into the potential uses of and also barriers for simulation and AI. The results are presented in the following Section III.

III. SURVEY RESULTS

Figure 5 shows the adoption level of simulation and AI. The responses range from *regularly* to *don't know*.

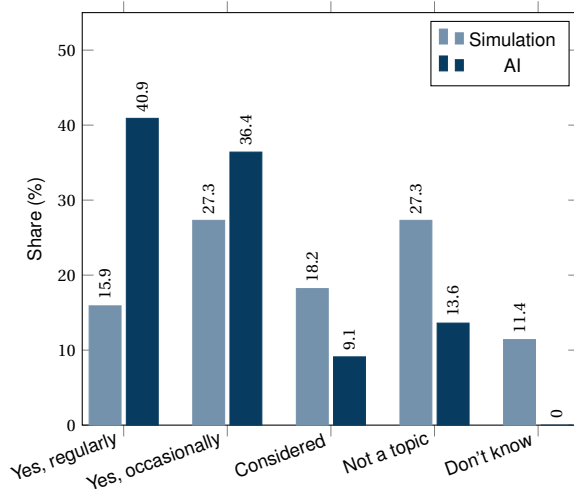


Figure 5. Adoption of simulation and AI.

The deployment areas of simulation and AI are considered next. Figure 6 shows the organizational units that use these techniques. The dominant application fields mentioned for simulation are *IT processes* and *logistics*. AI is mentioned relatively consistently across *IT processes*, *customer services*, *logistics*, and *production*.

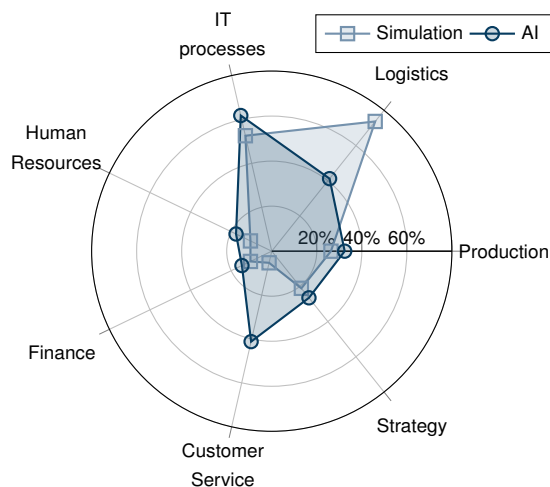


Figure 6. Deployment of simulation vs. AI.

Since a low number of participants was anticipated, no Kano analysis was conducted [9]. Expectations regarding the usefulness of the simulation were, however, surveyed and compared with its actual usage.

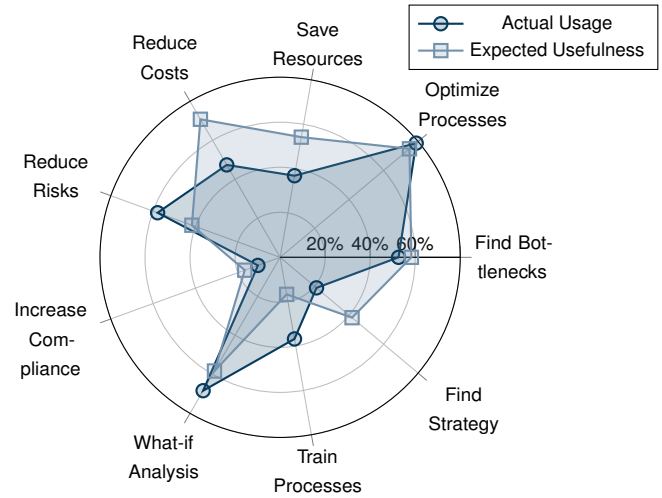


Figure 7. Expected usefulness vs. actual usage of simulations.

Figure 7 shows the answer distribution. It seems that for the respondents, simulation does not fulfill their expectations regarding financially relevant aspects such as *reduce costs*, *save resources*, or *find strategy*.

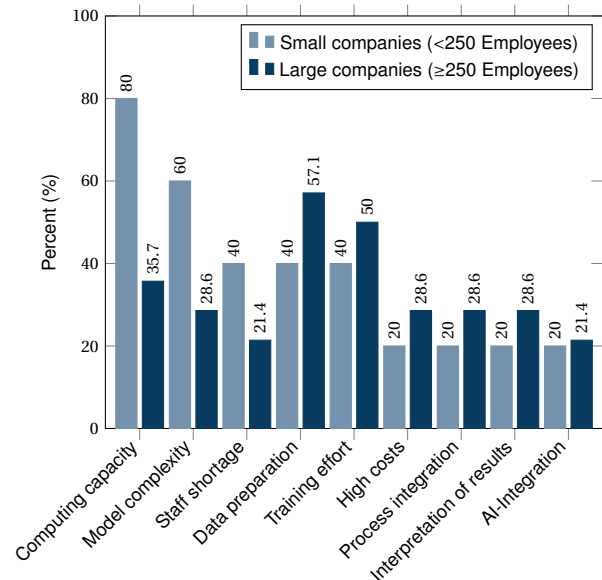


Figure 8. Difficulties in simulation by company size.

Former users of simulation were asked about the difficulties they encountered. Their answers differed according to the respondents' company sizes (see Figure 8). Small companies frequently mentioned the required computing capacity as a reason. This and the mentioned problem of model complexity may indicate that they attempted to simulate large systems, possibly production lines. The survey, however, does not provide evidence for this. Larger

companies reported struggling with the availability of well-prepared data. All companies identified *training effort* and *staff shortage* as key challenges.

Figure 9 shows that responses from large companies indicate deficits in formal methods, which hinder the use of simulation despite its recognized value. This may explain the estimated high costs.

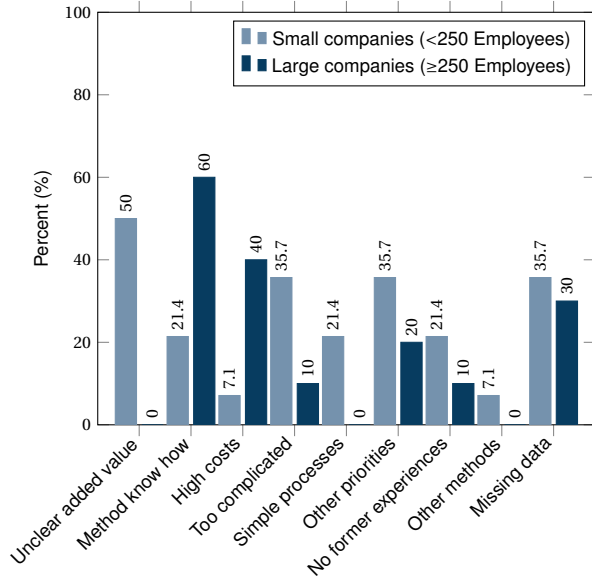


Figure 9. Reasons against simulation usage by company size.

Another major problem is the absence of suitable data. This may be the reason why no alternative approaches (such as linear optimization) are mentioned, since all other approaches also rely on valid data.

Another question considered the possibility to overcome these limitations with AI methods. The answers presented in Figure 10 differ with respect to the participants self-assessed AI competence. Furthermore, they can be classified into (1) AI support to find reasons for simulation and (2) AI support to simplify the simulation process:

- 1) Risk identification as a reason for simulation is primarily cited by respondents familiar with AI. Other reasons, such as *reducing quantities*, *identifying bottlenecks*, or *calculating costs*, are more often mentioned by respondents unfamiliar with AI. LLM-based AI technologies seem to predominate in both applications.
- 2) AI methods to improve the simulation process such as improving *data integration* or *estimating values* which could be achieved with ML based methods like classification or regression are mostly mentioned by participants already familiar with AI.

The aspect *suggest improvements* does not fall into these categories but opens a new field: Could AI also be used to explain the simulation results? The final outcome of the survey presented here is an estimated assessment of the extent to which AI methods, and especially LLMs, could be used for this purpose.

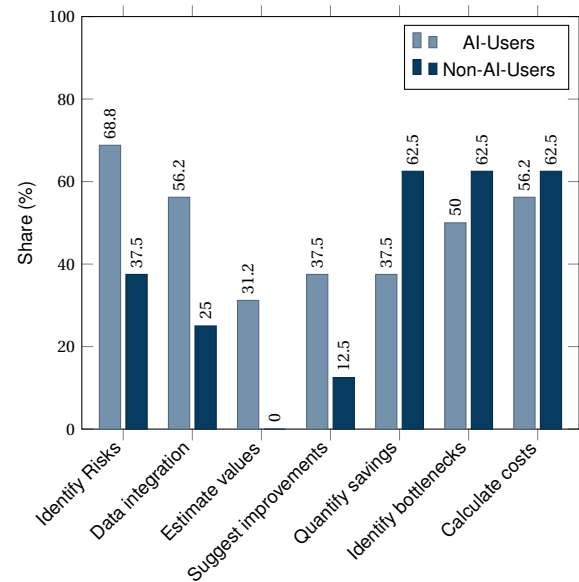


Figure 10. When non-users would simulate, by AI experience.

Figure 11 shows the respondents' expected ability to generate *reports*, *diagrams*, *explanations*, and *summaries*. The lowest potential is seen for interactive chats with AI systems that explain simulation and its results.

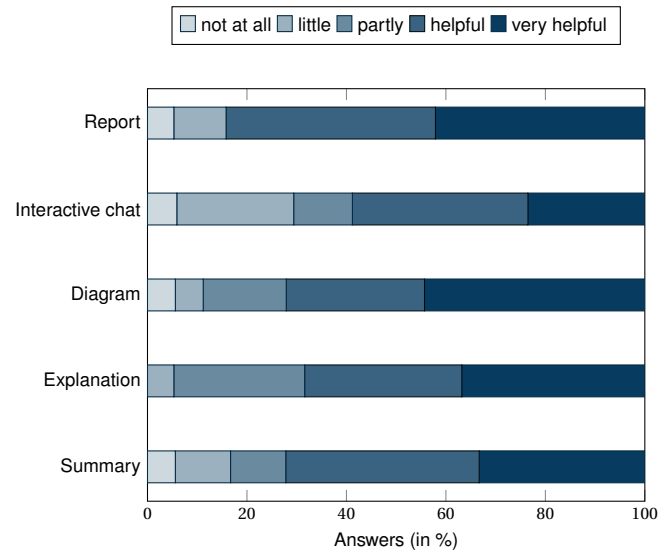


Figure 11. Usefulness of AI to understand simulation results.

The survey results presented here have revealed several barriers to the adoption of simulation. At the same time, AI methods appear to be gaining increasing acceptance and may help to overcome these barriers. Despite the limitations of the survey, it delivers answers to the first two research questions and at least uncovers tendencies. The findings can now be used to develop a structured approach to combining both worlds and potentially integrating the dual nature of both methods (causality vs. correlation). This is done in the following Section IV.

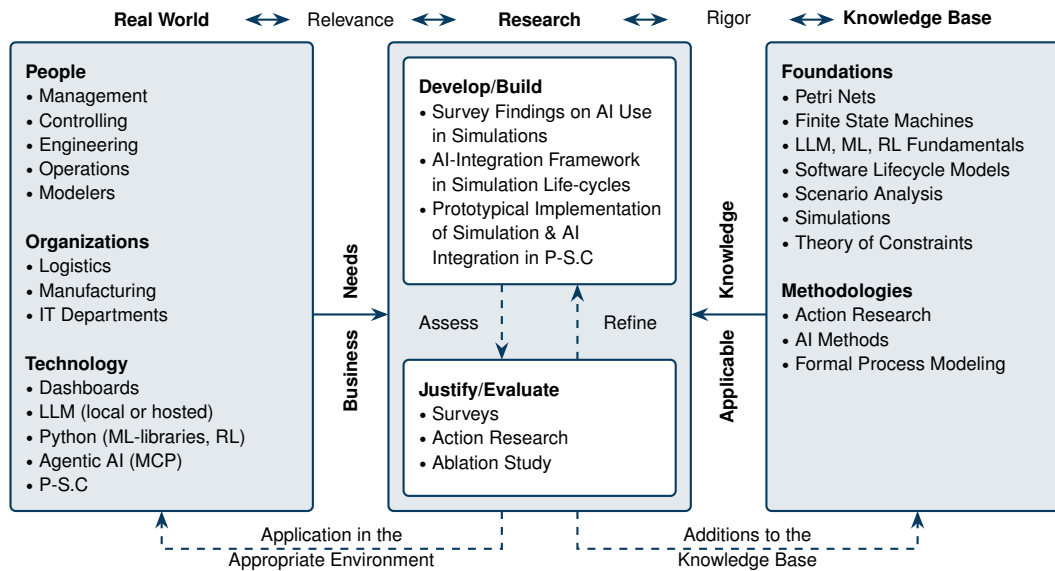


Figure 12. Applied DSR Framework. (based on [10, p. 80])

IV. A RESEARCH AGENDA FOR FUTURE WORK

The duality of simulation, which is based on causality, and AI, which is based on correlation, makes their integration a challenging task. It requires a research agenda to guide this integration in a structured, repeatable, and validated way. This paper is the second of three planned publications in preparation for the actual research.

The research agenda is grounded in Design Science Research (DSR) [10, p.75], which provides a structured approach to the creation and evaluation of IT artifacts. Figure 12 shows the adaptation of the framework, addressing three areas: *real-world environment*, *research activities*, and *knowledge base*. A first research agenda is formulated in [4].

The survey sharpens two components of this agenda:

- The organizational context should be set to IT, logistics, and manufacturing, as they emerge as the dominant deployment areas of the survey's participants (see Figure 6).
- The role of ML in data preparation and value estimation raises questions regarding its importance, as respondents familiar with AI cited it most frequently (see Figure 10). This opens new possibilities for simulation: instead of being conducted for a single scenario it can be conducted for a set of scenarios most likely for the future. Simulation can then help to find resilient strategies that remain stable in the event of unplanned changes.

To stay focused, the AI-method cluster mentioned in *Foundations* in Figure 12 is narrowed from a broad list to the two method families actually reflected in the survey responses: ML/RL for data and estimation, LLM-based methods for explanation.

Moreover, concrete requirements for a prototypical implementation of simulation & AI integration can be formulated. An integrated simulation environment must provide mecha-

nisms to explain and summarize the simulation results with the aid of AI components (see Figure 11).

The aim is to test the integration within the Process Simulation Center (P-S.C). This platform utilizes Petri nets to model and simulate system behavior in order to assess the implications of strategic decisions on processes in administration, production, and logistics [11]. This environment enables an integration from the other direction: during process simulation, the simulated system can interact with AI components. This technique is called Agentic AI and uses the Model Context Protocol (MCP) as an API [12].

The scientific contribution comes from combining causal models with probabilistic ones: embedding AI methods into formal process structures and vice versa to keep correlation-based components grounded in process semantics. This bidirectional integration is the open research question the agenda is designed to address.

V. CONCLUSION AND FUTURE WORK

A literature analysis of more than 1900 papers from simulation-related conferences and journals showed a growing number of studies combining simulation and AI and motivated a survey on this topic. Its results must be interpreted as exploratory indications rather than representative evidence, because of the small number of participants locally restricted to Germany. Nonetheless, the results can at least be interpreted as tendencies and hence as initial answers to the first and second research questions. They are therefore used as suggestions for completing a formally defined research agenda described in Section IV.

The next research step is a third paper. Based on a further literature analysis, existing AI solutions and implementations are mapped to the phases of a V-Model for process simulation. It is expected that further requirements can be derived

for the integration of simulation and AI in a modeling and simulation tool.

After completion of this third preparatory step, the actual research as described in the research agenda of Figure 12 can be conducted. According to current research status, the actual research is estimated to last about three years.

REFERENCES

- [1] SIMUL, “International conference on advances in system simulation”, 2026, [retrieved: May, 2026]. [Online]. Available: <https://www.thinkmind.org/library/SIMUL>
- [2] IARIA, “International journal on advances in systems and measurements”, 2026, [retrieved: May, 2026]. [Online]. Available: https://www.iariajournals.org/systems_and_measurements/index.html
- [3] ECMS, “Digital library of the european council for modelling and simulation”, 2026, [retrieved: May, 2026]. [Online]. Available: <https://www.scs-europe.net/dlib/dl-index.htm>
- [4] C. Simon and M. Hladik, “A research agenda to improve process simulation life-cycles for academia and practice”, in *40th International ECMS Conference on Modelling and Simulation*, Grimstad (Norway), 2026, pp. 519–525.
- [5] T. Kampik et al., “Large process models: A vision for business process management in the age of generative AI”, *KI - Künstliche Intelligenz*, 2024. DOI: 10.1007/s13218-024-00863-8
- [6] H. Kourani, A. Berti, D. Schuster, and W. M. P. van der Aalst, “Process modeling with large language models”, in *Enterprise, Business-Process and Information Systems Modeling (BPMDS 2024)*, ser. Lecture Notes in Business Information Processing, vol. 511, Springer, 2024, pp. 229–244. DOI: 10.1007/978-3-031-61007-3_18
- [7] S. Razavi, B. A. Tolson, and D. H. Burn, “Review of surrogate modeling in water resources”, *Water Resources Research*, vol. 48, no. 7, 2012.
- [8] M. Hladik, “Künstliche Intelligenz in Prozesssimulationen”, German, transl. Artificial Intelligence in Process Simulation, M.S. thesis, Hochschule Worms, 2026.
- [9] N. Kano, N. Seraku, F. Takahashi, and S. Tsuji, “Attractive quality and must-be quality”, *Journal of the Japanese Society for Quality Control*, vol. 14, no. 2, pp. 147–156, 1984.
- [10] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research”, in *Management Information Systems Quarterly*, vol. 28, 2004, pp. 75–106.
- [11] C. Simon, M. Hladik, and S. Haag, “Process Management Compass and Competence Ladder to Empower Process Management Skills”, in *39th International ECMS Conference on Modelling and Simulation*, Catania (Italy), 2025, pp. 486–492.
- [12] Model Context Protocol a Series of LF Projects, LLC., “Model context protocol specification”, 2025, [retrieved: May, 2026]. [Online]. Available: <https://modelcontextprotocol.io/specification/2025-11-25>

Assessing the Accessibility of Public Healthcare Web Services in Spain

Verónica Alvarez-Lopez

Sub-Dir. of Information Systems and Tech. Infrastructures
SESPA
Oviedo, Spain
email: veronica.alvarezllo@sespa.es

Daniel Fernandez-Lanvin

Department of Computing
University of Oviedo
Oviedo, Spain
email: dflanvin@uniovi.es

Abstract—This paper presents a large-scale automated accessibility evaluation of public healthcare web services in Spain. The Spanish healthcare system is decentralized across autonomous regions, which has resulted in the development of multiple digital platforms for citizen interaction. We analyze the impact of such decentralization on the quality of the services provided to healthcare users in critical services like primary care. In this work, we compiled a dataset of 57 official healthcare web services covering all autonomous communities and the cities of Ceuta and Melilla. Automated audits were performed using Google Lighthouse to collect accessibility, performance, and structural metrics. The analysis revealed important variability across both territorial healthcare platforms and service categories. Informational healthcare pages generally obtained the highest accessibility scores, whereas transactional systems exhibited greater variability and lower average accessibility performance.

Keywords—Web Accessibility; eHealth; Public Healthcare Services; Automated Accessibility Evaluation; eGovernment; Healthcare Portals.

I. INTRODUCTION

Digital public services have become a central component of modern eGovernment infrastructures. Healthcare systems rely heavily on online platforms to provide citizens with essential services, such as appointment booking, access to medical information, and interaction with healthcare providers. Ensuring that these services are accessible to all citizens is therefore a critical requirement for public administration. In Spain, the public healthcare system is decentralized across autonomous communities, each responsible for developing and maintaining its own digital health platforms. This organizational structure has resulted in a heterogeneous ecosystem of healthcare web services with different technological architectures and design approaches.

The goal of this study is to provide a large-scale automated evaluation of accessibility across Spanish public healthcare web services to obtain a broad overview of the accessibility status of these critical digital healthcare services. By analyzing a dataset of 57 official services covering all Spanish territories, we aim to identify general accessibility patterns, differences between service categories, and potential relationships between structural complexity and accessibility outcomes.

The remainder of this paper is organized as follows. Section II presents related work, Section III describes the methodology, Section IV presents the experimental setup,

Section V discusses the results, and Section VI concludes the paper and outlines future work.

II. RELATED WORK

Web accessibility has been extensively studied in the context of e-government systems, where public-sector websites frequently present barriers affecting navigation, form interaction, and compliance with accessibility standards. Systematic reviews of accessibility evaluation studies have shown that government portals are commonly assessed through a combination of automated tools and standards-based approaches based on the Web Content Accessibility Guidelines (WCAG) [1]. However, these reviews also highlight the methodological heterogeneity of existing studies, which often complicates cross-country comparisons of public digital services [2].

In the healthcare domain, accessibility is particularly critical because digital barriers may directly affect citizens' access to essential services. Previous studies evaluating eHealth portals have identified recurring accessibility problems in health-related websites across Europe, especially in systems that include critical interactive functionalities such as appointment booking or patient portals [3][4]. More recent literature reviews also emphasize that accessibility is still inconsistently incorporated throughout the design and evaluation lifecycle of digital health systems, despite the widespread adoption of WCAG as a reference framework [5]. From a methodological perspective, automated accessibility auditing tools such as Lighthouse [6], WAVE [7], and Axe [8] are frequently used for large-scale evaluations of web services. These tools enable scalable benchmarking of accessibility across large sets of websites, although they are typically considered complementary to manual inspection and user-based evaluation [9]. Despite the growing number of studies on web accessibility in both e-government and eHealth, there is still limited comparative evidence focusing on national public healthcare ecosystems as a whole.

In this context, the present work provides a large-scale automated audit of accessibility across Spanish public healthcare web services, covering all autonomous communities and the autonomous cities of Ceuta and Melilla. By systematically evaluating a dataset of official digital health services, this study aims to provide a comparative overview of accessibility across the Spanish public healthcare digital ecosystem.

III. DATASET

To conduct the accessibility evaluation, we compiled a dataset of official public healthcare web services belonging to the Spanish National Health System. Spain's healthcare system is decentralized across autonomous communities, each responsible for the management and operation of its regional healthcare services. As a consequence, each region has developed its own digital platforms to provide citizens with online healthcare services. The dataset includes 57 official web services covering all 17 Spanish autonomous communities as well as the autonomous cities of Ceuta and Melilla. The selection focused on citizen-facing services that represent the main interaction points between users and public healthcare systems. To ensure a representative sample of the digital health ecosystem, the dataset includes three main categories of services:

- Appointment booking systems, which allow citizens to schedule medical consultations.
- Patient portals, which provide access to personal health information and administrative services.
- Informational healthcare pages, which provide guidance or access points to digital health services.

These services were identified through the official websites of the regional healthcare services and the corresponding government portals. Only publicly accessible services were included in the dataset, ensuring that all evaluated pages could be accessed without authentication barriers. The final dataset therefore represents a broad overview of the Spanish public healthcare digital ecosystem, capturing both transactional services and informational platforms used by citizens across the country.

IV. METHODOLOGY

To evaluate the accessibility of the selected web services, we performed an automated accessibility audit using Google Lighthouse, an open-source auditing tool widely used for assessing web quality and accessibility. Lighthouse evaluates web pages according to several technical metrics, including accessibility, performance, and structural characteristics of the web interface. For each web service included in the dataset, an automated audit was executed using a headless browser environment. The evaluation process consisted of loading each web page and extracting the accessibility and performance metrics generated by Lighthouse. The following metrics were collected for each evaluated service:

- Accessibility score, which provides an aggregated indicator of accessibility compliance based on automated WCAG-related checks.
- Document Object Model (DOM) size, representing the structural complexity of the web page.

These metrics provide a quantitative overview of both accessibility compliance and technical characteristics of the evaluated services. The automated audit process was executed using a scripted pipeline that sequentially evaluated all URLs in the dataset and stored the resulting metrics in a structured dataset. This approach allows for

reproducibility and facilitates large-scale comparative analysis across services and territories.

After collecting the metrics, descriptive statistical analysis was performed to examine accessibility levels across the dataset. Additional analyses were conducted to explore differences between service categories and territories, as well as potential relationships between structural complexity and accessibility outcomes.

V. RESULTS

This section presents the results of the automated accessibility evaluation conducted on the dataset of Spanish public healthcare web services. We first provide an overview of the global accessibility levels observed in the dataset, followed by a comparative analysis across service categories and territories. Finally, we examine potential relationships between structural complexity and accessibility outcomes.

A. Overall Accessibility Levels

The accessibility evaluation reveals generally high accessibility levels across the healthcare services analyzed. The mean accessibility score across the dataset is approximately 0.90, with values ranging between 0.66 and 1.00. Most websites achieve scores above 0.85, indicating relatively good compliance with automated accessibility checks. However, the presence of lower scores suggests that accessibility issues still persist in some services. To better understand these results, we analyzed the most frequently detected accessibility violations reported by the automated evaluation tool (Table 1). The most common issues identified across the evaluated services include:

- Insufficient color contrast between text and background, which may reduce readability for users with visual impairments.
- Missing alternative text for images, preventing screen readers from properly conveying visual information.
- Form elements without associated labels, making it difficult for assistive technologies to interpret input fields.
- Buttons or links without accessible names, which can hinder navigation for screen reader users.
- Improper heading hierarchy, affecting the structural navigation of pages through assistive technologies.

Among these issues, insufficient color contrast and missing form labels appear as the most frequent violations across the evaluated services. These problems are particularly common in interactive healthcare services such as appointment booking systems, where complex forms and dynamic interface elements are widely used.

Overall, the distribution of accessibility scores suggests that while many healthcare services achieve relatively high automated accessibility scores, recurrent violations related to contrast, form accessibility, and semantic structure still affect a subset of services.

TABLE 1. MOST FREQUENT ACCESSIBILITY VIOLATIONS DETECTED ACROSS EVALUATED SERVICES

Accessibility issue	WCAG reference	Description	Frequency (services affected)
Insufficient color contrast	WCAG 1.4.3	Text elements do not provide sufficient contrast against their background	32
Form elements without associated labels	WCAG 1.3.1 / 4.1.2	Input fields are missing programmatically associated labels	28
Links without accessible names	WCAG 2.4.4 / 4.1.2	Anchor elements lack descriptive accessible text	24
Buttons without accessible names	WCAG 4.1.2	Interactive buttons are not properly labeled for assistive technologies	21
Images without alternative text	WCAG 1.1.1	Images lack descriptive alt attributes	19
Improper heading hierarchy	WCAG 1.3.1	Headings are not used in a consistent hierarchical order	17
Missing form input purpose	WCAG 1.3.5	Form fields do not specify autocomplete attributes	15
ARIA roles used incorrectly	WCAG 4.1.2	Incorrect or conflicting ARIA attributes detected	13
Low contrast in UI components	WCAG 1.4.11	UI controls or icons do not meet contrast requirements	12
Missing document language	WCAG 3.1.1	The HTML document does not declare the primary language	9

Table 1 summarizes the most frequent accessibility violations detected during the automated evaluation. The results indicate that contrast-related problems and form accessibility issues are the most common barriers across the evaluated healthcare services. In particular, insufficient color contrast and missing form labels appear in a substantial proportion of services. These problems are especially prevalent in interactive systems such as appointment booking platforms and patient portals, where complex forms and dynamic interface components are widely used.

B. Accessibility by Service Category

To complement the territorial analysis, the evaluated healthcare services were grouped into three major functional categories frequently identified across the analyzed regional healthcare ecosystems: appointment booking systems (“*cita previa*”), patient portals (“*portal paciente*”), and informational/service pages. The classification was derived from the functional taxonomy of the analyzed public healthcare services.

Informational and service-oriented pages achieved the highest overall accessibility performance, with a mean Lighthouse accessibility score of 0.913 (Standard Deviation

(SD) = 0.065, $n = 7$). Patient portals obtained an average score of 0.885 (SD = 0.095, $n = 18$), while appointment booking systems showed the lowest average accessibility score, with a mean value of 0.859 (SD = 0.088, $n = 18$). These results suggest that transactional healthcare services, particularly appointment booking systems, tend to present greater accessibility challenges than informational portals. This trend may be related to the higher interaction complexity and the presence of dynamic interface elements in appointment-management workflows. By contrast, informational pages generally contain more static content and simpler navigation structures, which may facilitate compliance with accessibility guidelines. Overall, these results suggest that the complexity of user interaction plays an important role in accessibility outcomes, particularly in services that require users to complete forms or navigate dynamic interfaces.

C. Accessibility Across Territories

Because the Spanish healthcare system is organized at the regional level, it is also relevant to examine whether accessibility levels differ between territories.

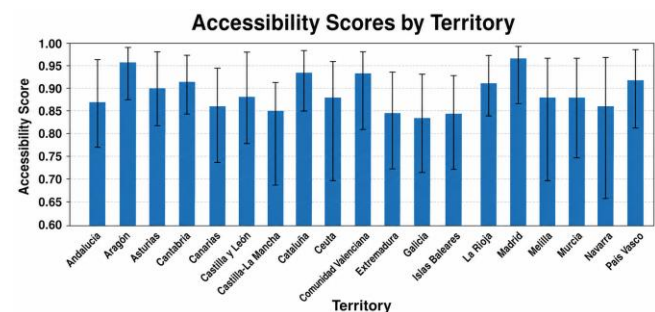


Figure 1. Average accessibility scores across territories.

Figure 1 presents the mean accessibility score for the healthcare web services associated with each autonomous community and autonomous city. Although most territories achieve relatively high accessibility levels, some regional healthcare platforms exhibit lower scores than others. These differences likely reflect variations in technological platforms, development practices, and accessibility policies adopted by regional healthcare administrations. Given the decentralized structure of the Spanish healthcare system, each autonomous community independently develops and maintains its digital healthcare services. As a result, accessibility outcomes may depend on regional design standards, technological frameworks, and development resources. Despite these differences, the results show that most territories achieved mean accessibility scores above 0.80 across the evaluated services, suggesting a generally acceptable level of automated accessibility compliance across the national healthcare ecosystem.

D. Structural Complexity and Accessibility

To explore the relationship between structural complexity and accessibility outcomes, we analyzed the size of the DOM for each evaluated service and compared it with the corresponding accessibility score obtained from the

automated Lighthouse audit. Across the evaluated dataset, DOM sizes showed substantial variability between services, ranging from relatively small informational pages to highly interactive transactional healthcare systems. Appointment booking services and patient portals generally exhibited larger DOM structures due to the presence of dynamic components, forms, authentication mechanisms, and multi-step interaction workflows. To statistically examine this relationship, a Spearman correlation analysis was conducted between DOM size and accessibility scores across the evaluated services ($n = 57$). The analysis revealed a statistically significant moderate positive correlation ($\rho = 0.54$, $p < 0.001$). Although larger DOM structures are commonly associated with increased interface complexity, the observed results suggest that structurally larger services in this dataset often incorporated more mature interface implementations, richer semantic organization, and more accessibility-oriented design practices. In contrast, some smaller services still exhibited important accessibility limitations despite their lower structural complexity. These findings indicate that DOM size alone is not a reliable predictor of accessibility quality. Instead, accessibility outcomes appear to depend more strongly on implementation practices, semantic correctness, and interaction design decisions than on structural size itself.

VI. CONCLUSIONS AND FUTURE WORK

This paper presented a large-scale automated evaluation of accessibility in Spanish public healthcare web services. By compiling a dataset of 57 official healthcare services covering all Spanish autonomous communities and the autonomous cities of Ceuta and Melilla, we conducted a comparative analysis of accessibility outcomes using automated Lighthouse audits. The results show that, overall, accessibility levels across the Spanish public healthcare digital ecosystem are relatively high, with a mean accessibility score close to 0.90. Most evaluated services achieved scores above 0.85, suggesting a generally good level of compliance with automated accessibility checks. However, the analysis also revealed important variability across both territorial healthcare platforms and service categories. Informational healthcare pages generally obtained the highest accessibility scores, whereas transactional systems such as appointment booking services and patient portals exhibited greater variability and lower average accessibility performance. These findings suggest that interaction-intensive services continue to pose greater accessibility challenges due to the presence of dynamic interfaces, forms, authentication workflows, and interactive components. The study additionally identified recurrent accessibility violations affecting multiple platforms, particularly insufficient color contrast, missing form labels, and interactive elements lacking accessible names. These problems were especially frequent in services involving user interaction and multi-step workflows, reinforcing the importance of accessibility-aware implementation practices in transactional healthcare systems. A correlation analysis

between DOM size and accessibility scores revealed a statistically significant moderate positive correlation ($\rho = 0.54$, $p < 0.001$). These results suggest that larger healthcare services in the dataset often incorporated more accessibility-oriented development practices, indicating that DOM size alone is not a reliable predictor of accessibility quality. From a broader perspective, the results highlight the challenges of ensuring accessibility in decentralized public digital ecosystems, where multiple regional administrations independently develop and maintain healthcare services. While the overall accessibility levels observed in this study are encouraging, the persistence of recurrent violations indicates that accessibility practices remain inconsistently implemented across platforms. Future work should extend the analysis beyond automated auditing by incorporating manual expert evaluation, user-centered accessibility testing involving people with disabilities, and longitudinal monitoring of accessibility evolution over time. Additional studies could also explore the relationship between accessibility, usability, service complexity, and user interaction patterns in public healthcare digital services.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Innovation and Universities (project internal code MCIU/AEI/10.13039/501100011033) and co-funded by the European Regional Development Fund (ERDF) under project PID2024-155586OB-I00. Additional support was provided by the Government of the Principality of Asturias through grant GRU-GIC-24-070.

REFERENCES

- [1] W3C, *Web Content Accessibility Guidelines (WCAG) 2.2*, W3C Recommendation, Oct. 2023. [Online]. Available: <https://www.w3.org/TR/WCAG22/>.
- [2] D. Cisneros, F. Huamán Monzón, and F. Paz, "Accessibility Evaluation of E-Government Web Applications: A Systematic Review," in *Lecture Notes in Computer Science*, vol. 12780, Human-Computer Interaction – HCII 2021, Cham, Switzerland: Springer, 2021, pp. 210–223.
- [3] C. Sik-Lányi and E. Orbán-Mihálykó, "Accessibility Testing of European Health-Related Websites," *Arabian Journal for Science and Engineering*, vol. 44, pp. 9175–9189, 2019.
- [4] J. Martins, J. Gonçalves, and M. Branco, "A full scope web accessibility evaluation procedure proposal based on Iberian eHealth websites," *Computers in Human Behavior*, vol. 73, pp. 676–684, 2017.
- [5] M. Jonsson, S. Johansson, D. Hussain, J. Gulliksen and C. Gustavsson, "Development and Evaluation of eHealth Services for People With Impairments: Scoping Literature Review" *Journal of Medical Internet Research*, vol. 25, pp. 1–16, 2023.
- [6] Google Chrome Developers, Lighthouse. [Online]. Available: <https://developer.chrome.com/docs/lighthouse/>.
- [7] WebAIM, WAVE Web Accessibility Evaluation Tools. [Online]. Available: <https://wave.webaim.org/>.
- [8] Deque Systems, axe DevTools. [Online]. Available: <https://www.deque.com/axe/>.
- [9] M. Vigo, J. Brown, and V. Conway, "Benchmarking Web Accessibility Evaluation Tools" *ACM Transactions on the Web*, vol. 7, no. 4, pp 1-10. 2013.

From Relational and Property Graph Data to Large Language Models

Malcolm Crowe

University of the West of Scotland (retired)
Paisley, UK
e-mail: Malcolm.Crowe@uws.ac.uk

Fritz Laux

Reutlingen University (retired)
Reutlingen, Germany
e-mail: Friedrich.Laux@reutlingen.de

Abstract — A data management server that combines the relational data model found in traditional databases with the graphical models being developed by newer data products can provide its knowledge model to generative tools for large language models. Both models have been reconciled at the implementation level by replacing foreign keys by reference values. This boosts performance and allows effectively following links instead of lookup joins. The reference value will never change as it uses the record address in the log file. This paper reports on progress in providing such a synthesis and includes an example of a knowledge base that is triple-based and would also be supported in such a system.

Keywords—database standatrimple-based graphical database system; data model; knowledge graph.

I. INTRODUCTION

It has been established by experiment [1] that the best way of prompting Artificial Intelligence (AI) generative models is by giving them the vocabulary appropriate to a problem scenario, and the best way of communicating it is as a knowledge graph using serialization format, such as JavaScript Object Notation (JSON). Unfortunately, constructing any sort of knowledge graph for a real problem area is itself a complex and time-consuming task. A good way is to begin with whatever structured data is available. If it is already in a database, get the database to generate JSON from it. There are many products to assist in the process, but they all seem to have complex rules. They should be made easier to use, and this presentation aims to walk through a realistic case study.

This paper is part of an open-source research project [2] to support the approaches of both ISO/IEC 9075 Database Language SQL (Structured Query Language) and ISO/IEC 39075 Database Language GQL (Graph Query Language) in a single stand-alone database server. Section II outlines the current state of the art. The key contribution of this paper is outlined in Section III, and Section IV describes a case study showing some of the new approaches to data thus enabled. Some important highlights of the implementation are detailed in Section V. Section VI provides some performance measurements, and Section VII details some conclusions and planned future work. This will include contributions towards the emerging technologies for using large language models.

II. RELATED WORK

SQL/PGQ [3], [4], [5] and Neo4j [6] offer standalone server-based implementations of graph databases, but neither was designed to provide the advanced semantics envisaged in the GQL standard, such as MATCH over repeating graph

patterns and horizontal aggregations. The review of the field offered by Kondylakis et al. [7] identified current challenges in the field: (a) Support for schema constraints remains limited and fragmented in Graph Database Management Systems (GDBMSs) and applications, hindering data integration, query optimization and Machine Learning (ML) feature extraction; (b) Data integration is particularly difficult with Resource Description Framework (RDF) graphs and requires either mapping both models to a common unifying representation, while direct transformations between the two have difficulty in ensuring that RDF triples are accurately mapped to the nodes, edges, and properties of property graphs and vice versa; (c) Normalization and standard forms are fundamentally under-researched (but see [8]); (d) GQL was designed to allow massively parallel execution, but graph query answering is relatively slow in current implementations. It has been observed [9], that the only mechanism in SQL for linking data is foreign keys, but SQL does not make foreign keys usable in queries.

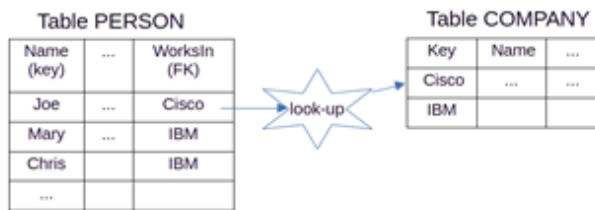
Meantime, almost all graphical databases are built by chaining together transformations provided by different products. The standardization process continues however, and new editions of the international standards are expected in 2027.

III. THE RESEARCH CONTRIBUTION

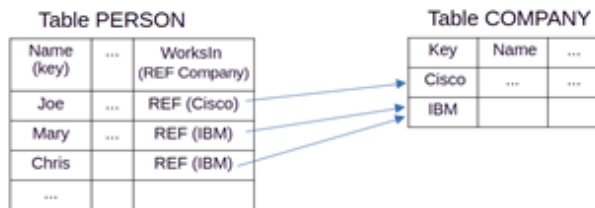
As mentioned above, SQL uses key values to enable tabular data to be joined, while edges in GQL join individual nodes. In previous work [10], we have shown that for record-based data the concept of reference values can support both approaches, so that a unified approach to linked data management should use structural reference values instead of keys. A bonus is that a relational system holding such data can then support a range of alternative graph models. That work outlined an ongoing implementation of a SQL server that supports GQL syntax, graph insertion, and graph queries. The main contribution of our paper is to explain the implementation of the hybrid graph-relational database Pyrrho V8. Some details of this design are summarized in Section IV below.

In a record-based Database Management System (DBMS), an edge type record contains one or more references, e.g., (1 reference) a record might identify a company (e.g., in a Person record, WorksIn: Cisco), (2 references) a record of employment of a person in a company (a WorksIn record, Person: Joe, Company: Cisco), and so on. The edge type can contain non-reference properties (e.g., dates). Figure 1 shows a simple example.

a) Relations with Foreign Key (FK)



b) Relations with reference value (REF)



c) Property graph instance and model



Figure 1. REF and foreign key values

Figure 1. REF and foreign key values.

In the (Person worksIn company) example, the question is – what sort of record is this? It could be an entry in a list of contacts or job applications (Fred, current employer Cisco, born in Scotland..). Do these values identify records in tables of persons, companies, countries? If no, then these properties are not actual references. If yes, then the reference identifies a particular (current, etc.) company for the person, etc.

Given a database with such records, we can define graph models by ignoring some node and edge types, ignoring particular properties, and/or identifying two references FROM and TO in the edge types we retain. In graph models, we use labels for different sorts of information, typically preceded by a colon, and property lists using braces, such as {product:'Fuel', supplier:'Exxon'}.

There are numerous graph database products that allow the definition of node and edge types: several products support graph construction using “Ascii-art”, graph patterns where nodes are indicated by parentheses and edges by square brackets: e.g.,

```
INSERT (Joe)-[WorksIn]->(Cisco)
```

The goal of this research is to handle all three cases of Figure 1 using reference columns in relational tables. Section V below outlines how a foreign key as in Figure 1(a) can be implemented using reference columns. It is fairly clear that an edge type as Figure 1(c) implies references to the source and destination nodes. In the next section we explore the

third case, corresponding to Figure 1(b), where the reference values are bound during insertion of a graph pattern.

IV. CASE STUDY [11]

One way of modeling data is design by example: and constructing graph patterns for a new database works this way. Let us look at a complex data model, taking the ACM

AR. Architecture and Organization (0 Core-Tier1 hours, 16 Core-Tier2 hours)

	Core-Tier1 hours	Core-Tier2 Hours	Includes Elective
AR/Digital Logic and Digital Systems		3	N
AR/Machine Level Representation of Data		3	N
AR/Assembly Level Machine Organization		6	N
AR/Memory System Organization and Architecture		3	N
AR/Interfacing and Communication		1	N
AR/Functional Organization			Y
AR/Multiprocessing and Alternative Architectures			Y
AR/Performance Enhancements			Y

AR/Digital Logic and Digital Systems

[3 Core-Tier2 hours]

Topics:

- Overview and history of computer architecture
- Combinational vs. sequential logic/Field programmable gate arrays as a fundamental combinational + sequential logic building block
- Multiple representations/layers of interpretation (hardware is just another layer)
- Computer-aided design tools that process hardware and architectural representations
- Register transfer notation/Hardware Description Language (Verilog/VHDL)
- Physical constraints (gate delays, fan-in, fan-out, energy/power)

Learning outcomes:

1. Describe the progression of computer technology components from vacuum tubes to VLSI, from mainframe computer architectures to the organization of warehouse-scale computers. [Familiarity]
2. Comprehend the trend of modern computer architectures towards multi-core and that parallelism is inherent in all hardware systems. [Familiarity]

Figure 2: An excerpt from the CS2013 document [3]

Figure 2. An excerpt from CS2013 document [3].

CS2013 syllabus report [11] as an example. It has 18 Knowledge Areas with dozens of subject areas (called Knowledge Units): each unit has Topics and Learning Outcomes, categorized as Core-Tier1, Core-Tier2, and Elective. Let us assume we can add a number of similar items by nesting in curly brackets. We suddenly find we are quite naturally getting a new sort of model, that is neither relational nor the usual kind of graph.

However, the nature of the data as records is quite clear: there can be a new or modified knowledge area, a new knowledge unit, new or modified level measures or degree programs, or new topics or learning outcomes in a given knowledge unit.

The university probably wants to arrange these into degree programs with modules and years of study. How to get started? We can simply start adding the data we have. If we start with Appendix A of the report, we can simply introduce some bindings for the Categories and Learning Levels and just start pasting in some of the details we see. As we do this, our DBMS can infer the node and edge types we are defining. As above, property lists are enclosed in braces, and nodes are enclosed in parentheses:

```

Insert (c1:Category{name:'Core-Tier1'}),
(c2:Category{name:'Core-Tier2'}),
(e1:Category{name:'Elective'}),
(fa:Level{name:'Familiarity'}),
("as":Level{name:'Assessment'}),
(us:Level{name:'Usage'}),
(al:KnowledgeArea{name:'AL'})<-{
(:KnowledgeUnit{name:'AL/Basic Analysis'})<-{
(:Topic{detail:'Differences among best, expected, and
worst case behaviors of an algorithm',category:c1}),
(:Topic{detail:'Asymptotic analysis of upper and
expected complexity bounds',category:c1}),
(:Topic{detail:'Big O notation: formal
definition',category:c1}),
(:Topic{detail:'Complexity classes, such as constant,
logarithmic, linear, quadratic, and
exponential',category:c1}),
(:Topic{detail:'Empirical measurements of
performance',category:c1}),
(:Topic{detail:'Time and space trade-offs in
algorithms',category:c1}),
(:Topic{detail:'Big O notation: use',category:c2}),
(:Topic{detail:'Little o, big omega and big theta
notation',category:c2}),
(:Topic{detail:'Recurrence relations',category:c2}),
(:Topic{detail:'Analysis of iterative and recursive
algorithms',category:c2}),
(:Topic{detail:'Some version of a Master
Theorem',category:c2}),
(:LearningOutcome{detail:'Explain what is meant by
"best", "expected", and "worst" case behavior of an
algorithm.',category:c1,Level:fa}),
(:LearningOutcome{detail:'In the context of specific
algorithms, identify the characteristics of data and/or
other conditions or assumptions that lead to different
behaviors.',category:c1,Level:"as"}),
(:LearningOutcome{detail:'Determine informally the
time and space complexity of simple algorithms.',
category:c1,Level:us})

```

There are hundreds of pages, but it is possible to stop and examine the model once some data has been loaded: simply close both braces: `}}`.

So far, the DBMS transaction log is as shown in Table I.

There is a single transaction here that has created 6 node types, 4 reference types, 8 named properties, and 22 records. Of course there is much more of CS2013 to insert.

To resume the load, (a) use Match to reestablish the bindings for c1, c2, e1, fa, as, us, and al, (b) add a binding for the knowledge unit currently worked on, which was 'AL/Basic Analysis', and (c) resume the insert operation. If it was switched off, the database reloads the work done already:

```

Match (c1:Category{name:'Core-Tier1'}),
(c2:Category{name:'Core-Tier2'}),
(e1:Category{name:'Elective'}),
(fa:Level{name:'Familiarity'}),
("as":Level{name:'Assessment'}),
(us:Level{name:'Usage'}),
(alba:KnowledgeUnit{name:'AL/Basic Analysis'})
Insert (alba)<-{
(:LearningOutcome{detail:'State the formal definition
of big O.',category:c1,Level:fa}),
(:LearningOutcome{detail:'List and contrast standard
complexity classes.',category:c1,Level:fa})}

```

This time we are only one brace deep. Let us look at the data model we have so far:

```

KnowledgeArea: (name) <- KnowledgeUnit: (name) <-
{ Topic: (detail,category),

```

```

LearningOutcome: (detail, category, level) }

```

We can already see that some data relationships are being handled by simple properties, and others, such as the reference of a KnowledgeUnit (knowledge unit) to its knowledge area, by simple edges. Figure 3 attempts to show this structure.

TABLE I. TRANSACTION LOG FOR CS2013

5	PTransaction for 44 Role=-502 User=-501 Time=14/04/2026 09:39:04
23	PNodeType CATEGORY NODETYPE
49	PNodeType LEVEL NODETYPE
72	PNodeType KNOWLEDGEAREA NODETYPE
103	PNodeType KNOWLEDGEUNIT NODETYPE
134	PNodeType TOPIC NODETYPE
157	Domain REF 23
171	Domain REF 103
185	PNodeType LEARNINGOUTCOME NODETYPE
218	Domain REF 49
232	Domain REF 72
246	PColumn3 NAME for 23(0)[CHAR] NOT NULL
269	Record 269[23] 246=Core-Tier1[CHAR]
292	Record 292[23] 246=Core-Tier2[CHAR]
315	Record 315[23] 246=Elective[CHAR]
336	PColumn3 NAME for 49(0)[CHAR] NOT NULL
359	Record 359[49] 336=Familiarity[CHAR]
383	Record 383[49] 336=Assessment[CHAR]
406	Record 406[49] 336=Usage[CHAR]
424	PColumn3 NAME for 72(0)[CHAR] NOT NULL
447	Record 447[72] 424=AL[CHAR]
462	PColumn3 KNOWLEDGEAREA for 103(0)[232] NOT NULL
495	PColumn3 NAME for 103(1)[CHAR] NOT NULL
519	Record 519[103] 462=447[INTEGER],495=AL/Basic Analysis[CHAR]
556	PColumn3 CATEGORY for 134(0)[157] NOT NULL
585	PColumn3 DETAIL for 134(1)[CHAR] NOT NULL
612	PColumn3 KNOWLEDGEUNIT for 134(2)[171] NOT NULL
647	Record 647[134] 556=269[INTEGER],585= Differences among best, expected, and worst case behaviors of an algorithm[CHAR],612=519[INTEGER]
750	Record 750[134] 556=269[INTEGER],585= Asymptotic analysis of upper and expected complexity bounds[CHAR],612=519[INTEGER]
838	Record 838[134] 556=269[INTEGER],585= Big O notation: formal definition[CHAR],612=519[INTEGER]
900	Record 900[134] 556=269[INTEGER],585= Complexity classes, such as constant, logarithmic, linear, quadratic, and exponential[CHAR],612=519[INTEGER]
1014	Record 1014[134] 556=269[INTEGER],585= Empirical measurements of performance[CHAR],612=519[INTEGER]
1080	Record 1080[134] 556=269[INTEGER],585= Time and space trade-offs in algorithms[CHAR],612=519[INTEGER]
1148	Record 1148[134] 556=292[INTEGER],585= Big O notation: use[CHAR],612=519[INTEGER]
1196	Record 1196[134] 556=292[INTEGER],585= Little o, big omega and big theta notation[CHAR],612=519[INTEGER]
1267	Record 1267[134] 556=292[INTEGER],585= Recurrence relations[CHAR],612=519[INTEGER]
1316	Record 1316[134] 556=292[INTEGER],585= Analysis of iterative and recursive algorithms[CHAR],612=519[INTEGER]
1391	Record 1391[134] 556=292[INTEGER],585= Some version of a Master Theorem[CHAR],612=519[INTEGER]
1452	PColumn3 CATEGORY for 185(0)[157] NOT NULL
1481	PColumn3 DETAIL for 185(1)[CHAR] NOT NULL
1508	PColumn3 KNOWLEDGEUNIT for 185(2)[171] NOT NULL
1543	PColumn3 LEVEL for 185(3)[218] NOT NULL
1570	Record 1570[185] 1452=269[INTEGER],1481= Explain what is meant by "best", "expected", and "worst" case behavior of an algorithm.[CHAR],1508=519[INTEGER],1543=359[INTEGER]
1705	Record 1705[185] 1452=269[INTEGER],1481= In the context of specific algorithms, identify the characteristics of data and/or other conditions or assumptions that lead to different behaviors.[CHAR],1508=519[INTEGER],1543=383[INTEGER]
1890	Record 1890[185] 1452=269[INTEGER],1481= Determine informally the time and space complexity of simple algorithms.[CHAR],1508=519[INTEGER],1543=406[INTEGER]

The data captured in the above statements do not conform to ISO39075 node/edge graphs but can be simply stored in SQL tables with reference columns as described above. Then any of the formats discussed in section II above can easily be exported using JSON for input to other systems or formatted back into a readable document rather like the ACM report we began with. More importantly, it can be queried and analyzed, and further data and metadata can be added about authorship, updates, provenance, delivery by universities, and outcomes.

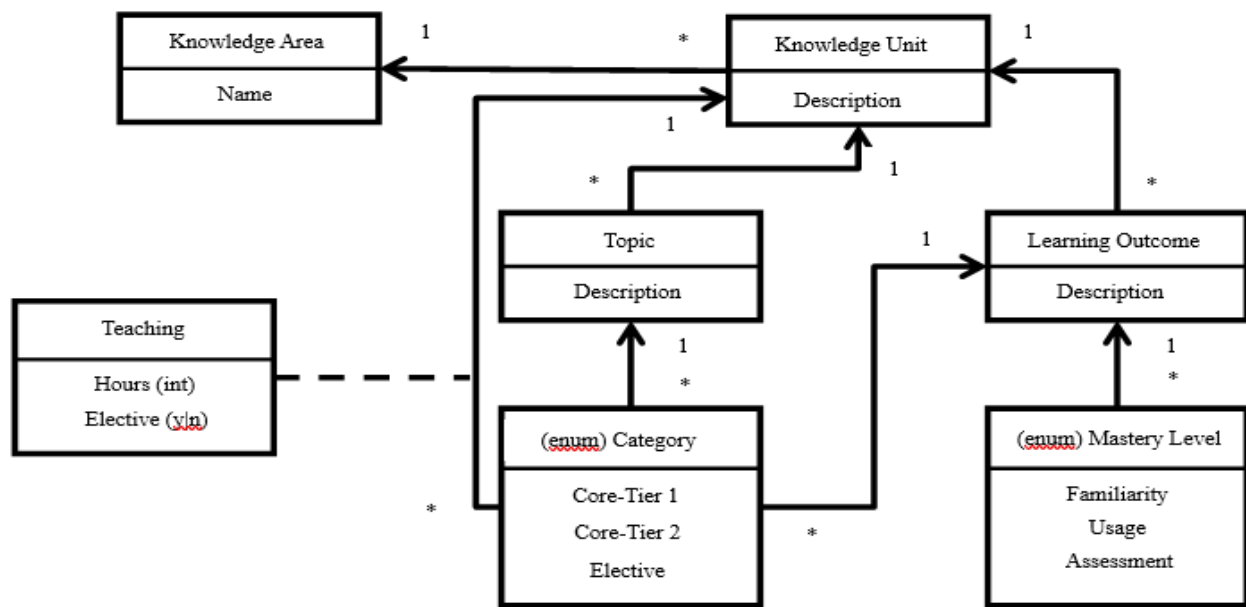


Figure 3. The implied data model

In more complex graphs, the ability to label arrows as in -AnnualFee-> is useful. Such arrow labels can be used as an alternative to labelling node types, and in combination can provide facilities for generalization and/or subtyping.

Reverse arrows can also be allowed, and in every case the node at the arrow base will contain a reference to the node at the arrowhead. With these conventions, a new label can be bound by its use on an arrow or as a node type, and on commit such bindings will be saved as a modification to the current schema (or role) in the database.

A graph schema consists of a working set of node types, edge types, arrow types and graph instances. The schema can be constructed by declaration (a closed schema), or incrementally as in the above example.

Figure 3 shows the schema implied by the above steps, with some tentative additions suggested by the text of the report [10] (a) multiplicity constraints, (b) refinement to property types (e.g., specifying the Level values as an enumeration), and (c) adding the teaching information shown in the top part of Figure 2 as a GQL-style edge.

The addition of constraints and refinement of types as in (a) and (b) here can be performed non-destructively provided the existing values satisfy the constraints or can be coerced to the new types: such schema changes in Pyrrho will modify the associated data in a cascade, whose failure will prevent any changes occurring.

V. IMPLEMENTATION HIGHLIGHTS

In this section, we briefly review a few aspects of the implementation relevant to the above discussion of the transaction log and foreign keys.

The open-source DBMS server is v8 of PyrrhoDBMS and is available on Github. It is a record-based system with a persistent append-storage transaction log: this guarantees

fully serialized isolated transactions. The transaction log is a binary file containing a sequence of objects representing a transaction start, a table definition, a column definition, a data record and so on: examples of these objects are illustrated below.

Transactions can include a mixture of data changes and schema changes. Since 2019, the implementation has used shareable data structures: these are tree structures that behave like ordinary strings to the extent that any change results in a new tree. Importantly, the server maintains these in memory as long as they are in use, and the new tree continues to use nearly all the same nodes of the old version. The only new nodes form a path from the new root to the node or modified node that has just been added. Figure 4, from [12], explains this mechanism. The result of this architecture is that every change to a database of size N takes $O(\log N)$ of processing time and storage.

This design is implemented in Pyrrho [2] using a generic tree structure $Ctree\langle K, V \rangle$. It is called $Ctree$ rather than $Btree$ if both the generic parameters K and V contain values that can be compared using $<$, $>$, and $=$.

In SQL, a primary key declares a lookup index, while a foreign key declares a sort of reverse lookup: Pyrrho's implementation replaces the foreign key index with a hidden reference column. For example, the SQL statements

```
QL> create table company (name string primary key)
QL> create table person (name string primary
key, worksin string references company)
result in the following entries in the transaction log:
23 PTable COMPANY
35 PColumn3 NAME for 23(0)[Domain CHAR]
58 PIndex COMPANY on 23(35)PrimaryKey
98 PTable PERSON
```

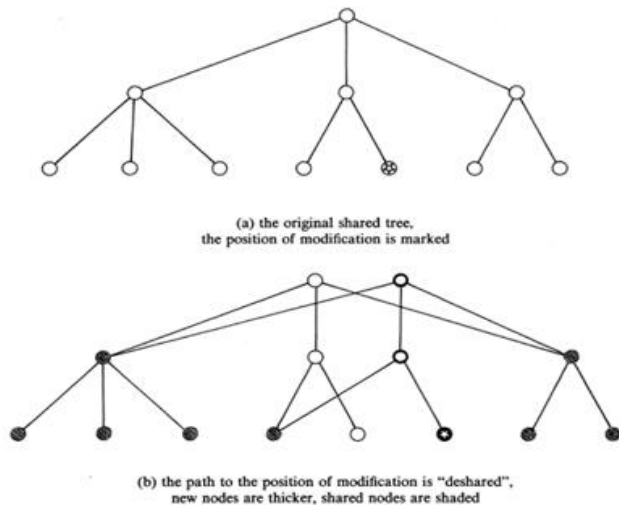


Figure 4. The shareable data concept. [12]

Figure 4. The shareable data concept [12].

```

109 PColumn3 NAME for 98(0)[Domain CHAR]
132 PIndex PERSON on 98(109)PrimaryKey
153 PColumn3 WORKSIN for 98(1)[Domain CHAR]
180 PIndex COMPANY on 23(153) RestrictUpdate,
RestrictDelete
204 Domain REF 23
218 PColumn3 COMPANY for 98(2)[204] RestrictUpdate,
RestrictDelete keymap 180

```

The last two entries define a reference column that is to be constructed from the value of the WORKSIN column. In the database itself, there is no object corresponding to the reverse index 180, and the Person table displays the first two columns NAME and WORKSIN:

```

{COMPANY TABLE (35) Display=1 [35,Domain CHAR]
Indexes:((35)58)}
{TableColumn NAME 35 Domain CHAR optional Table=23
RefCols: [218]}
{Index 58 for 23 Key:Domain ROW (35)
Display=1[35,Domain CHAR]
{PERSON TABLE (109,153|218) Display=2 [109,Domain
CHAR], [153,Domain CHAR], [218,Domain REF 23]
Indexes:((109)132) ColRefs: 23=[218]}
{TableColumn NAME 109 Domain CHAR optional Table=98}
{Index 132 for 98 Key:Domain ROW (109)
Display=1[109,Domain CHAR]
{TableColumn WORKSIN 153 Domain CHAR optional Table=98}
{TableColumn 218 Domain REF 23 optional Table=98 keyMap
(153) RestrictUpdate, RestrictDelete}

```

Column 218 contains values generated from the WORKSIN column 153 using the primary index of table 23. The mechanism for keeping track of references using structural row uids is detailed below. It should now be relatively clear how reference values support edge connections.

The implementation of graph types in Pyrrho is role-based and is intended to be intuitive. Given a relational database in Pyrrho and the right privileges, new graph data can be created within the current role by defining element types (under or by adding metadata to existing tables or types) or incrementally using INSERT. If the user has sufficient privileges to modify existing tables, new properties can be added explicitly or implicitly, and new edge

connections implemented (with default properties FROM, TO and WITH. By default, new graph properties and references are not optional.

Existing relational data in the database can be accessed using a default graph model whereby tables without foreign keys can be accessed as node types, and tables with at least two foreign keys can be used as edge types: the first two foreign key columns become the default FROM and TO connections, and additional or replacement element types can be constructed or modified implicitly as above or by explicit create and alter statements. Importantly, in all cases the identification of graph data in the database is role-specific, sharing the underlying tables and types with other roles subject to privileges granted to them.

This approach is pragmatic rather than purist: while by abuse of language SQL types are treated internally as a subclass of SQL tables and GQL edge types a subclass of GQL node types, they occupy different namespaces and for most purposes overloading of names will not be a problem; for example, a graph pattern will search for a node label in the set of nodetypes or an edge label in the set of edgetypes, while SQL uses different syntax for tables and types. An explicit CREATE with the same name as another object of a different class will create a new object with a different uid, and any name conflicts can be resolved by role-based renaming and/or the use of numeric uids for objects hidden by name overloading. SQL tables and SQL types have different semantics for columns and fields respectively: for example, a table column may contain nulls unless declared NOT NULL while a type field is required unless declared OPTIONAL. These aspects remain true even if the table is treated as a node type, or a node type is treated as an edgetype. Importantly, SQL types support methods and UNDER, which is needed when the set of columns is extended by the the subtype ("width subtyping"), while GQL types defined on an SQL table are only able to use subtypes defined by a "depth-subtyping" constraint such as CHECK FROM IS OF(person), WHERE MATCH ..., or the recently-proposed REQUIRE syntax.

There are some resulting extensions to SQL, allowing multiple inheritance to be permitted for graph types in addition to nesting. Pyrrho also permits n-ary edge type. A graph type can be constructed implicitly by an inline insert/create statement if it has a label without &. Pyrrho does not have a distinguished ID column: if such a column is declared it has no special status beyond being the default name for the node in visualisations of portions of the graph.

In Section IV, we saw how the arrow notation automatically adopted (or created) reference links between records. The implementation maintains a list of reference connections for each table, and these can be upgraded to edge types.

Assuming the current user has the privileges to do so (e.g., database owner), the data can be dynamically extended during data insertion: extra properties added to nodes, and single-valued properties can be converted to set-valued properties. GQL allows sets of node labels, and these simply inherit the union of properties associated with these labels. In

Section III, the introduction of an edge property between, say, KnowledgeUnit and Topic, as in

`-[:Syllabus{at:'Yale'}]->` would infer the creation of a table of edges specific to one or more universities, and also give a name to such edges (section III left them unnamed).

Once the data model is built, JSON can be generated directly by the server suitable for import into a large language model. Currently the server has system tables that contain these JSON items for the schema, for example table "Role\$Json". In the above example, this gives the schemata shown in Tables II and III.

TABLE II. SCHEMA FOR CS2013

Name	Definition
CATEGORY	{Fields:[NAME:{Domain:CHAR}],Display:1}
LEVEL	{Fields:[NAME:{Domain:CHAR}],Display:1}
KNOWLEDGEAREA	{Fields:[NAME:{Domain:CHAR}],Display:1}
KNOWLEDGEUNIT	{Fields:[KNOWLEDGEAREA:{Domain:REF,ToType:'KNOWLEDGEAREA'},NAME:{Domain:CHAR}],Display:2}
TOPIC	{Fields:[CATEGORY:{Domain:REF,ToType:'CATEGORY'},DETAIL:{Domain:CHAR},KNOWLEDGEUNIT:{Domain:REF,ToType:'KNOWLEDGEUNIT'}],Display:3}
LEARNINGOUTCOME	{Fields:[CATEGORY:{Domain:REF,ToType:'CATEGORY'},DETAIL:{Domain:CHAR},KNOWLEDGEUNIT:{Domain:REF,ToType:'KNOWLEDGEUNIT'},LEVEL:{Domain:REF,ToType:'LEVEL'}],Display:4}

TABLE III. FOR THE AUTHOR AND BOOK EXAMPLE

Name	Definition
AUTHOR	{Fields:[ID:{Domain:INTEGER},ANAME:{Domain:CHAR}],Display:2}
BOOK	{Fields:[ID:{Domain:INTEGER},AUTHID:{Domain:INTEGER},TITLE:{Domain:CHAR},AUTHOR:{Domain:REF,ToType:'AUTHOR'}],Display:3}

In this way, graphical database concepts are simply added to the basic SQL implementation. NodeTypes are a subclass of user defined types, and EdgeTypes are a subclass of NodeTypes, but the role manages separate namespaces for each kind of object. We recall that database objects and their contents can be renamed in roles.

Graphical data is nevertheless record-based data, and so all these classes are implemented as tables. Tables and user-defined types can acquire graphical properties by adding metadata or by the operation of the graphical INSERT statement. However, properties are always implemented by columns in tables or fields in types, and edges are implemented behind the scenes using reference values, and graphical operations and specifications must be compatible with (or, in the hands of an object owner, modify) existing database structure. By default, columns in tables represent optional properties, and columns and references in user-defined types are not optional.

Pyrrho enhances the GQL and SQL model by means of subtypes and unions, on the principle that additional properties imply subtypes. A new node type in GQL can have a set of type labels, and if these are existing types or tables in the database, all of their properties are properties of

the new type. An unlabeled node in a Match statement is constrained by adjacent nodes and edges to belong to a particular table or type.

In this way, information about database objects can be role-specific and dynamic. The database objects affected have system-managed properties that reflect graphical aspects of the current database contents as follows:

SysRefIndex keeps track of reference values. It is a database-level property of any referenced Table T of form `Ctree<long,Ctree<long,Ctree<long,bool>>>`. The first long is the uid of a referenced row rr in T, the second is the uid of a referencing column rc in another Table S, and the third is the uid of a referencing row ss in S. Then `rr = ss.vals[rc]`, so that SysRefIndex provides a reverse index to referencing data.

ColRefs is a list of referencing columns in a given Domain D, of form `Ctree<long,Ctree<long,bool>>`. The first long is a Table referenced by D, while the second is the referencing column in D.

Model is a role-dependent (or schema-dependent) property of any database object S that contains references accessed using type names or named arrows, of form `Ctree<string,Ctree<Qlx,long>>`. Here the string is the name of a referenced table T, the token is an arrow, and the long identifies a reference column of S. A useful convention is to give a simple reference column the name of the referenced table, and the default arrow name is the empty string so the statement parser can use a simple algorithm to compute the reference column for each reference.

The server is a multi-threaded TCP server using platform- and culture-independent strong primitive data types. It is implemented for Microsoft .NET Core 10.0 and so will run on Windows, Linux and MacOS.

VI. PERFORMANCE MEASUREMENTS

As mentioned above, there are good reasons to believe that the performance of the engine as a DBMS will be satisfactory. PyrrhoDBMS builds databases completely in memory using disk storage merely for the persistent transaction log. For the Graphics Data Council Financial Benchmark [13], we obtained the results shown in Table IV.

TABLE IV. BUILDING THE FINANCIAL BENCHMARK

Scale	Build time (seconds)	Ratio	Log Size (KB)	Ratio
SF001	83		5768	
SF01	868	10.46	57383	9.95
SF1	9507	10.96	579151	10.09

For this experiment, the benchmark was modified as shown in the Appendix to showcase subtypes and ternary edges and use a combination of SQL and GQL syntax. The task was to build the database from the spreadsheets that use the original data model provided by LDBC. For all of the many ternary edges (subclassing TRANSFER) a third reference property was calculated using a default expression that uses the Match statement, highlighted in the Appendix. Performance of the DBMS shows very good scalability. It

showed that the performance was nearly linear in the construction time for graphical models with transaction log size 5 MB, 50 MB, 500 MB. The amount of memory used to hold the database was also nearly linear, at roughly 10 times the transaction log size. Timings for a more general workload will depend on the complexity of the workload and (for example) whether multiplicity constraints require to be rechecked on every data modifying transaction.

VII. CONCLUSIONS AND FUTURE WORK

This work is primarily a continuation of our integration of graphical database ideas into our SQL implementation, and the provision of support for resource description using JSON. It seems clear that the use of references to data entities as records is simpler than using lists of key values, especially when navigating links between records. For this reason, we believe that our work will support greater integration of real-world data into systems that seek to build models of expected behavior, such as resource augmented generation.

REFERENCES

- [1] S. Bhatt et al., "Knowledge graph semantic enhancement of input data for improving AI." *IEEE Internet Computing* 24.2, 2020, pp. 66-72.
- [2] M. K. Crowe: The Pyrrho DBMS, <https://github.com/MalcolmCrowe/ShareableDataStructures> subfolder PyrrhoV8alpha, (accessed 2026.05.07).
- [3] O. van Rest, Property Graphs in Oracle Database 23ai: The SQL/PGQ standard, Oracle Database Insider <https://blogs.oracle.com/database/property-graphs-in-oracle-database-23ai-the-sql-pgq-standard> (March 2025) (accessed 2026.05.07).
- [4] D. ten Wolde, DuckPGQ.org, <https://duckpgq.org/> (accessed 2026.05.07)
- [5] Google Cloud Spanner <https://cloud.google.com/products/spanner/graph> (accessed 2026.05.07)
- [6] Neo4j, Neo4j welcomes New GQL International Standard (April 2024)
- [7] H. Kondylakis et al. "Property Graph Standards: State of the Art & Open Challenges." *Proceedings of the VLDB Endowment* 18.12, 2025, pp. 5477-5481.
- [8] P. Skavantzios and S. Link, Normalizing Property Graphs. *Proc. VLDB Endow.*, 2023.
- [9] J. Shute, C. Zheng, and R. Kuditkar, "Semantic Data Modeling, Graph Query, and SQL, Together at Last?" *VLDB* 2026.
- [10] M. K. Crowe and F. Laux, Property Graph Techniques in Relational Databases, *DBKDA 2026, The Eighteenth International Conference on Advances in Databases, Knowledge, and Data Applications*, March 2026, Valencia, ISBN: 978-1-68558-359-0 p.7-12.
- [11] ACM/IEEE: Computer Science Curricula 2013, ISBN 978-1-4503-2309-3.
- [12] T. Krijnen and G. L. T. Meertens, "Making B-Trees work for B". Amsterdam: Stichting Mathematisch Centrum, 1982, Technical Report IW 219.
- [13] Graph Data Council (formerly Linked Data Benchmark Council), Financial Benchmark <https://ldbouncil.org/benchmarks/finbench/> (Accessed 2026.05.07).

APPENDIX

```
create type legalentity as (name string, isBlocked
boolean, createtime timestamp, country string, city
string) nodetype
create type person under legalentity as (id int primary
key, gender string, birthday timestamp)
create type company under legalentity as (cid int
primary key, business string, description string,url
string)
create type actbase as (createtime timestamp, accttype
string) nodetype
alter table actbase add balance float64 default 0.0
create type account under actbase as (id int primary
key, isBlocked boolean, nickname string, phonenumber
string, email string, freqlogintype string,
lastlogintime timestamp, acctlevel string)
create type loan under actbase as (lid int primary key,
loanAmt float64, interest float32)
create type medium as (id int primary key, type string,
isblocked boolean, createtime timestamp, lastlogin
timestamp, risklevel string) nodetype
create type apply as (createTime timestamp,
organization string) edgetype (legalentity, loan)
create type personapplyloan under apply edgetype
(person, loan)
create type companyapplyloan under apply edgetype
(company, loan)
create type guarantee as (createTime timestamp,
relationship string) edgetype (legalentity,legalentity)
create type personguaranteeperson under guarantee
edgetype (person, person)
create type invest as (ratio float64, createTime
timestamp) edgetype (legalentity, legalentity)
create type personinvestcompany under invest edgetype
(person, company)
create type owns as (createTime timestamp) edgetype
(legalentity, actbase)
create type personownsaccount under owns edgetype
(person, account)
create type activatedfor as (createtime timestamp,
location string) edgetype (med=medium, account)
create type transfer as (amount float64, createTime
timestamp, orderNumber int optional, comment string
optional, payType string optional, goodsType string
optional) edgetype (from actbase with activatedfor
optional to actbase)
alter table transfer alter with1 to usedFor
alter table transfer alter usedFor add default (match
any 1 (from1)<-[a:activatedfor where
a.createtime>=from1.createtime]-() yield a)
create type accountrepayloan under transfer edgetype
(from account to loan)
create type accounttransferaccount under transfer
edgetype (from account to account)
create type accountwithdrawaccount under transfer
edgetype (from account to account)
create type companyguaranteecompany under guarantee
edgetype (company, company)
create type companyinvestcompany under invest edgetype
(company, company)
create type companyownsaccount under owns edgetype
(company, account)
create type loandepositaccount under transfer edgetype
(from loan with activatedfor optional to account)
```

PyWIB: A Python Library for a Multi-Modal Approach to Web Interaction Behavior Analysis

Guillermo Dylan Carvajal-Aza¹, Alejandro Alvarez-Varela¹, Javier De Andres², Martin Gonzalez-Rodriguez¹, Daniel Fernandez-Lanvin¹, Mercedes Paino³

¹Dept. of Computer Science.

²Dept. of Accounting.

³Dept. of Psychology.

University of Oviedo, Oviedo, Spain

Email: {carvajalguillermo, avarela, jdandres, martin, dflanvin, mpaino}@uniovi.es

Abstract— This paper addresses the need for efficient tools in multi-modal web interaction analysis, as existing frameworks often lack integration for diverse data streams. PyWIB (Python Web Interaction Behavior) is a new Python library designed for extracting and analyzing metrics from user interactions with web pages. It aims to streamline the analysis of user interaction data in the field of Human-Computer Interaction, providing researchers with a comprehensive set of methods to process event logs and compute metrics essential for such research. PyWIB is open-source and available on GitHub. The results demonstrate that PyWIB effectively unifies mouse-tracking and keystroke dynamics into a single processing pipeline.

Keywords—Human-Computer Interaction; interaction analysis; web interaction patterns; web behavior; python.

I. INTRODUCTION

The analysis of web interaction has been a cornerstone of Human-Computer Interaction (HCI) research for over two decades, evolving from early clickstream discovery [1] to the automated usability evaluation frameworks that characterize modern research [2]. Among the various techniques used to analyze the interaction of a user with a web page, mouse-tracking and keystroke dynamics [3] are the two most commonly used methods for obtaining insights into the user's specific characteristics.

These two types of interaction analysis have been widely conducted in HCI to infer several aspects of the user's behavior and interaction with web and computer interfaces. Both dynamics have been proven useful in different fields of human behavior. Mouse dynamics have been employed in the analysis of human behavioral patterns and factors [4], cognitive and physical conditions affecting the user [5], [6], [7], in user identification [8] and in studies aiming to measure the influence of age or gender in common computer tasks [9], [10]. In contrast, keystroke dynamics have been widely used for the recognition of users' emotions and personalities [5], as well as user authentication, behavioral biometrics, and affective computing [11].

These types of analysis require a lengthy implementation process, from data collection to its processing; the data generated from a mouse-tracking session of 60 seconds can record up to 900 entries (recording mouse events with

intervals of 100 milliseconds), making it difficult to scale and manage without appropriate computational methods.

As data science gains importance within the field of HCI and scientific tools become increasingly accessible, there is a clear need for a platform that provides a straightforward way to extract metrics from datasets. Furthermore, a well-documented repository of user interaction metrics is essential to help researchers identify which ones are best suited for their specific studies.

While there already exist other software libraries dedicated to extracting metrics from users' data (e.g., Mousetrap) [12], to the extent of our knowledge, none have been implemented in a language as versatile as Python, which allows the creation of Application Programming Interfaces (APIs) that can be deployed in web environments and integrated into complex data science pipelines. Consequently, PyWIB facilitates a more efficient research workflow.

Additionally, visualization tools for mouse and keystroke events are not often available to everyone, as the existing tools tend to hide these features behind paywalls after their free trial (such as Hotjar [13], Mouseflow [14] or Microsoft Clarity [15]). These visualization tools are essential in some cases for the interpretation of users' results or analyses carried out using computer vision techniques [4], forcing researchers to implement their own solution to this problem.

To address the existing software limitations, PyWIB has been developed by the Human Communication Interaction Research Group from the University of Oviedo as an open-source implementation available for download as a Python library. This library has been designed as a modular library, which provides the user with a set of functions, for which we have made available extensive documentation on both their technical and mathematical implementation, on the official documentation site [16].

By using this modular approach, we have created a library that can not only be easily expanded to incorporate new computations but can also be used by other programs through the creation of an API, enhancing interoperability.

To demonstrate its utility, PyWIB facilitates representative use cases in common HCI research contexts. First, it supports usability and interaction efficiency studies, where mouse-trajectory metrics quantify motor precision during pointing or drag-and-drop tasks [10], allowing researchers to objectively compare interaction strategies

across different interface designs. Additionally, the library supports research on cognitive and affective states. Keystroke dynamics can be analyzed to investigate cognitive load, emotional patterns, or error-correction behavior, providing insights into user performance beyond basic execution times [5], [17].

The remainder of this paper is structured as follows. Section II describes metric calculation and validation. Section III details the software design and framework. Section IV explains the segmentation process. Section V presents visualization features. Finally, Section VI discusses limitations, concludes the paper, and outlines future work.

II. METRICS CALCULATION AND VALIDATION

One of the main problems when dealing with a library that aims to cover the computation of the most common metrics in HCI research is the validation of those metrics. For this reason, PyWIB has been developed considering the most relevant metrics used in research works, which have been proven to be representative of user behavior in different contexts. The following metrics are available for computation using tracking datasets in the current PyWIB release:

- A. **Velocity:** Rate of change of position over time.
- B. **Acceleration:** Rate of change of the velocity over time.
- C. **Jerkiness:** Rate of change of the acceleration over time.
- D. **Path:** Total distance traveled.
- E. **Area Under the Curve (AUC) Ratio:** Deviation of the user's trajectory with respect to the optimal path.
- F. **Maximum Absolute Deviation (MAD):** Maximum perpendicular offset from the ideal straight-line trajectory.
- G. **Average Absolute Deviation (AAD):** Mean perpendicular displacement from the optimal straight-line path.
- H. **Mouse:** Click counts and activation-to-release deviations.
 - a. Number of clicks
 - b. Click Slip
- I. **Keyboard:** Typing dynamics and error-correction patterns.
 - a. Typing Duration
 - b. Typing Speed
 - c. Backspace Usage
- J. **Time:** Execution, active movement, and pause count.
 - a. Execution Time
 - b. Movement Time
 - c. Number of Pauses

A. Velocity

Velocity is the rate of change of position of the user's cursor with respect to the observed time of the different points that mark the user's interaction events, measured in pixels per second [18].

The applied formula for this calculation is as follows:

$$v_i = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{t_i - t_{i-1}} \quad (1)$$

where (x_i, y_i) are the coordinates and t_i is the timestamp on the analyzed point.

B. Acceleration

Acceleration is calculated as the change of velocity with respect to time [18]. For this metric, it is imperative to compute the velocity of each session trace by applying the previous formula.

The applied acceleration formula is as follows:

$$a_i = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \quad (2)$$

where v_i is the velocity at the examined time and t_i is the value of the timestamp for the given point in the user's interaction.

C. Jerkiness

Jerkiness is a metric derived from acceleration, computed as the change in acceleration per unit of time [19]. This metric, while not widely used, has been applied in studies to identify bursts of speed and identify irregularities in motor control, reflecting the degree of movement fluidity or hesitation.

$$j_i = \frac{a_i - a_{i-1}}{t_i - t_{i-1}} \quad (3)$$

D. Path

The path is the total distance travelled, a key metric computed as the total distance between consecutive points in the user interaction [4], [6], [18], [20], calculated as:

$$\sum_{i=0}^n \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \quad (4)$$

E. Area Under the Curve

The AUC is a metric that quantifies the deviation of the user's trajectory with respect to the optimal path from the start point to the end point of his interaction. This metric has been widely used in HCI when analyzing users' trajectory [6], [18]. PyWIB provides two approaches for obtaining this metric. Given the presence of loops and backtracking in the user's trajectory, these approaches allow researchers to choose whether to penalize motor inefficiencies and user control by accounting for such behaviors or to mitigate their impact through a geometrical simplification of the trajectory.

F. Maximum Absolute Deviation

The MAD is a metric calculated as the maximum perpendicular distance from any point on the user's actual path to the straight line that connects the starting and ending points of his movement (that is, the optimal trajectory the user should follow) [4], [6], [18].

G. Average Absolute Deviation

The AAD consists of the average of the perpendicular distances from each point of the trajectory to the straight line connecting the start and end points of the movement (the optimal trajectory), as defined in [6].

H. Mouse

1) Number of Clicks

This metric consists of the total number of times an event of type `EVENT_ON_CLICK` is registered for the user's session or trace [21].

2) Click Slip

The click slip is a metric that measures the distance of click positions from the intended targets, as defined by [7], computed taking the initial point where an event of type `EVENT_ON_MOUSE_DOWN` takes place and the distance to its next event of type `EVENT_ON_MOUSE_UP`.

I. Keyboard

The metrics obtained from keyboard interaction have been widely used in various applications, including user authentication, behavioral biometrics, affective computing, and HCI studies [3], [5], [17], [22].

To the best of our knowledge, there is limited research on the use of keystroke dynamics and behavioral user patterns. The keystroke metrics here defined are derived from the works developed by Katerina and Nicolaos [4] and Khanna and Sasikumar [17], where the authors explore the use of keystroke dynamics in relation to the recognition of behavioral user attributes from their stroke patterns.

1) Typing Duration

A time-based metric for keystrokes that captures the time interval between an event of a key-press and an event of type key-release [4], [17].

2) Typing Speed

This metric refers to the average number of typed Characters Per Minute (CPM) by a user during a session, defined and used in research by [17].

3) Backspace Usage

The number of times a user presses the backspace (or any deletion key) during a session or trace. This metric can be helpful to assess a negative emotional state and users' error-correction behavior as done by [17].

J. Time

1) Execution Time

Also referred to as "total time" or "session time", as it represents the overall duration of a user session from the first recorded event to the last, as described in [7]. It can be useful when filtering sessions based on their length if the task being studied has a defined expected minimum duration. It is important to note that this metric does not account for periods of inactivity or pauses within the session.

2) Movement Time

Defined by [7], this metric focuses on active movement events during a session, excluding idle times, and provides a more accurate representation of the time spent in motion by the user. It is useful when focusing on the time of the actual interaction, rather than filtering based on total session duration.

3) Number of Pauses

As the name suggests, the number of pauses refers to occurrences in which the user halts interaction for a minimum duration. This threshold is typically measured in milliseconds, commonly set between 100 and 200 milliseconds [7], [17], as these values represent the boundary between automatic motor execution and cognitive processing delays.

Regarding the correctness of metric computation, PyWIB has been designed with validation as a core principle. The implementation of all metrics follows their formal mathematical definitions as reported in the literature, and input data are systematically validated to ensure consistency and completeness prior to computation. Moreover, unit tests have been implemented for all metric modules using the `pytest` framework, covering both valid and invalid input scenarios to verify numerical correctness and appropriate error handling.

Not all metrics are equally valid for all experimental contexts. For example, metrics that are valid for analyzing mouse movements in a desktop web application may not be valid for analyzing touch interactions on mobile devices, as it could make no sense to analyze users speed if they can only tap the screen or drag and drop elements. Therefore, it is crucial to consider the context in which the metrics will be applied and to verify that their use is correct.

Moreover, the setup of an experiment itself can influence the validity of certain metrics [23], [24], which is why PyWIB encourages users to verify the correct use of the metrics they compute in their specific contexts and experimental setups.

III. SOFTWARE DESIGN, FRAMEWORK AND FEATURES

The PyWIB package has been built using a modular approach in Python 3.9, inspired by mathematical and data analysis libraries like NumPy and Matplotlib, which employ a hybrid approach with Object Orientation and Procedural programming aspects. Even if, at the start, our library has only found the need for describing modules, this does not mean it cannot lean towards a hybrid approach, following the necessities of the scientific community and the specific cases it tackles in the field of HCI.

The code also adopts a defensive programming paradigm [25], where all inputs are validated prior to processing. This ensures early detection of invalid states, such as missing columns or *null* DataFrames. Unit tests have also been implemented to verify the correct workflow of inputs using the Python testing framework `pytest`. In these tests, validation functions were tested against both valid and invalid inputs to ensure correct exception handling and error reporting, among

all metrics provided by the library. Together, the validation layer and unit testing framework ensure that the software behaves predictably under a wide range of conditions, reducing runtime errors and increasing reproducibility.

The PyWIB library has dependencies on the following packages, with a minimum requirement of their versions:

- Pandas' versions equal or greater than 1.3.0
- NumPy versions equal or greater than 1.21.0
- Matplotlib versions equal or greater than 3.5.0
- OpenCV-python versions equal or greater than 4.12.0.88
- Seaborn version equal or greater than 0.13.2

As for its current version (1.0.0), its modular approach allows for easy implementation of new modules and flexibility to be combined with deeper algorithms or new visualization techniques.

The structure of this library has been separated into packages based on use cases related to the tracking of user interaction. The main packages that compose the library are represented in Figure 1, which presents a UML package diagram of the software framework. This separation of concerns improves maintainability, extensibility, and testability.

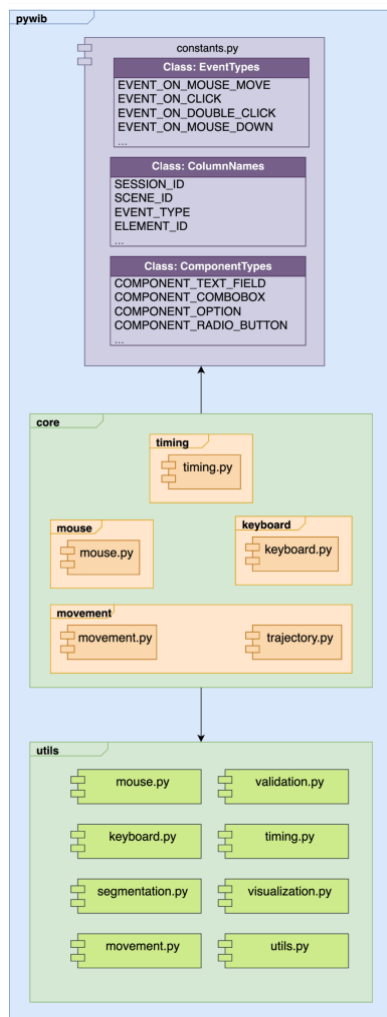


Figure 1. PyWIB Package Diagram.

These main packages can be described as follows:

- From PyWIB core package:
 - Keyboard: containing metrics related to keystroke analysis.
 - Mouse: containing metrics related to mouse clicking, such as number of clicks
- Movement: containing metrics related to both the movement the user describes over the web interface and his described trajectory. The metrics included in this package are applicable to both mouse-tracking on computers and tracking over mobile and tablet devices.
- Timing: containing metrics related to time, such as execution time and movement time (time the user spends moving through the web interface, without idle times).
- From PyWIB utils package:
 - Visualization: containing methods used to create visual representations of the user trajectory, heatmaps over his keystrokes, etc.
 - Segmentation: containing methods related to the segmentation of datasets into different traces.

The library also contains a package that lists a set of constants related to events, key codes, columns used to compute metrics and component types used during preprocessing and analysis. These constants have been set according to the ones produced by the Interaction Lab tool [26], used by our team to facilitate compatibility with the tool, which are intended to represent numerically all possible JavaScript events on computers and mobile devices.

IV. SEGMENTATION

One of the key elements of PyWIB is the analysis based on the segmentation of long user sessions into small movement traces. Given a Data Frame that contains the user's interaction, a trace is considered as a sequence of consecutive (non-stopped) mouse/touch movements between two non-movement events.

This type of segmentation has been widely employed in HCI research to obtain better insights into the users' movement and trajectories [7], [27].

This movement segmentation is implemented over two functions: *extract_trace* and *extract_traces_by_session*, both take a Data Frame as input and return a list or a dictionary containing each session as a key and the list of traces as a value.

These segmentations are not limited to metrics related to movement, as for the current version, there is also a method that extracts traces from keystroke interaction.

The method *extract_keystroke_traces_by_session* reduces a Data Frame to a dictionary that contains, as values, lists of segments of the user's interaction in which the interactions occur only through the keyboard. This method can be useful in cases where keystroke dynamics require analysis for speed or other time-based metrics.

V. VISUALIZATION

In its current version, PyWIB offers three types of visualization that can be created from the interaction data, by making use of the libraries *seaborn* and *matplotlib*.

Trace visualization, through method *visualize_trace*, creates an image using the user's movement to represent the movement of the mouse over the screen. This method can be used either with a full DataFrame, containing the entirety of the user's interaction, or with segmented traces, to analyze a concrete movement, as seen in Figure 2.

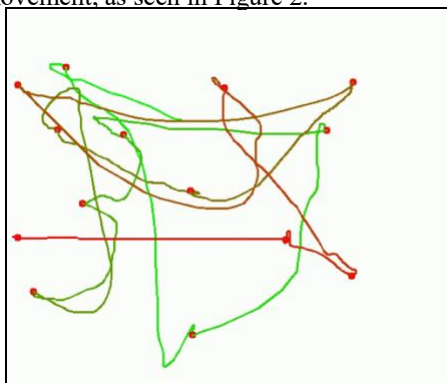


Figure 2. Visualization of the Mouse over the Screen.

Using a similar approach, *video_from_trace* method generates an mp4 video from the user's interaction using its timestamps to show his movement over time.

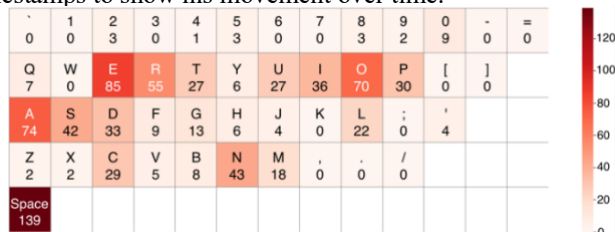


Figure 3. Heatmap of Keystrokes.

Finally, a keystroke heatmap is generated via the *keyboard_heatmap* function to visualize typing intensity across the layout. This visualization is exemplified in Figure 3, showcasing the spatial distribution of key presses.

VI. LIMITATIONS, CONCLUSION AND FUTURE WORK

In this paper, we have presented PyWIB, a novel Python library designed to support HCI researchers in the preprocessing and analysis of user interaction tracking data. Its modular architecture facilitates seamless integration with existing Python scripts, APIs, or other tools, making it a flexible and extensible solution that enables users to manipulate data according to their specific needs. Moreover, it offers a transparent and accessible alternative to proprietary software, providing useful visualization tools such as heatmaps and trace reconstruction without the constraints of paywalls. By adopting a defensive programming approach and offering standardized and validated metrics, PyWIB enhances reproducibility.

While PyWIB is intended to ease the workload of HCI researchers alongside their existing toolchains, its current

implementation still leaves room for improvement. Specifically, further optimization is required to reduce computation time when handling large-scale datasets.

As PyWIB was made publicly available as an open-source project on GitHub in January 2026, increasing its recognition and adoption within the research community constitutes an important direction for future work. Planned efforts include dissemination through scientific publications and conference venues, as well as the development of example-driven documentation and research-oriented use cases to promote adoption by researchers and practitioners.

Future work will focus on the systematic evaluation of PyWIB to empirically validate its effectiveness. Planned studies will assess metric correctness, computational efficiency, and robustness using large-scale, real-world interaction datasets to demonstrate the library's practical utility in HCI research.

To directly address these computational limitations, future development of this software should primarily focus on optimizing performance, leveraging available hardware (e.g., GPUs), and integrating frameworks such as PySpark to enable real-time, large-scale data processing in a distributed environment using Python and Apache Spark. Looking ahead, there is also significant potential to further enhance PyWIB by expanding its computational capabilities and incorporating new metrics, such as fixation duration for eye-tracking integration, scroll-depth analysis, or touch-gesture complexity for mobile environments, as the needs of the scientific community grow. Additionally, adding multi-format export options (e.g., JSON and CSV) and basic statistical built-in functions, such as T-tests or ANOVA, for automated metric comparison, will ensure better compatibility and expanded analysis capabilities.

ACKNOWLEDGMENT

This research has been partially funded by the Carlos III Health Institute (reference FISS-23-PI23-00416) co-funded by the European Union, and by the Ministry of Science, Innovation and Universities and co-funded by the European Regional Development Fund (ERDF) under project PID2024-155586OB-I00 (MCIU/AEI/10.13039/501100011033). Additional support was provided by the Government of the Principality of Asturias through grant GRU-GIC-24-070.

REFERENCES

- [1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data," *SIGKDD Explor. Newsl.*, vol. 1, no. 2, pp. 12–23, Jan. 2000, doi: 10.1145/846183.846188.
- [2] M. Y. Ivory and M. A. Hearst, "The state of the art in automating usability evaluation of user interfaces," *ACM Comput. Surv.*, vol. 33, no. 4, pp. 470–516, Dec. 2001, doi: 10.1145/503112.503114.
- [3] L. M. Vizer, L. Zhou, and A. Sears, "Automated stress detection using keystroke and linguistic features: An exploratory study," *Int. J. Hum.-Comput. Stud.*, vol. 67, no. 10, pp. 870–886, Oct. 2009, doi: 10.1016/j.ijhcs.2009.07.005.

- [4] T. Katerina and P. Nicolaos, "Mouse behavioral patterns and keystroke dynamics in End-User Development: What can they tell us about users' behavioral attributes?," *Comput. Hum. Behav.*, vol. 83, pp. 288–305, Jun. 2018, doi: 10.1016/j.chb.2018.02.012.
- [5] I. A. Khan, W.-P. Brinkman, N. Fine, and R. M. Hierons, "Measuring personality from keyboard and mouse use," in *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*, Funchal Portugal: ACM, Jan. 2008, pp. 1–8. doi: 10.1145/1473018.1473066.
- [6] J. Rhim, J. Kim, and G. Gweon, "Using Geometric Features of Drag-and-Drop Trajectories to Understand Students' Learning," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 2023)*, New York: Assoc Computing Machinery, 2023, p. 706. doi: 10.1145/3544548.3581143.
- [7] V. Pandey, N. C. Khan, A. S. Gupta, and K. Z. Gajos, "Accuracy and Reliability of At-Home Quantification of Motor Impairments Using a Computer-Based Pointing Task with Children with Ataxia-Telangiectasia," *ACM Trans. Access. Comput.*, vol. 16, no. 1, pp. 1–25, Mar. 2023, doi: 10.1145/3581790.
- [8] M. Karim and Md. Hasanuzzaman, "A Study on Mouse Movement Features to Identify User," *Sci. Res. J.*, vol. 08, no. 04, pp. 77–82, Apr. 2020, doi: 10.31364/SCIRJ/v8.i4.2020.P0420766.
- [9] P. Kratky and D. Chuda, "Estimating Gender and Age of Web Page Visitors from the Way They Use Their Mouse," in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, Montreal, Quebec, Canada: ACM Press, 2016, pp. 61–62. doi: 10.1145/2872518.2889384.
- [10] B. Pariente-Martinez, M. Gonzalez-Rodriguez, D. Fernandez-Lanvin, and J. De Andres-Suarez, "Measuring the role of age in user performance during interaction with computers," *Univers. Access Inf. Soc.*, vol. 15, no. 2, pp. 237–247, Jun. 2016, doi: 10.1007/s10209-014-0388-6.
- [11] F. Monroe and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Gener. Comput. Syst.*, vol. 16, no. 4, pp. 351–359, Feb. 2000, doi: 10.1016/S0167-739X(99)00059-X.
- [12] D. U. Wulff, P. J. Kieslich, F. Henninger, J. M. B. Haslbeck, and M. Schulte-Mecklenbeck, "Movement tracking of psychological processes: A tutorial using mousetrap," *Behav. Res. Methods*, vol. 57, no. 11, p. 307, Oct. 2025, doi: 10.3758/s13428-025-02695-2.
- [13] Hotjar Ltd., *Hotjar*. (2026). [Online]. Available: <https://www.hotjar.com>
- [14] Mouseflow ApS, *Mouseflow*. (2026). [Online]. Available: <https://mouseflow.com>
- [15] Microsoft Corporation, *Microsoft Clarity*. (2026). [Online]. Available: <https://microsoft.com>
- [16] Human Communication Interaction Group, *pywib: Human Communication Interaction*. (2026). [Online]. Available: <https://humancommunicationinteraction.github.io/pywib>
- [17] P. Khanna and M. Sasikumar, "Recognising Emotions from Keyboard Stroke Pattern," *Int. J. Comput. Appl.*, vol. 11, no. 9, pp. 1–5, Dec. 2010, doi: 10.5120/1614-2170.
- [18] P. J. Kieslich, F. Henninger, D. U. Wulff, J. M. B. Haslbeck, and M. Schulte-Mecklenbeck, "Mouse-Tracking: A Practical Guide to Implementation and Analysis 1," in *A Handbook of Process Tracing Methods*, 2nd ed., Routledge, 2019.
- [19] G. Muthumari, R. Shenbagaraj, and M. Blessa Binolin Pepsi, "Mouse gesture based authentication using machine learning algorithm," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, India: IEEE, May 2014, pp. 492–496. doi: 10.1109/ICACCCT.2014.7019492.
- [20] A. Seelye et al., "Computer mouse movement patterns: A potential marker of mild cognitive impairment," *Alzheimers Dement. Amst. Neth.*, vol. 1, no. 4, pp. 472–480, Dec. 2015, doi: 10.1016/j.dadm.2015.09.006.
- [21] A. Strzelecki and A. Miklosik, "Device-dependent click-through rate estimation in Google organic search results based on clicks and impressions data," *Aslib J. Inf. Manag.*, vol. 77, no. 3, pp. 513–529, Apr. 2025, doi: 10.1108/AJIM-04-2023-0107.
- [22] C. Epp, M. Lippold, and R. L. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver BC Canada: ACM, May 2011, pp. 715–724. doi: 10.1145/1978942.1979046.
- [23] M. Schoemann, M. Lüken, T. Grage, P. J. Kieslich, and S. Scherbaum, "Validating mouse-tracking: How design factors influence action dynamics in intertemporal decision making," *Behav. Res. Methods*, vol. 51, no. 5, pp. 2356–2377, Oct. 2019, doi: 10.3758/s13428-018-1179-4.
- [24] E. Kuric, P. Demcak, M. Krajcovic, and P. Nemcek, "Is mouse dynamics information credible for user behavior research? An empirical investigation," *Comput. Stand. Interfaces*, vol. 90, p. 103849, Aug. 2024, doi: 10.1016/j.csi.2024.103849.
- [25] J.-L. Boulanger, "Technique to Manage Software Safety," in *Certifiable Software Applications 1*, Elsevier, 2016, pp. 125–156. doi: 10.1016/B978-1-78548-117-8.50006-4.
- [26] D. Fernández-Lanvin, J. De Andrés, M. González-Rodríguez, and P. Torre, "Interaction Lab: Web User Interaction Tracking and Analysis Tool," in *Proceedings of the 18th International Conference on Web Information Systems and Technologies*, Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2022, pp. 152–158. doi: 10.5220/0011567100003318.
- [27] K. Gajos, K. Reinecke, and C. Herrmann, "Accurate measurements of pointing performance from in situ observations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin Texas USA: ACM, May 2012, pp. 3157–3166. doi: 10.1145/2207676.2208733.

Engagement-Aware Response Generation for Educational Robots Using Grounded LLM Prompting

Nayanika Reddy Rajoli, Vennela Akula, Daniela Marghitu

Department of Computer Science and Software Engineering

Auburn University

Auburn, Alabama, USA

e-mail: nzh0054@auburn.edu, vza0024@auburn.edu, marghda@auburn.edu

Abstract—Educational robots could make learning more personal, but most current systems use pre-written responses that do not adjust to how a student is engaging or responding. Theory of Mind is the ability to infer what someone is paying attention to, confused by, or trying to do from observable behavior, and it is a necessary foundation for robots that tutor effectively in social settings. This paper presents a work-in-progress multimodal pipeline that implements a simplified, behavior-based approximation of Theory of Mind for an educational robot by observing a user’s behavioral signals in real time and generating context-aware verbal responses through a small language model. The pipeline is designed to be configurable across domains and audiences, and is evaluated here on a single mathematics-quiz scenario: it uses MediaPipe for face detection, gaze estimation, head pose tracking, and hand-raise recognition, aggregates these signals into an engagement state through a priority-ordered trigger system, and constructs observation-grounded prompts for the model. We compare two prompt design strategies: the Grounded Mode, which provides both the sensor observation dictionary and an explicit trigger-type label, and the Ungrounded Mode, which provides the observation dictionary alone. The pipeline runs on both a desktop PC and a Raspberry Pi 5, with the model served from a remote graphics processing endpoint. This work falls under the conference Robotics track, specifically cognitive robotics and mobile assistive robots, combining real-time multimodal sensing with language-model reasoning on resource-constrained hardware. Preliminary results show that the Grounded Mode achieves 86.1% response relevance compared to 23.3% for the Ungrounded Mode, demonstrating that small language models require an explicit semantic-translation layer between numeric sensor output and the prompt to produce contextually appropriate responses.

Keywords—theory of mind; educational robotics; engagement detection; large language models; multimodal sensing.

I. INTRODUCTION

Educational robots are increasingly being used as tutoring companions, but one major problem that still exists is that educational robots are still not able to perceive and respond to the mental state of a user during a session. A user who looks away, raises a hand, or stops responding may be signaling disengagement, confusion, or a desire to ask a question. The ability to perceive and respond to a user’s mental state is something that a human tutor can accomplish effortlessly through a cognitive ability known as Theory of Mind (ToM) [1]. These behavioral cues are proxies: they are observable signals rather than direct measurements of an internal state. Robots that lack this capacity default to scripted interactions that cannot adapt to the learner’s evolving needs.

Recent advances in Large Language Models (LLMs) offer a path toward more adaptive robot behavior. LLMs can generate contextually relevant natural language responses when provided with appropriate situational context [2]. However, integrating LLMs into real-time robotic systems requires solving two problems simultaneously: perceiving the user’s behavioral state through sensors, and constructing prompts that convey that state to the LLM in a form it can act upon.

This paper addresses both problems. The system uses MediaPipe [3] for real-time face detection, gaze estimation, head pose tracking, and hand-raise detection via pose landmarks. These sensor signals are processed by an engagement manager that detects five types of disengagement in priority order: `hand_raised`, `face_absent`, `no_answer`, `head_turned`, and `no_movement`. When a trigger is detected, the current sensor readings are combined into a prompt and sent to a Phi-3-mini LLM (3.8B parameters) [4], which generates a short spoken response to re-engage the user or answer their question. While the pipeline architecture is intended to be configurable across domains and audiences, this work evaluates it only on a mathematics quiz. The prompts and quiz content can be adjusted for other target populations in future studies.

We evaluate the pipeline using an adaptive mathematics quiz for children aged 7 to 14 as the initial test scenario. The Grounded Mode provides the LLM with both the raw observation dictionary and an explicit trigger-type label (e.g., “The user’s face is not visible” or “The user is looking away”). The Ungrounded Mode provides only the observation dictionary, requiring the LLM to infer the trigger type from sensor values alone. Preliminary results show that the Grounded Mode produces substantially more relevant responses, confirming that small LLMs cannot reliably infer engagement states from raw sensor data without explicit labeling.

The pipeline currently runs on both a desktop PC and a Raspberry Pi 5. Language-model inference is currently offloaded to a remote endpoint rather than run on the Pi itself, so the system is not yet fully self-contained at the edge. On-device inference is part of future work. The target deployment platform is a Marty V2 educational robot [5], which supports a Raspberry Pi as an onboard expansion module. The Raspberry Pi 5 runs Robot Operating System 2 (ROS 2) Jazzy [6], to control the robot’s motors and sensors.

The remainder of this paper is organized as follows. Sec-

tion II reviews related work on ToM in robotics, engagement detection, and LLMs in educational settings. Section III describes the system architecture. Section IV presents the Grounded Mode versus Ungrounded Mode prompt design comparison. Section V reports preliminary results. Section VI concludes the paper and outlines future work.

II. RELATED WORK

Theory of Mind was originally defined by Premack and Woodruff [1] as the ability to impute mental states to oneself and others and to use those attributions to predict behavior. In robotics, ToM has been explored primarily through probabilistic models, neural approaches, and reinforcement learning frameworks. Scassellati [7] proposed one of the earliest architectures for endowing a humanoid robot with ToM, focusing on joint attention and perspective-taking. More recently, Chen et al. [8] demonstrated that an observer robot could model another agent's behavior through visual processing alone, achieving 98.5% success in predicting future actions. Hellou et al. [9] surveyed the field and identified three main classes of computational ToM: probabilistic models using Bayesian inference, neural network approaches, and hybrid methods. They noted that implementations in Human-Robot Interaction (HRI) remain limited, particularly for educational contexts.

Engagement detection in educational settings has been studied using various sensor modalities. Wang et al. [10] developed a MediaPipe-based system for detecting behavioral and emotional engagement in e-learning environments, fusing features from gaze direction, head pose, and facial expressions. Their work demonstrated that MediaPipe provides a low-cost, low-latency framework suitable for real-time deployment on edge devices.

The integration of LLMs into social and educational robots is an active area of research. Studies have shown that LLM-powered robots can maintain open-domain interactions across multiple sessions [11], and that cloud-based LLMs significantly outperform scripted approaches in interaction quality and user satisfaction [12]. In the domain of child-robot interaction, LLMs have been fine-tuned to exhibit creative storytelling behavior with positive effects on engagement [13]. However, existing work has not combined real-time multi-modal engagement sensing with observation-grounded LLM response generation in a single educational robot pipeline. This paper addresses that gap by building a pipeline that perceives behavioral cues, infers an engagement state, and generates contextually appropriate verbal responses through an LLM, implementing a simplified form of ToM.

III. SYSTEM ARCHITECTURE

The pipeline consists of seven modules that operate in a continuous loop during a tutoring session. Figure 1 provides an overview of the architecture. The architecture is designed to separate sensing, engagement detection, and response generation from the specific quiz content, although this paper evaluates only the mathematics-quiz configuration. For the

current evaluation, the system is configured to run an adaptive mathematics quiz targeting children aged 7 to 14.

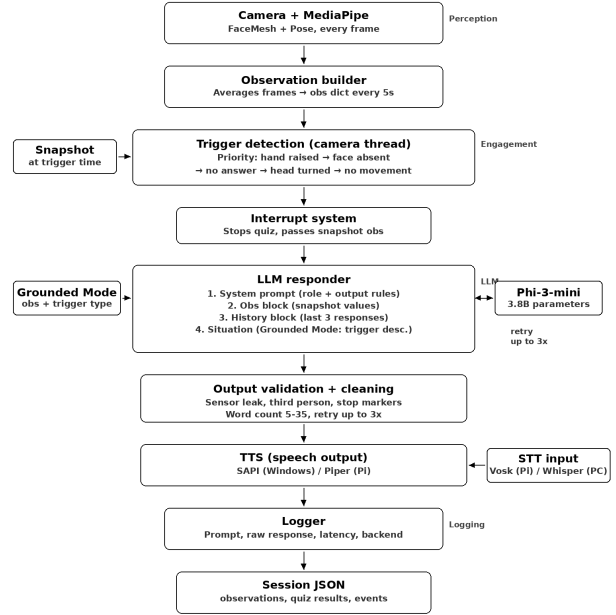


Figure 1. Pipeline architecture showing data flow from camera input through perception, engagement detection, LLM response generation, output validation, speech synthesis, and session logging.

A. Perception Layer

The MediaPipe Processor module captures video frames from a webcam and runs two MediaPipe task models in a background thread: FaceMesh for facial landmark detection and Pose Landmarker Lite for body and hand tracking [3]. Frames are processed continuously, and an observation dictionary is built by averaging sensor readings over a 5-second interval. The dictionary includes the following fields: `face_detected` (boolean), `gaze_on_robot` (0.0 to 1.0, estimated from head yaw as a proxy rather than true eye tracking), `head_yaw_deg` and `head_pitch_deg` (head orientation in degrees), `head_moving` (boolean, based on a 2-degree threshold), `body_detected` (boolean), `hand_raised` and `hand_raise_side` (detected from pose landmarks when a wrist is above shoulder level), `no_movement_sec` (seconds since last detected movement), and `speech_energy`. Several of these fields are deliberately coarse proxies for the underlying behavior, so the engagement states derived from them should be read as behavioral approximations rather than confirmed mental states. The Speech Input module uses Whisper on PC and Vosk on Raspberry Pi for speech-to-text (STT) transcription, with automatic platform detection.

B. Adaptive Quiz Engine

The current test scenario involves a quiz manager with 90 questions in mathematics for different levels of difficulty.

These levels are defined as Level 1 for easy questions, Level 2 for medium questions, and Level 3 for hard questions for ages between 7 and 14 years old. The difficulty level increases or decreases based on whether there are three consecutive correct or incorrect answers. The quiz context, including the current question, topic, difficulty level, and user's previous answers, is provided as input to the LLM prompt so that responses are grounded in the current learning context. The quiz module is replaceable according to different domains or ages.

C. Engagement Manager

The Engagement Manager aggregates sensor readings from the MediaPipe Processor into discrete engagement events. It monitors five trigger conditions in a hardcoded priority order. (1) `hand_raised`, the user's hand is detected above shoulder level. (2) `face_absent`, no face detected for a configurable duration (default 3 seconds). (3) `no_answer`, the user has not responded to a quiz question within a timeout period (default 15 seconds). (4) `head_turned`, the gaze score falls below a threshold (default 0.3 out of 1.0), and (5) `no_movement`, no body movement detected for a configurable duration (default 10 seconds). The priority order is a deliberate design decision: a hand raise always takes precedence over passive disengagement signals because it represents an active intent to communicate.

When the trigger fires, the Engagement Manager captures a snapshot of the current state of the sensor as an observation dictionary. This ensures the observation reflects conditions at the time the trigger occurred, rather than at the time the LLM responds, which could be several seconds later. A 15-second cooldown period prevents the same trigger from being repeated. The trigger priority order, the 5-second observation window, and the timeout, threshold, and cooldown values are untuned engineering defaults chosen from pilot observation. A systematic ablation of these choices is left for future work.

D. LLM Responder

The LLM Responder constructs a structured prompt from four components: (1) a system prompt defining the robot's role and strict output rules, (2) the observation snapshot with current sensor values, (3) a history block containing the last three LLM responses with an instruction not to repeat them, and (4) in the Grounded Mode, a plain-English situation description of the trigger type. The complete prompt is sent to a Phi-3-mini (3.8B parameter) LLM [4], served from a remote endpoint that is accessible from both the PC and the Raspberry Pi. The module enforces a word count between 5 and 35 words and validates the output to reject responses that contain raw sensor field names (e.g., `gaze_ratio` or `head_yaw`), which would be meaningless to the user. If a response fails validation, the module retries up to three times with increasing temperature values (0.7, 0.9, 1.1). If all retries fail, a pre-written fallback phrase specific to the trigger type is used instead.

E. Speech Output and Logging

The Speech Output module converts the LLM-generated response to speech using platform-specific text-to-speech (TTS) engines: Windows Speech API (SAPI) on PC and Piper TTS on Raspberry Pi. A stop method allows mid-speech interruption when a hand raise is detected during playback. The Logger module records all session data into a structured JSON file, including every quiz interaction, sensor reading, engagement trigger, the full prompt sent to the LLM, the raw LLM response, retry logs, response timing, and the LLM backend identifier for post-session analysis.

F. Privacy and Ethical Considerations

Because the intended end users are children and the system uses a webcam, a microphone, and session logging, privacy and safeguarding are first-order design concerns. Camera frames are processed in memory and are not stored. Only the derived observation dictionary, the transcribed text, and the generated responses are written to the session log, and no raw video or audio is retained. All testing reported in this paper used consenting adult volunteers, and no minors have participated. Any future study with children will proceed only under an approved Institutional Review Board (IRB) protocol, with informed parental consent and child assent, defined retention limits for the JSON logs, and the option to delete a session on request. We also acknowledge that face and pose detection can exhibit accuracy bias across skin tones, lighting, and seating geometry, which could systematically disadvantage some users. Auditing the perception layer for such bias is part of planned work. For safeguarding, the output validation layer (Section III-D) and the trigger-specific fallback phrases bound the range of responses the robot can speak, and the system neither requests nor stores personally identifying information.

IV. PROMPT DESIGN: GROUNDED MODE VS UNGROUNDED MODE

A central design question is how much contextual information the LLM needs to generate appropriate responses. We evaluated two prompt strategies using the Phi-3-mini model:

Grounded Mode (Observation Dictionary and Trigger Type): The prompt includes both the sensor observation snapshot and an explicit plain-English trigger description (e.g., "The user's face is not visible to the camera during the math quiz. Mention you can't see their face and ask them to come back."). This provides the LLM with both low-level sensor data and a human-interpretable summary of the engagement state.

Ungrounded Mode (Observation Dictionary Only): The prompt includes only the sensor observation snapshot without any trigger description. The LLM must infer the user's engagement state from the numerical sensor values alone.

Both modes include identical quiz context (current question, topic, user age), identical behavioral constraints (5 to 35 word limit, no sensor values, audience-appropriate language), and the same history block of previous responses. The comparison

isolates the effect of explicit trigger labeling on response relevance. By construction, the Grounded Mode receives information the Ungrounded Mode does not, so the comparison is not a like-for-like contest between equally informed systems. It is a measurement of how much the explicit label contributes to relevance.

We conducted a preliminary manual relevance review of all LLM responses across both modes. Each response was labeled relevant if it addressed the correct trigger context, and not relevant otherwise, following a written rubric. A single author performed the relevance scoring against this rubric. Independent scoring by multiple annotators, with a measure of inter-rater agreement, is planned as part of future validation. This review measures only whether a response matched the trigger context. It does not assess pedagogical quality, tone, or actual re-engagement effect.

TABLE I. GROUNDED MODE VS UNGROUNDED MODE RESPONSE RELEVANCE BY TRIGGER TYPE (MANUAL REVIEW).

Trigger Type	Grounded		Ungrounded	
	R / Total	%	R / Total	%
Hand raised	11 / 13	84.6	1 / 14	7.1
Face absent	11 / 13	84.6	0 / 12	0.0
No answer	9 / 11	81.8	3 / 10	30.0
No movement	14 / 16	87.5	9 / 19	47.4
Head turned	17 / 19	89.5	4 / 18	22.2
Total	62 / 72	86.1	17 / 73	23.3

R = relevant responses scored by manual review.

The difference is substantial. In the Ungrounded Mode, the LLM frequently misidentified the engagement trigger. For example, when a user's face was absent from the frame, Ungrounded Mode responses sometimes addressed gaze direction or encouraged continued focus, rather than acknowledging the user's absence. When a user raised their hand, the Ungrounded Mode could not connect the boolean flag `hand_raised: yes` to the concept of a user wanting to ask a question, because the model's training data contains no mapping between that field name and the associated intent. Grounded Mode responses consistently matched the trigger type because the explicit plain-English description removed the inference burden from the LLM.

This finding is consistent with known limitations of LLMs in ToM reasoning. Ullman [14] demonstrated that even large LLMs fail on trivial alterations to standard Theory of Mind tasks, suggesting that a 3.8B parameter model cannot be expected to infer engagement states from raw sensor values without explicit guidance. More broadly, this result is an empirical confirmation that an explicit semantic-translation layer is required between numeric sensor output and an LLM: the model behaves as a semantic reasoning engine rather than a numeric or boolean classifier, so feeding it raw sensor arrays bypasses its core strengths. The 86.1% relevance of the Grounded Mode is achieved precisely because the plain-English trigger description supplies that semantic abstraction.

For our pipeline, the Grounded Mode is therefore the required configuration.

V. PRELIMINARY RESULTS

We conducted 8 test sessions with consenting adult volunteers using the Grounded Mode on both a desktop PC (Windows) and a Raspberry Pi 5 (Ubuntu 24.04, 8 GB RAM, Vosk for speech-to-text, Piper for text-to-speech). Table II compares platform-level performance.

TABLE II. PC VS RASPBERRY PI LATENCY (GROUNDED MODE, AVERAGED ACROSS SESSIONS).

Metric	PC	Raspberry Pi 5
Avg TTS Question (s)	3.91	5.44
Avg TTS Feedback (s)	1.72	3.29
Avg STT (s)	0.93 (Whisper)	0.08 (Vosk)
Avg LLM Response (s)	0.89	0.54

LLM latency weighted across 42 (PC) and 27 (Pi) events.

The pipeline achieved zero observation mismatches and zero sensor field leaks across all test sessions. This confirms that the snapshot approach (Section III-C) and the output validation layer (Section III-D) function correctly on both hardware configurations.

These results should be read as a small pilot rather than a controlled study. The relevance review covers 72 grounded and 73 ungrounded responses, and the latency figures aggregate 42 PC and 27 Pi trigger events. Responses within a single session are not fully independent, since they share the same user and learning context. The trigger-specific fallback phrases (Section III-D) also act as an implicit templated baseline, but a controlled comparison against a full templated-response system and against a cloud-hosted LLM is left for future work (Section VI).

VI. CONCLUSION AND FUTURE WORK

This paper presented a multimodal pipeline that implements a simplified, behavior-based approximation of Theory of Mind for educational robots by combining real-time engagement sensing with observation-grounded LLM response generation. The Grounded Mode vs Ungrounded Mode comparison confirms that small LLMs require explicit trigger labeling, achieving 86.1% relevance compared to 23.3% without it, which empirically validates the need for a semantic-translation layer between sensor output and the model. Although the architecture is configurable for other domains and audiences, this paper evaluates it only in a mathematics-quiz scenario for children aged 7 to 14. Therefore, such generalization remains to be demonstrated beyond the present scenario.

Several important challenges remain unresolved. First, the quality and pedagogical appropriateness of LLM responses have not been formally validated. While a manual relevance review confirmed that Grounded Mode responses address the correct trigger context at 86.1%, this does not assess whether responses are educationally effective, appropriate in tone, or helpful for re-engaging the user. Four validation approaches

are planned. The first is LLM-as-judge evaluation, using a larger LLM to score responses against pedagogical rubrics. The second is expert review by education specialists. The third is IRB-approved studies with participants in the target age range to measure actual re-engagement effects. The fourth is controlled baseline comparisons against a templated-response system and a cloud-hosted LLM, to quantify how much response quality the local small model gives up.

Second, the current system relies on a remote graphics processing unit (GPU) endpoint for LLM inference. Moving to on-device inference on the Raspberry Pi 5, potentially using a neural processing unit accelerator, such as the Hailo HAT+2, would eliminate network dependency and enable fully offline operation, which is important for classroom deployments with limited connectivity.

Third, the pipeline currently operates without a physical robot. The target platform, a Marty V2 robot [5] connected via ROS 2 Jazzy [6], will provide physical embodiment through walking and gesturing synchronized with verbal responses. Physical embodiment has been shown to increase engagement and trust in human-robot interaction [13].

Fourth, the current ToM model only detects behavioral engagement states, such as attentive, distracted, absent, and questioning. Adding the ability to infer emotional states, such as frustration, boredom, or confusion, through facial expression analysis and speech tone would allow the system to generate more targeted responses. For example, a frustrated student needs a different response than a bored student, even if both show similar gaze patterns. This would move the system from detecting where the user is looking to approximating how the user is feeling.

REFERENCES

- [1] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978.
- [2] J. Wang *et al.*, "Large language models for robotics: Opportunities, challenges, and perspectives," 2024. arXiv: 2401.04334 [cs.RO].
- [3] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," 2019. arXiv: 1906.08172 [cs.DC].
- [4] M. Abidin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," 2024. arXiv: 2404.14219 [cs.CL].
- [5] Robotical Ltd., "Marty the robot V2," 2026, [Online]. Available: <https://robotical.io/product/marty-the-robot-v2/> (visited on 03/31/2026).
- [6] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, eabm6074, 2022.
- [7] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.
- [8] B. Chen, C. Vondrick, and H. Lipson, "Visual behavior modelling for robotic theory of mind," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.
- [9] M. Hellou, S. Vinanzi, and A. Cangelosi, "A review of theory of mind and robotics: Mind reading in human-robot interaction for proactive social robots," in *Human-Friendly Robotics 2024 (HFR 2024)*, ser. Springer Proceedings in Advanced Robotics, vol. 35, Springer, 2025, pp. 197–211. DOI: 10.1007/978-3-031-81688-8_15.
- [10] J. Wang, S. Yuan, T. Lu, H. Zhao, and Y. Zhao, "Video-based real-time monitoring of engagement in e-learning using MediaPipe through multi-feature analysis," *Expert Systems with Applications*, vol. 267, p. 128 239, 2025. DOI: 10.1016/j.eswa.2025.128239.
- [11] M. Mauliana, A. Ashok, D. Czernochowski, and K. Berns, "Exploring LLM-powered multi-session human-robot interactions with university students," *Frontiers in Robotics and AI*, vol. 12, p. 1 585 589, 2025. DOI: 10.3389/frobt.2025.1585589.
- [12] K. Smit, S. Leewis, H. Almoustafa, K. Yildirim, and T. Uymaz, "Enhancing educational dynamics: Integrating large language models with a social robot," in *Proceedings of the 2024 8th International Conference on Software and e-Business (ICSeB '24)*, ACM, 2024, pp. 87–94. DOI: 10.1145/3715885.3715889.
- [13] M. Elgarf, H. Salam, and C. Peters, "Fostering children's creativity through LLM-driven storytelling with a social robot," *Frontiers in Robotics and AI*, vol. 11, p. 1 457 429, 2024. DOI: 10.3389/frobt.2024.1457429.
- [14] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," 2023. arXiv: 2302.08399 [cs.CL].

Parkinson's Disease Severity Assessment on the Unified Parkinson's Disease Rating Scale Using Text Embeddings Generated from Speech

Christina Koumpogianni and Constantine Kotropoulos
 Department of Informatics, Aristotle University of Thessaloniki
 Thessaloniki 54124, Greece
 email: {ckoumpo, costas}@csd.auth.gr

Abstract—Parkinson's disease (PD) is a movement disorder of the central nervous system that affects patients' mobility over time. PD can be assessed using the Unified Parkinson's Disease Rating Scale (UPDRS), a 50-item, 4-part assessment used by clinicians and researchers to evaluate the severity and progression of Parkinson's disease. It measures motor, non-motor, cognitive, and activities-of-daily-living functions, with higher scores indicating greater impairment; it is commonly used to guide treatment decisions. The paper aims to predict the UPDRS using a complete pipeline that integrates a hybrid audio-to-text approach, in which speech signals are processed with deep learning models for transcription and embedding generation, followed by a Support Vector Regression model for prediction. The experimental results show that this hybrid model is a sufficiently accurate tool for UPDRS prediction.

Keywords—Parkinson's disease; Natural language processing; Transformer-based embeddings; Self-supervised learning; Regression.

I. INTRODUCTION

Parkinson's Disease (PD) is a neurological disorder caused by a combination of genetic and environmental factors. Although no definitive cure currently exists, available treatments focus on symptom management and are primarily effective in the early stages of the disease. A significant proportion of individuals with PD experience dysarthria, with reported prevalence ranging from 70% to 90% [1].

In this paper, patients with PD who exhibit dysarthria are assigned their corresponding Unified Parkinson's Disease Rating Scale (UPDRS) scores, and the objective is to predict these scores as accurately as possible using deep learning-based audio-to-text embeddings combined with a machine-learning regression model. The main contributions are the design of an end-to-end pipeline that converts pathological speech into semantic representations using a pre-trained speech recognition and transcription model, followed by deep learning-based sentence-embedding extraction. These embeddings are then used as input to the regression model to predict UPDRS scores. This approach moves beyond traditional handcrafted acoustic features by capturing higher-level linguistic and contextual information, while enabling a more robust evaluation of the relationship between speech impairments and clinical severity in PD. Recent studies have addressed the same regression task primarily using Support Vector Regression (SVR), audio-to-text embeddings, and handcrafted feature vectors [2].

A. Related Work

Speech analysis has been a topic of interest to many scientists and clinicians. The paper aims to automate PD screening

using text embeddings generated from speech. A large percentage of patients with PD experience dysarthria. Prior studies have explored both acoustic and linguistic representations of speech for PD detection and severity estimation. Several studies have addressed this task by extracting acoustic features from voice signals. Speech features are studied that are related to how sounds are produced and articulated in both sustained vowels and continuous speech [2]. These features can reliably reflect speech-motor impairments associated with PD. In a similar line of research, non-negative matrix factorization [3] is used to extract time-frequency features from voice signals, thereby capturing patterns associated with vocal degradation in PD. These manually designed acoustic features are typically used with traditional machine learning models, such as Support Vector Machines (SVMs) or regression methods, to estimate the presence or severity of the disease.

However, most of these studies focus on PD detection rather than fine-grained severity prediction. The prediction of the modified Frenchay dysarthria level assessment (m-FDA) was addressed in [4], [5]. The Hilbert Cepstral Coefficients were used as input features in three regression models: Radial Basis Function (RBF) SVR, Random Forest Regressor [6], and Multilayer Perceptron Regressor [7]. Performance was evaluated by computing the Spearman correlation between the predicted and ground-truth m-FDA scores. This paper focuses on transcribing audio into text using a transformer [8]-based pretrained model, the Whisper [9], which is a large-scale pretrained model trained on multilingual and multitask data, demonstrating strong robustness across diverse speech conditions. Subsequently, textual representations are encoded with a transformer-based sentence-embedding model (e.g., mixedbread-ai/mxbai-embed-large-v1 [10]), which maps transcripts to high-dimensional semantic vectors. The final prediction in UPDRS is performed using SVR.

The remainder of the paper is organized as follows. Section II describes the dataset, data preparation, audio-to-text transcription, text embedding, regression, and UPDRS prediction. Section III presents the experimental results. Section IV concludes the paper and outlines directions for future research.

II. METHODOLOGY

A synopsis of the proposed pipeline is illustrated in Figure 1. The PC-GITA speech dataset [11] was loaded, and its meta-data were extracted for each .wav recording. The clinical meta-data from the provided Excel spreadsheet was merged with

the audio index using speaker identifiers, retaining only PD patients with complete information. Audio preprocessing included mono conversion, resampling to 16 kHz for compatibility with the Whisper model's input, amplitude normalization, and segmentation into overlapping 30-second chunks with a 5-second overlap. Speech transcriptions were generated in Spanish using the Whisper model [9]. Sentence-level semantic embeddings were subsequently extracted from the transcriptions using the `mixedbread-ai/mxbai-embed-large-v1` transformer model [10]. For each speaker, utterance embeddings were aggregated using the median representation and ℓ_2 -normalized to obtain a robust speaker-level embedding vector. The PD severity prediction was performed using SVR within a pipeline consisting of standardization, Principal Component Analysis (PCA), and hyperparameter optimization via grid search. Model evaluation was performed using Leave-One-Out Cross-Validation (LOOCV), and performance was assessed using the Root Mean Squared Error (RMSE).

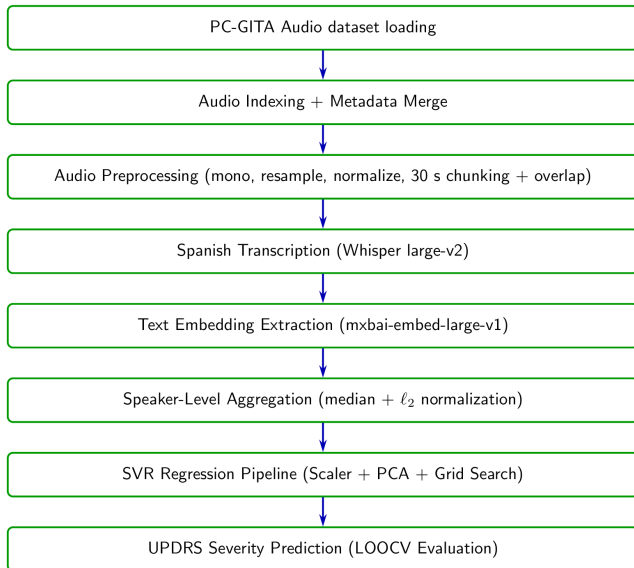


Figure 1. Proposed pipeline.

A. Dataset

Experiments have been conducted using the PC-GITA dataset [11], a Spanish-language balanced speech dataset used in PD research. It includes 50 speakers with PD and 50 healthy control (HC) speakers, comprising a total of 6405 recordings collected under controlled recording conditions. Specifically, the dataset contains 3214 raw recordings from speakers with PD and 3200 raw recordings from HC speakers. Each participant performed a standardized set of speech tasks designed to elicit a wide range of vocal behavior. Each participant's recording includes the following characteristics: unique ID, UPDRS_total, UPDRS_speech, Hoehn and Yahr scale, gender, age, and Time since diagnosis. The Hoehn and Yahr scale is a clinical staging scale used to describe the severity of PD. In the PC GITA dataset, this score is included as a clinical

label for each patient, indicating disease stage. It describes the five stages of PD's symptoms, i.e., Stage 1: Symptoms on one side of the body only; Stage 2: Symptoms on both sides with no balance problems; Stage 3: Balance impairment but physically independent; Stage 4: Severe disability with the ability to walk or stand with assistance; and Stage 5: Wheelchair-bound or bedridden unless assisted [12]. The Time since diagnosis refers to the number of years since each participant was diagnosed with PD. UPDRS_total refers to the overall UPDRS score that clinicians have assigned based on a full clinical assessment of disease severity. UPDRS_speech is a sub-score within the UPDRS scale that focuses only on speech impairment. This sub-score typically ranges on a much smaller scale (e.g., 0–4), reflecting the severity of speech impairment due to PD. In the paper, only the UPDRS_total is used, which ranges from 6 (mild motor impairment) to 93 (severe disease progression). The PC-GITA dataset comprises a variety of speech tasks, including sustained vowel phonation, diadochokinetic (DDK) tasks, spontaneous monologue, and text reading, which together capture different aspects of dysarthria. Importantly, PD speakers are annotated with UPDRS scores, allowing for conducting a regression task. That is, to make a reasonable prediction of the UPDRS score using an automated method for assessing disease severity in patients. Since multiple recordings are available for each speaker, their corresponding embeddings are aggregated into a single representative feature vector by taking the median across all of these embeddings. However, the dataset remains relatively small and balanced, which will limit generalization.

B. Data Preprocessing

The PC-GITA dataset is first indexed by recursively scanning the directory structure and collecting all available 6405 audio files in Waveform Audio File Format (.wav) along with their corresponding task labels, speaker identifiers, and file paths, to create a structured CSV catalog of all .wav files. Speaker-level pathological condition labels (PD or HC) are inferred from folder and file naming conventions and later validated using the provided metadata. Clinical information, including total UPDRS score, Hoehn and Yahr scale, age, gender, and time since diagnosis, is extracted from the original PC-GITA metadata and merged with the audio file index using speaker identifiers. After merging with speaker-level metadata and filtering for complete information (UPDRS, age, gender, Hoehn-Yahr scale, and time since diagnosis), only 50 PD speakers retained all characteristics without any missing data, whereas all HC speakers were excluded. The .wav files associated with the 50 PD speakers total 1099. Samples with missing or inconsistent clinical information were excluded to maintain data quality. This accounts for the reduction in size in the PD and HC subsets. As shown in Table I, the dataset is indexed by task, pathological condition, and speaker. Each entry in the final dataset corresponds to a single speech recording produced by a specific speaker during a given task. For example, in the DDK analysis task, recordings consist of rapid syllable repetitions (e.g., "ka-ka-ka"), which are

TABLE I. EXAMPLE OF DATA STRUCTURE AFTER METADATE MERGING.

Task	Label	Speaker_ID	Relative Path	Full Path	UPDRS_total	UPDRS_Speech	Hoehn-Yahr scale	Gender	Age	Time since diagnosis
DDK analysis	PD	AVPEPUDEA0053	DDK analysis/ka-ka-ka/...	/content/drive/...	33.0	2.0	2.0	M	47.0	2.0
DDK analysis	PD	AVPEPUDEA0052	DDK analysis/ka-ka-ka/...	/content/drive/...	23.0	1.0	2.0	F	70.0	12.0
DDK analysis	PD	AVPEPUDEA0042	DDK analysis/ka-ka-ka/...	/content/drive/...	54.0	1.0	3.0	F	65.0	8.0
DDK analysis	PD	AVPEPUDEA0023	DDK analysis/ka-ka-ka/...	/content/drive/...	14.0	0.0	1.0	M	68.0	1.0
DDK analysis	PD	AVPEPUDEA0047	DDK analysis/ka-ka-ka/...	/content/drive/...	40.0	1.0	2.0	F	64.0	3.0

known to reflect motor control and articulatory precision. This structure allows multiple recordings from the same speaker to be linked to a single clinical profile, facilitating speaker-level aggregation. The UPDRS scores in the dataset range from 6 (mild motor impairment) to 93 (severe disease progression). This broad distribution of scores supports regression analysis rather than coarse disease classification.

C. Audio-to-Text Transcription

First, Whisper-large-v2 [9], a pre-trained transcription model, is loaded. Whisper is a pre-trained model for Automatic Speech Recognition (ASR) and speech translation/transcription. Trained on 680k hours of labeled data, Whisper models demonstrate a strong ability to generalize to many datasets without the need for fine-tuning. The updated large-v2 version was chosen for its improved recognition performance and greater model capacity. The audio .wav recordings are loaded using the torchaudio [13] library and, when necessary, converted to a single channel by averaging across channels. To be compatible with the transcription model, which is designed to operate on 16 kHz audio, all signals are resampled to 16 kHz. The waveform amplitudes are then normalized to reduce variability due to recording conditions. The preprocessed audio waveform is first converted into a one-dimensional array by removing the channel dimension, yielding a continuous sequence of samples representing the entire recording. To facilitate transcription by the Whisper model, the waveform is then segmented into fixed-length chunks. Each chunk corresponds to a specified duration of 30 seconds, which is converted to the equivalent number of samples based on the 16 kHz sampling rate (30×16000 samples). However, after this process, it must be ensured that no recordings are lost or truncated due to 30-second chunking, which is why an overlapping strategy was used. Consecutive chunks are designed to partially overlap for 5 seconds, ensuring that no speech information is lost. Specifically, the starting index of each new chunk is advanced by the chunk length minus the overlap, producing overlapping segments of the original waveform. Each segment is stored as a separate one-dimensional array, forming a list of chunks that collectively represent the full recording while preserving continuity and context for subsequent transcription. The transcription of audio recordings begins by loading metadata containing the file paths of all recordings into a pandas DataFrame. After each audio file is segmented into overlapping chunks, as described previously, the Whisper model performs automatic ASR on each chunk to produce a Spanish transcript. Non-empty transcripts from all chunks of a single recording are concatenated to generate a complete transcript for that file. This procedure is applied iteratively to all recordings. The resulting transcripts are stored

in a new column of the DataFrame and saved to a CSV file, ensuring that the full audio content is preserved and ready for text embedding generation.

D. Text Embedding Generation and Speaker-Level Feature Aggregation

Text embeddings were generated using the pre-trained model mixedbread-ai/mxbai-embed-large-v1 [10] of size 1024 based on a Transformer encoder architecture for generating contextual sentence embeddings. Each transcript was first checked to ensure that it was a non-empty string of sufficient length, and a contextual prefix: “Parkinson speech severity prediction”, was added to guide the embedding model toward the clinical prediction task. Analytically, each token in the transcript interacts with the prefix tokens during self-attention, altering the hidden representations throughout the network. This results in the final 1024-dimensional embedding being slightly biased toward speech patterns associated with severity, such as slurring, pauses, or monotone delivery. Although all recordings are Parkinson-related, the prefix helps highlight subtle variations that are most informative for predicting UPDRS scores, producing embeddings that are more discriminative for downstream regression modeling. The cleaned transcripts were then encoded using this pre-trained model in batches of 16, producing 1024-dimensional dense vectors that capture semantic and contextual information, enabling the integration of linguistic information into machine learning models to predict clinical scores. After generating embeddings for each transcript, speaker-level feature vectors were constructed to represent an individual’s recordings. The dataset was filtered to include only PD patients. For each speaker, all their transcript embeddings were retrieved and aggregated by taking the median across embeddings, yielding a robust representation that reduces the influence of outlier segments. The aggregated vector was then ℓ_2 -normalized, ensuring the resulting feature vector had unit length, thereby standardizing the input scale for downstream models. Each speaker-level embedding was stored in a feature matrix $\mathbf{X}$, and the corresponding UPDRS score was stored in the target vector $\mathbf{y}$. This process resulted in a structured dataset of speaker-level embeddings and labels suitable for regression analysis, with $\mathbf{X}$ having shape (50×1024) and $\mathbf{y}$ of size 50 containing the associated UPDRS scores. Let $t_{i,j}$ denote the j th transcript of speaker i . Each transcript is augmented by appending the task-specific prefix p :

$$\tilde{t}_{i,j} = [p; t_{i,j}]. \quad (1)$$

The embedding model [10] $f_\theta(\cdot)$ converts each input sequence to a 1024-dimensional vector:

$$\mathbf{e}_{i,j} = f_\theta(\tilde{t}_{i,j}) \in \mathbb{R}^{1024}. \quad (2)$$

For each speaker, the embeddings are aggregated component-wise using the median:

$$\mathbf{s}_i = \text{median}(\{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \dots, \mathbf{e}_{i,n_i}\}). \quad (3)$$

The resulting speaker representation is then ℓ_2 -normalized:

$$\hat{\mathbf{s}}_i = \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|_2} \quad (4)$$

and the final dataset is defined as:

$$\mathcal{D} = \{(\hat{\mathbf{s}}_i, y_i)\}_{i=1}^N, \quad N = 50. \quad (5)$$

E. Regression and UPDRS Prediction

To predict the UPDRS units from speaker-level embeddings, SVR was applied within a leave-one-out cross-validation (LOOCV) framework combined with hyperparameter tuning (GridSearchCV). In each fold, one speaker from the 50 PD speakers is held out for testing, while the remaining speakers are used for training, ensuring the model does not see the test embeddings during training. The pipeline for this task included feature standardization via `StandardScaler()` [14] to prevent scale issues in SVR, optional dimensionality reduction via PCA [15], and SVR. Hyperparameters for both PCA and SVR were optimized using a 5-fold grid search on the training set. For PCA, the number of components tested was 5, 10, 20, and 40. For SVR, the grid search performed for choosing the kernel type $\in \{\text{linear, RBF, sigmoid, and polynomial}\}$, the regularization parameter $C \in \{0.1, 1, 10, 100, 200\}$, and the scale parameter $\gamma \in \{10^{-3}, 10^{-2}, 10^{-4}\}$. For the sigmoid kernel, `coef0` values $\in \{0, 0.1, 1\}$ were tested. For the polynomial kernel, degrees 2 and 3 were evaluated using `scikit-learn`. Precisely, inside each LOO iteration, the training set (49 speakers) is further split into 5 folds. The kernels used for this grid search were linear, RBF, and sigmoid. The grid search result corresponding to the best model was used to predict the UPDRS score for the held-out speaker, and the RMSE was recorded. This procedure was repeated for each speaker, resulting in a vector of RMSE scores. The mean RMSE is computed as well as its standard error (SE) defined as $\frac{std}{\sqrt{N}}$ where *std* denotes the standard deviation and *N* is the sample size (i.e., 50). The SE provides confidence in the estimated mean RMSE, i.e., how accurate the estimate is.

III. RESULTS

Table II provides insight into SVR for the linear and the RBF kernels. The average RMSE across all LOOCV folds is expressed in raw UPDRS units. The sigmoid kernel is excluded from Table II, because it was not selected in any fold. Among the evaluated kernels, the RBF kernel achieved the lowest mean RMSE of 14.94 with hyperparameters $n_{\text{components}} = 5$, $C = 100$, and $\gamma = 0.001$, indicating better average predictive performance than the linear kernel, which yielded a mean RMSE of 24.837. The total RMSE was

TABLE II. PERFORMANCE COMPARISON OF SVR KERNELS EXCLUDING THE POLYNOMIAL.

Kernel	Mean RMSE	SE
linear	24.837	10.962
RBF	14.94	1.4851

calculated across all folds and incorporated the predictions from all selected kernels. The total RMSE among all folds was 15.71, the MAE was 11.04, and the R^2 was 0.25. This result suggests that the relationship between the embeddings and UPDRS scores is probably nonlinear, which explains why the RBF kernel performs better.

The lowest mean RMSE was achieved with the RBF kernel (i.e., 14.94), compared to the linear kernel (i.e., 24.837), highlighting its potential to capture complex nonlinear relationships in certain cases. The SE of the RBF (1.4851) is lower than that of the linear kernel (10.962), indicating that the RBF kernel exhibits more stable behavior and reflects its ability to model complex patterns in the embedding space. When the polynomial kernel is excluded from the grid search, as demonstrated in Table II, the overall performance improves, with the RBF kernel achieving the lowest mean RMSE (i.e., 14.94).

When the polynomial kernel was frequently selected across folds during the grid search in inner cross-validation, as shown in Table III, it resulted in worse generalization performance. The polynomial kernel was selected 39 times during the grid search. The mean RMSE achieved was 17.7 from the polynomial kernel only, which is higher than the best mean RMSE of 14.94 from the RBF kernel in Table II. The total RMSE was also higher, at 16.752, with total SE of 1.623 and $R^2 = 0.146$, indicating that it does not capture the linguistic characteristics well. Also, the linear kernel was not selected in this case, not even once. These RMSE values differ because the first refers to the RMSE obtained using only the polynomial kernel. In contrast, the total RMSE/SE is calculated across all folds and incorporates predictions from all selected kernels. Polynomial kernels are highly flexible and can easily overfit small, high-dimensional datasets, especially in clinical datasets with a limited number of speakers. While the polynomial kernel may fit the training folds very closely, this increased complexity often leads to poorer performance on the unseen data. The exclusion of the polynomial kernel reduces the risk of overfitting by limiting the model's flexibility, thereby allowing the RBF kernel to better capture nonlinear relationships in the data.

However, the RBF kernel still exhibits variability, as indicated by its relatively high SE of 1.91, suggesting that performance remains sensitive to the specific training folds. However, the SE of the polynomial kernel was higher, i.e., 2.80, as demonstrated in Table III. A larger standard error indicates greater variability in model performance across cross-validation folds and, therefore, lower confidence in the estimated mean performance. Overall, these results indicate that the RBF kernel yields a smoother nonlinear decision function that captures complex relationships without overfitting noise or speaker-specific variations.

TABLE III. KERNEL PERFORMANCE COMPARISON.

Kernel	Mean RMSE	MAE	SE
polynomial	17.7	13.15	2.80
RBF	12.8558	9.3	1.91

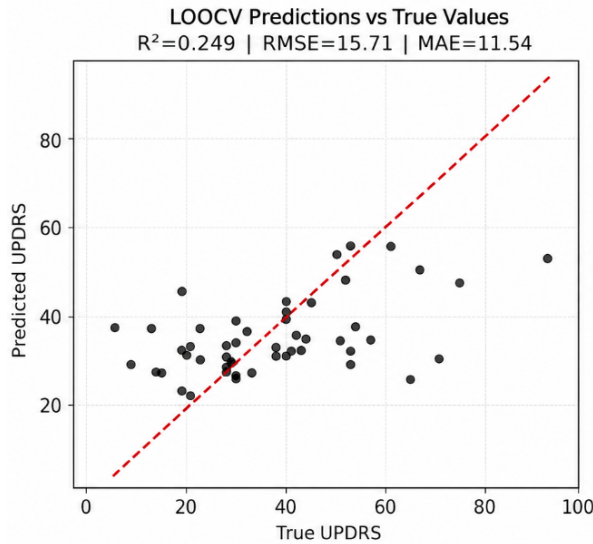


Figure 2. Predicted vs True values of the target variable across all LOOCV folds.

Figure 2 shows predicted versus true target values across LOOCV folds from the grid search procedure, with results summarized in Table II. The points are colored by the SVR kernel used (linear or RBF). The black dashed line represents perfect prediction. Points close to the line indicate accurate predictions, while deviations show errors. The majority of predictions, particularly for the RBF kernel, cluster around this diagonal line, but noticeable deviations indicate prediction errors. Notably, only a single point corresponds to the linear kernel. This occurs because model selection is performed independently within each fold via a grid search, and the RBF kernel outperformed the linear kernel in all other folds. Also, the plot shows that the model underestimates higher UPDRS scores and overestimates lower ones, as many points are compressed toward the center of the range. This effect is more pronounced with the RBF kernel, which, despite achieving a lower average RMSE, exhibits significant variability and scattered predictions across the range. In contrast, the linear kernel produces fewer predictions, with points clustering in a more consistent region, albeit with less flexibility in capturing extreme values. Overall, the plot confirms that while the RBF kernel can model nonlinear relationships, it is somewhat unstable, whereas the linear kernel provides more consistent but less expressive predictions.

Figure 3 shows the per-fold RMSE plot illustrating the model's performance variation across speakers under LOOCV. Each point represents the prediction error for a single test speaker across folds, while the dashed horizontal line indicates the mean RMSE of 15.71. The coefficient of determination R^2 was 0.25 when the best model was selected independently in

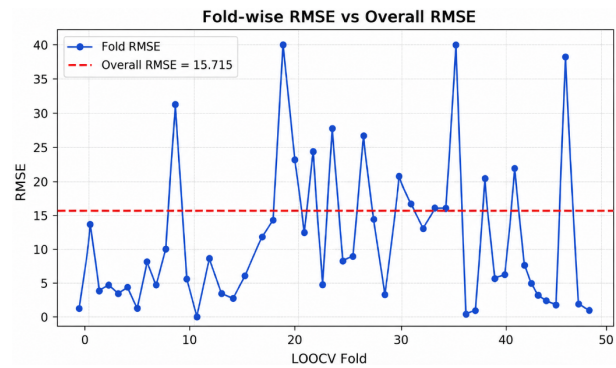


Figure 3. RMSE per fold corresponding to a speaker left out in cross-validation.

TABLE IV. MODEL PERFORMANCE WITH AND WITHOUT THE USE OF TEXTUAL PREFIX.

Setup	Dominant Kernel	RMSE
With prefix	RBF	15.71
Without prefix	RBF (29 times selected)	14.8628
Without prefix	linear (17 times selected)	18.4772
Without prefix	sigmoid (4 times selected)	24.6858

each fold, with a SE of 1.52, indicating moderate predictive accuracy and variability across train-test splits. The relatively low R^2 highlights the complexity of PD severity prediction and the influence of clinical and behavioral factors not fully captured by speech transcriptions. As can be seen, there is considerable variation in error across folds, with RMSE values ranging from very low to above 40. This suggests that the model performs quite accurately for some speakers but struggles with others, resulting in noticeably higher error rates. This variability indicates a strong dependence on speaker-specific characteristics, such as speech patterns, recording conditions, or disease severity, and highlights limited generalization across samples.

The textual prefix ("Parkinson speech severity prediction:") was introduced to guide the embedding model toward capturing clinically relevant semantic information. To verify its contribution to the results, two implementations were conducted: one using the prefix and one without it. The results presented in Table IV show that without the prefix, during the nested cross-validation procedure, all three kernel functions (linear, RBF, and sigmoid) were selected as the optimal SVR kernel in at least one fold and achieved an RMSE of 17.115, $R^2 = 0.11$ indicating no good generalization, whereas with the prefix, a non-linear RBF kernel was selected in most folds, leading to a lower total RMSE of 15.71. This indicates that the prefix is useful in this experiment because it alters the structure of the embedding space, enabling the model to capture more complex, non-linear relationships between speech-derived features and PD severity. These findings support the inclusion of the prefix as a simple yet effective strategy for improving representation quality and improving prediction.

The mean RMSE of the proposed method is compared with the mean RMSE disclosed in [16], where linguistic embeddings combined with SVR achieved mean RMSE values

between 13.5 for Bidirectional Encoder Representations from Transformers (BERT) [17] and 14.6 for XLNet [18], with the best performance of (12.9 ± 1.68) observed using the *text-embedding-3-large* model [19], employing embeddings of 768 dimensions and the sigmoid kernel. In this proposed approach, the model achieves a mean RMSE of (15.71 ± 1.52) with a coefficient of determination $R^2 = 0.25$, indicating improved predictive accuracy. However, the results also show good variability, particularly for the linear kernel, indicating less stable performance across folds, likely due to the small dataset and the choice of kernel. Additionally, hyperparameter tuning is performed separately within each fold, which can lead to different model configurations and further increase variability.

IV. CONCLUSION AND FUTURE WORK

The experimental results suggest that audio recordings transcribed to text can provide useful information about PD severity. The pipeline used for predicting UPDRS scores in PD included audio transcription, transformer-based embeddings, and SVR. On average, the predictions were approximately 16 points off the UPDRS score, with a minimum of 6 and a maximum of 93. The results indicate that linguistic representations extracted from transcribed speech contain relevant information for estimating clinical severity. Although this model has not achieved the highest precision among the previously described models, the coefficient of determination indicates a good prediction potential. It is not sufficient on its own for clinical estimation due to a 16-unit offset relative to an accurate estimate. The comparison between SVR kernels showed that the RBF kernel achieved a lower average RMSE. However, this improvement came with increased variability, indicating reduced robustness across folds. The linear kernel didn't produce more stable but less accurate predictions, highlighting a trade-off between model flexibility and generalization. Based on the results of the LOOCV folds, a tendency toward regression to the mean is relieved, with higher UPDRS scores being underestimated and lower scores overestimated. This behavior, along with the variability observed in per-fold RMSE, suggests that the model struggles to capture extreme cases and that performance is sensitive to individual speaker characteristics.

Some limitations include the small dataset (50 PD speakers), the use of text-only features, the aggregation of embeddings across recordings from the same speaker (which may obscure variability between recordings), and prediction errors of about 12 UPDRS points, which are not precise enough for detailed clinical use. Despite these challenges, the proposed approach demonstrates the potential of combining pre-trained models for transcription and embedding generation in clinical prediction tasks. The use of a contextual prefix further enhances the relevance of the extracted features with minimal complexity. Future work should focus on expanding the dataset through data augmentation, combining text with acoustic embeddings, and exploring more robust modeling approaches to improve performance and generalization.

ACKNOWLEDGMENT

The authors would like to thank Professor Juan Rafael Orozco Arroyave for sharing the PC-GITA dataset with them.

REFERENCES

- [1] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," *Research and Reviews in Parkinsonism*, vol. 9, pp. 9–16, 2019.
- [2] Y. Liu *et al.*, "Automatic assessment of Parkinson's Disease using speech representations of phonation and articulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 242–255, 2022.
- [3] B. Karana, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahtoa, "Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction," *Computer Speech & Language*, vol. 69, p. 101216, 2021.
- [4] B. Karan and S. S. Sahu, "An improved framework for Parkinson's Disease prediction using variational mode decomposition-Hilbert spectrum of speech signal," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 717–732, 2021.
- [5] J. C. Vázquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's Disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 2018.
- [6] Scikit-learn Developers, "Randomforestregressor," <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>, 2026, [retrieved: May, 2026].
- [7] —, "Mlpregressor," https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html, 2026, [retrieved: May, 2026].
- [8] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [9] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," 2022, [retrieved: May, 2026]. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [10] S. Lee, A. Shakir, D. Koenig, and J. Lipp, "Open source strikes bread - new fluffy embeddings model," 2024, [retrieved: May, 2026]. [Online]. Available: <https://www.mixedbread.ai/blog/mxbai-embed-large-v1>
- [11] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 342–347.
- [12] American Parkinson Disease Association, "Stages in Parkinson's Disease," 2013, [retrieved: May, 2026]. [Online]. Available: <https://www.apdaparkinson.org/article/stages-in-parkinsons>
- [13] Y.-Y. Yang *et al.*, "TorchAudio: Building blocks for audio and speech processing," 2021, [retrieved: May, 2026]. [Online]. Available: <https://arxiv.org/abs/2110.15018>
- [14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [15] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.
- [16] J. Crawford, "Linguistic changes in spontaneous speech for detecting parkinson's disease using large language models," 2024, [retrieved: May, 2026]. [Online]. Available: <https://arxiv.org/abs/2404.05160>
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, [retrieved: May, 2026]. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [18] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," 2020, [retrieved: May, 2026]. [Online]. Available: <https://arxiv.org/abs/1906.08237>
- [19] OpenAI, "OpenAI. new embedding models and api updates," 2020, [retrieved: May, 2026]. [Online]. Available: <https://developers.openai.com/api/docs/guides/embeddings>

Articulating Multidisciplinary Knowledge Bundles for Urban Planning

Robert Laurini
LIRIS, University of Lyon, France
email: Robert.Laurini@liris.cnrs.fr

Abstract—Conventional urban planning has long relied primarily on data, but contemporary practice increasingly integrates a wide array of knowledge bundles. These bundles originate from diverse disciplines and sources: expert analyses, regulatory frameworks, technical know-how, lessons learned from past projects, local historical in-sights, public participation outcomes, civil engineering, environmental planning, cartographic practices, and even the conventions of cartography and written communication. In addition, lessons learned through the simulation of city digital twin model must also be considered. Given this diversity, it becomes essential not only to identify the knowledge elements relevant to planning but also to articulate a coherent conceptual model capable of integrating them, taking account of both expert knowledge and shallow knowledge. To address this need, the paper proposes a visual model that represents the interactions among these heterogeneous knowledge bundles. Particular attention is given to the potential of generative artificial intelligence to synthesize citizen input and extract meaningful patterns and suggestions from public opinion. The paper concludes by outlining pathways for transitioning from data assisted to knowledge assisted urban planning, aligning this evolution with the broader ambition of fostering urban intelligence.

Keywords—Urban Planning; Knowledge Management; Visual Modeling; Smart Cities; Urban Intelligence.

I. INTRODUCTION

In urban planning, for decades, the dominant mantra has been “survey before planning” [1] with the connotation of data-assisted activity. So many cities have designed information systems for this purpose [2]. Now, with the advent of artificial intelligence and especially Knowledge Management (KM), a new era is coming under the umbrella of urban intelligence which is defined by the synergy between AI and human collective intelligence (Figure 1).

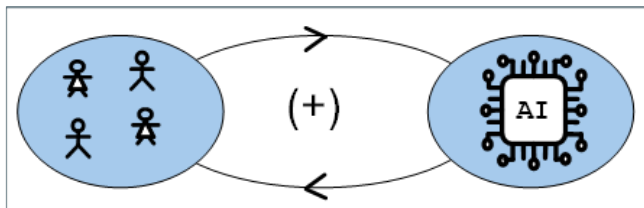


Figure 1. Synergy between artificial intelligence and human collective intelligence.

Since urban planning is a multidisciplinary activity, many different bundles of knowledge must be mobilized,

from architecture to economics, from civil engineering to legal issues, from pollution fights to sports and culture. And the goal of this paper is to identify and characterize articulations of those different bundles of knowledge within the scope of operability.

In computing, knowledge is usually defined as information that can guide action or solve problems [3]. Knowledge bundles can have different origins, i.e. regulations, written documents, reports, satellite images, aerial photos, expert’s know-how, data mining, etc. The major problem is not only compiling them but overall organizing them to help reasoning to find solutions.

However, much of the knowledge acquired by urban planners remains tacit or goes unrecognized as a vital city resource; it should instead be understood and managed as a shared collective wealth.

From a practical point of view, knowledge chunks can be rules (either physical or juridical), best practices, lessons learned, recognized patterns, verified statements, etc. In this category, one can add recommendations and adages. When it is possible, rules can be mathematically written so that they can be formalized in a computer language, i.e. made operational [4], otherwise their use is not so easy for automation.

The idea is not to develop an ontology describing cities or urban planning [5, 6], i.e. designing a formalized vocabulary dedicated to urban planning, but rather to organize existing knowledge chunks and bundles which can be useful to plan a smart city.

Bennet and Bennet [7] distinguish three levels of knowledge, surface, shallow and deep. In the field of urban planning, local authorities may be regarded as project owners, whereas for some aspects, specialized companies typically function as project managers or subcontractors. Put differently, these external actors possess deep knowledge, while planners within local authorities usually have more limited expertise, so shallow knowledge. So, in this study, only expert and shallow knowledge bundles in urban planning will be considered.

Let us make a comparison with car engines: mainstream drivers have some surface knowledge whereas a car vendor has shallow knowledge, and a car designer has an expert knowledge.

This paper is organized as follows: it will begin by a rapid state of review regarding the modeling of urban planning knowledge and then some issues regarding its background. Afterwards, the core of this paper will develop various bundles in the multidisciplinary domain linked to urban planning, together with the modeling for ending the planning process.

II. BUNDLE-ORIENTED MODEL OF KNOWLEDGE

In the state of the art about knowledge modeling, several major families of approaches exist, among other ones based on networks and on rules. Those based on networks allow an interesting model to describe a situation and to reason [8, 9] whereas rules permit to describe the results of a process [10]. Here, we need a model to help design a plan.

In this paper, a top-down approach will be used to present mainstream knowledge bundles for urban planning. Of course, about the context, additional bundles must be considered. In order to model the links between different bundles of knowledge, a visual vocabulary will be defined as in Figure 2 in which a hexagon represents a bundle or sub-bundle of knowledge and a scroll is used for any kind of regulations, existing or generated. A sub-bundle corresponds to a subset of knowledge. For modeling any kind of repository, a cylinder will be used, not only for representing any kind of multimedia databases, but also textual corpora. Two kinds of arrows will be used, the first one for the link between a sub-bundle of knowledge and its corresponding upper bundle, and another for the generation of new kind of knowledge. The process of generating knowledge can be human – for instance by discussing with an expert – or computer-based – for instance by data mining or through generative artificial intelligence.

The two “if-any” arrows (with dots) are introduced essentially because of national contexts which differ from country to country, either due to absence or from being not mandatory.

Colors will be used to denotate whether a bundle is the focus (main bundle), a sub-bundle or shallow knowledge

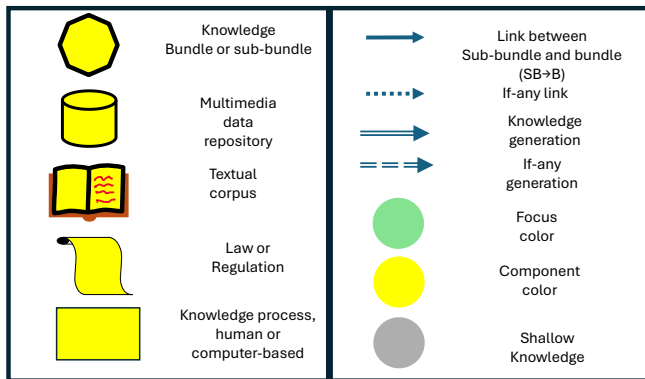


Figure 2. Visual vocabulary for modeling the links between urban knowledge bundles.

We will use this formalism to model the links between knowledge bundles in any organization such as companies, and here in planning.

III. BACKGROUND KNOWLEDGE BUNDLES IN URBAN PLANNING

Before turning to the planning process itself, it is helpful to outline the key elements that underpin it.

A. Juridical Framework

Commonly, urban planning is defined as the coordinated shaping of a city's physical spaces, infrastructure, and policies to support sustainable, functional, and livable communities. It can be seen as a feedforward component of urban governance. For that purpose, most countries have designed laws or more exactly regulations to organize urban planning. Figure 3 shows the national codes (if they exist), urban planning, environmental planning together with regulations concerning buildings and public participation.

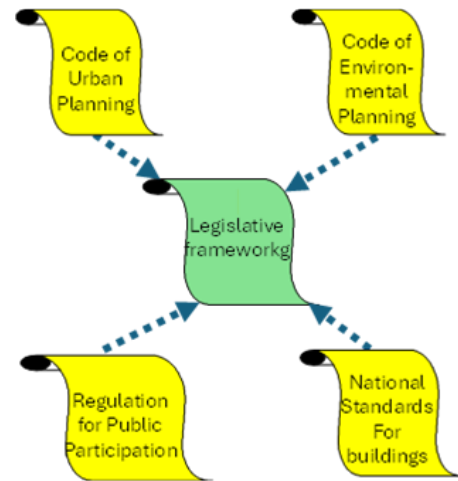


Figure 3. The legislative framework and its main components.

B. Definition of planning objectives

Each stakeholder has his own vision of the future of the city, based on his own knowledge about the city and some interesting experiences gained in other cities.

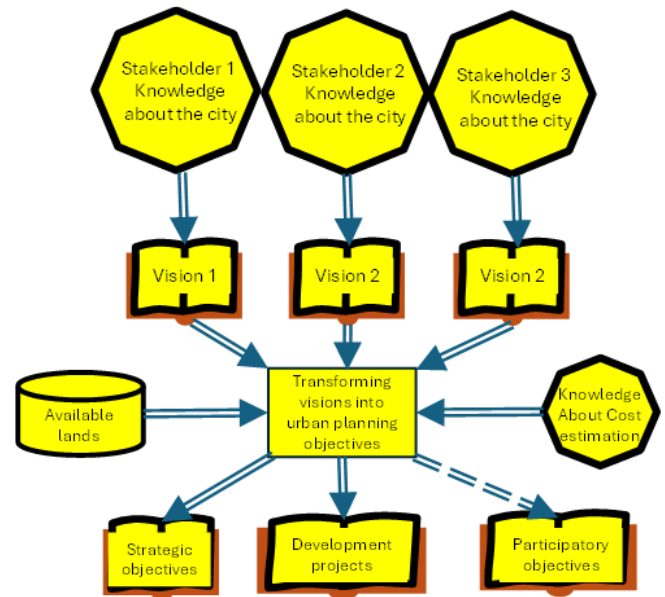


Figure 4. Transforming visions into planning instruments.

Among stakeholders, let us mention various administrative entities, political parties, some activist groups and other protagonists such as the local chamber of commerce. Maybe, some ideological issues can be included in the vision of some of them.

Anyhow, there is a mechanism to select those visions or to make a sort of compromise between them according to land opportunities and some estimation of cost. This mechanism can be some vote made by the local council. Three documents are written, the list of strategic objectives together with their priority, and some structuring projects, those corresponding to planning instruments (Figure 4). Possibly they are complemented by some recommendations regarding citizens' participation.

The case of development projects will not be detailed in this paper. Let us remember that local authorities can act either as a regulator checking whether projects follow all planning rules (e.g. for a gated community), or as a prime contractor (e.g. for a tram newline).

C. Knowledge involved in methodology

Once the strategic objectives are set, the methodology (Figure 5) will begin by studying the previous plans to discover parts currently on-going or not yet implemented, and by stating whether they are still in consideration. Then, it will be a sort of amalgamation of experts' know-how, best practices, conventional methodologies and lessons learned from the past (in the city itself and also elsewhere). If the city has implemented a digital twin (CDT), several simulations can give indications and tips about the potential outcomes.

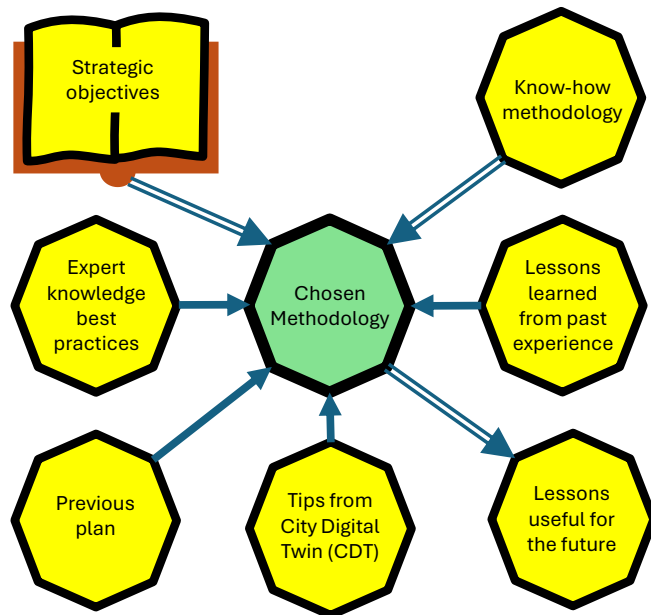


Figure 5. Planning methodology.

D. About internal and external geolocated knowledge

Recall Tobler's Law [11], often regarded as the First Law of Geography: "Everything is related to everything else, but

near things are more related than distant things". It implies not to limit to the jurisdiction, but also to consider external influences, in other words, external knowledge. For instance:

- The headquarters of a plant – located in my city – are in New York City and its CEO has decided to close it; so, increasing local unemployment.
- A city located in another country has implemented an original biking rental activity. This activity could be deployed in my own city.
- The region has decided to design a new railway system crossing my city, but the main station is not in my city.
- A natural or technological disaster in a neighboring city has consequences for my city.
- Etc.

So, we define external knowledge as information describing external influences that are relevant to a given territory. Several categories of such knowledge can be distinguished (Figure 6):

- Short distance or neighboring knowledge, i.e. knowledge located in the vicinity of the jurisdiction; a solution can be to determine an out-buffer zone around the jurisdiction and consider knowledge inside this buffer zone.
- Long distance knowledge includes the source and contents of external influences coming perhaps from decision made elsewhere. Here, we must consider that local authorities only have shallow knowledge of those influences.
- Technology and sociological watching can be included in external knowledge; however, in some cities, technology and sociological watching must be considered as shallow knowledge.

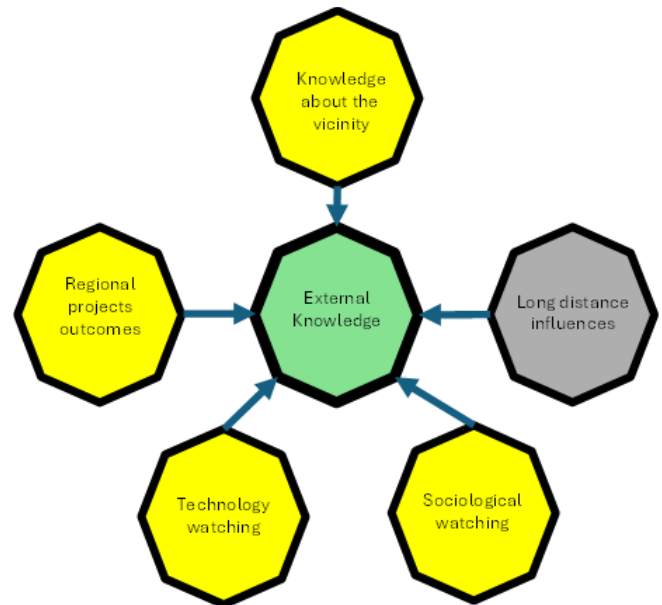


Figure 6. City external knowledge and its major sub-bundles.

IV. DETAILED MULTIDISCIPLINARY MODELS

In this section, several models of knowledge bundles will be presented. It is important to state that they are only for planning purposes, not for design. Indeed, for the design, for instance, of a shopping mall, several consulting companies will be involved with their own knowledge. For planning purposes, only shallow knowledge will be mobilized regarding many activities in the city.

A. General Knowledge about the City

This bundle regroups 6 sub-bundles, namely concerning local peculiarities, local history, city as a system, listed monuments, population needs and risks and disasters (Figure 7). Of course, it is necessary to mention knowledge coming from databases, geographic information systems including 2D and sometimes 3D data, aerial photos and perhaps satellite images.

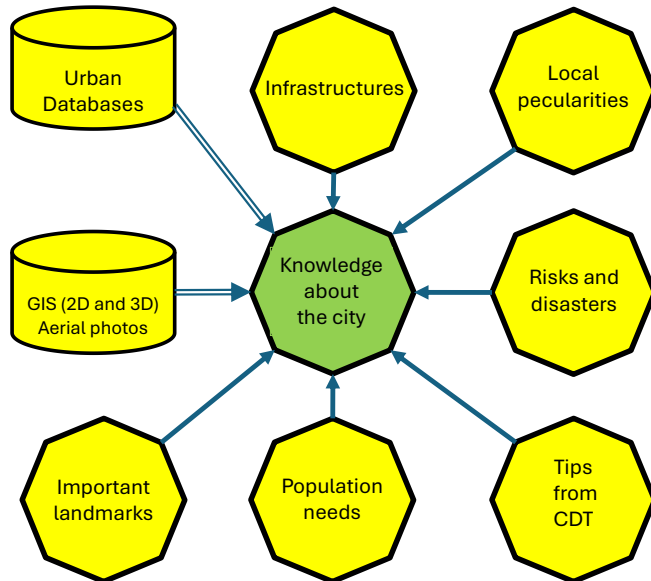


Figure 7. Knowledge about the city and its sub-bundles.

B. Local Peculiarities

Not all cities are similar; each has its local peculiarities due to history and its geographic location (mountain, sea, etc.). For instance, the problems generated by the existence of a harbor or from important touristic landmarks can influence its development, such as skiing resort, or an archaeological site. (Figure 8). Moreover, having an airport or a regional hospital are important assets for economic development.

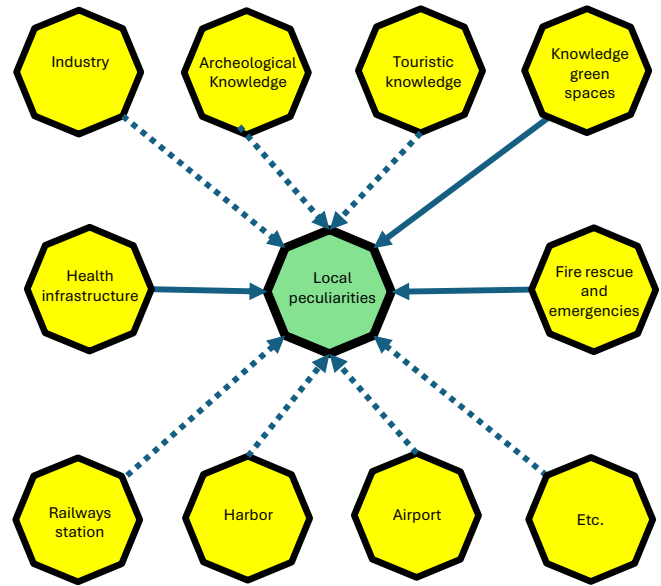


Figure 8. Local peculiarities.

C. Various multidisciplinary models

The wide range of urban activities allows for the design of various bundle models. Figure 9 illustrates a selection of these models; however, many additional variants can also be considered.

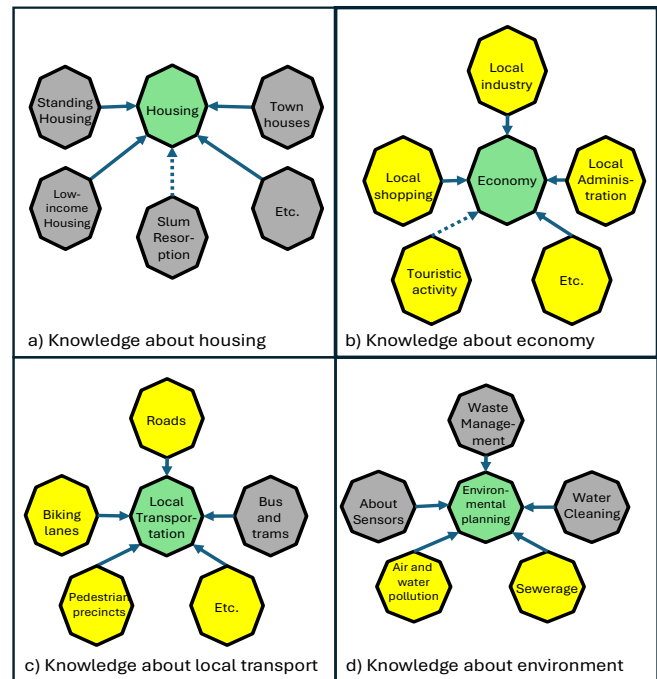


Figure 9. Models of various sub-bundles.

D. Tips Learned from City Digital Twins

Several definitions of the concept of City Digital Twin (CDT) exist [12]. In essence, a digital twin city represents a dynamic, data-driven virtual replica of a physical urban environment. Unlike static 3D models or conventional maps, this technology creates a continuously updated virtual counterpart that reflects real-time behavior, status, and characteristics of the actual city. The system integrates sensor data, satellite imagery, traffic flows, environmental readings, and infrastructure performance metrics to maintain accuracy and relevance.

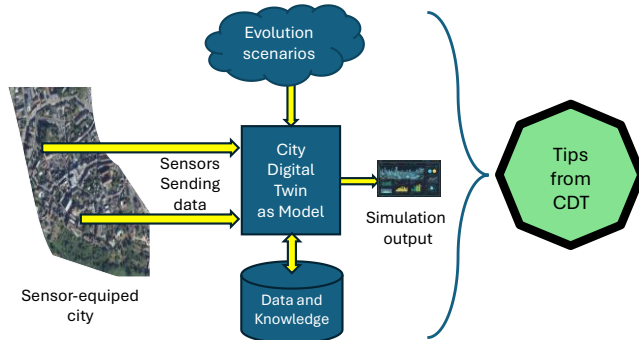


Figure 10. Knowledge (tips) issued from the use of city digital twin system.

The more important thing in our domain is the way that scientists, policymakers and planners interact with real cities with respect to their understanding, prediction and design [13]. The city digital twin system integrates a range of spatial models, from analytical simulations of local neighborhoods to large-scale systems of cities and city systems and briefly describe computational challenges that geospatial applications in cities pose.

Practically, users develop some evolution scenarios which are submitted to the CDT model and generate simulated outputs (Figure 10). Then, those outputs are evaluated through some criteria (efficiency, cost, etc.). At the end of this process, some tips learned are listed as knowledge bundles.

E. Public Participation Knowledge

Two kinds of knowledge must be considered: first, the way to deal with collected citizens' contribution and second, regarding its organizational structure.

In some countries, there are regulations regarding the organization, the documents to be presented to the citizens, the duration, the minutes, etc. However, the more effective the organization is, the more likely citizens are to feel that their opinions are being taken into account. In other words, this is a problem of communication.

Knowledge bundles should be maintained on the most effective ways to present the plan, including maps and explanatory materials, the city's evolution, future development scenarios, key projects, and the organization of public hearings. Some coaches can be trained to explain to

people the ins and outs of the plan and to encourage them to give their opinion.

It is important to mention that the presented bundles of knowledge must be considered as shallow knowledge in the countries in which citizen participation is not legally organized (Figure 11).

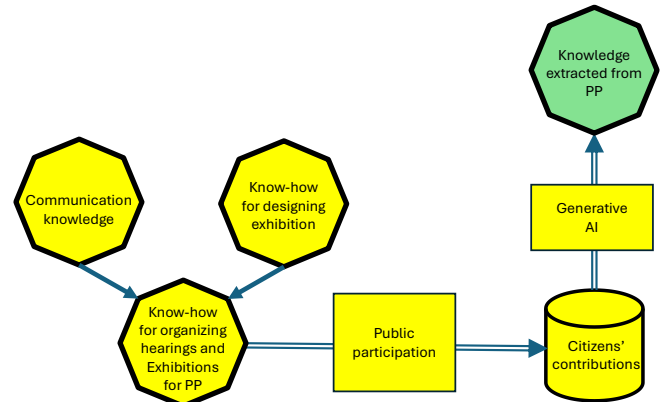


Figure 11. Knowledge bundles necessary to get citizens' contributions through Public Participation (PP).

In urban governance, it is common to organize public surveys, consultations or hearings in which citizens can give their opinions or suggestions on projects proposed in the plans. In other words, those suggestions are positioned through an area and can constitute knowledge chunks coming directly from people. Although there are nuances in the vocabulary, here the word "suggestions" will be used as a general term to describe contributions, opinions, complaints, objections, etc.

It looks like it is possible to use generative AI to extract salient patterns from large volumes of citizen suggestions, to generate summaries and highlight key concerns or good ideas and eventually simulate urban design scenarios based on public preferences.

F. Final Remarks

The model can also be extended to include other knowledge bundles. Examples include infrastructure systems, such as road networks and utilities (electricity, telecommunications, street lighting, traffic management, and district heating), as well as public services related to healthcare, culture, sports, local administration, policing, and recreation. In accordance with the legislative context, such aspects may not be managed by the city, but by other administrations or private companies. Within this model, however, they will often be represented as shallow knowledge for planners.

These shallow bundles are typically entrusted to partners with greater expertise in their respective domains. Two solutions can be mentioned, either to build a file of potential partners and to contact one of them, or to write an open call for tenders.

V. KNOWLEDGE MECHANISMS REQUIRED TO CONCLUDE THE PLANNING PROCESS

Once the planners have completed their work, they have to prepare several blueprints out of which two are mainstream: a written report and a map (Figure 12).

The written report will include the objectives of planning and the details for each zone; a substantial part of this report will be the basis for local regulations. For its part, the map must be consistent with the written report and must show visualization of the plan.

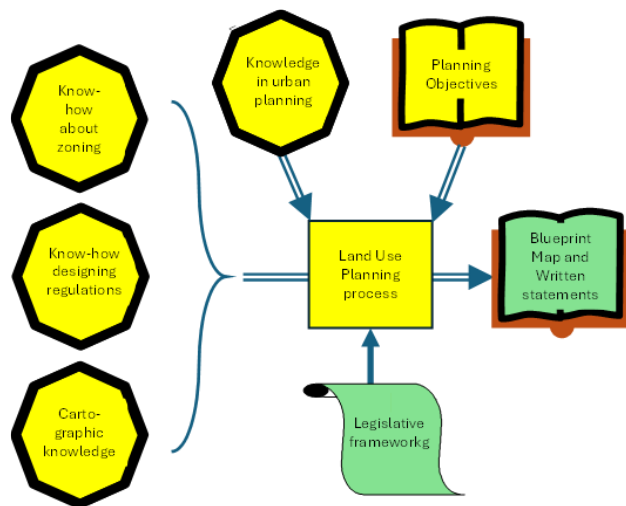


Figure 12. Final blueprint design.

According to the legislation, there is a mechanism for approving the blueprints; for instance, the local council may vote to approve it, thereby giving rise to an enforceable master plan. From the planner point of view, by designing the plan, some lessons have been learned, and a sort of learning feedback loop will be generated, as illustrated in Figure 13. This feedback can affect any kind of knowledge bundles listed in this paper.

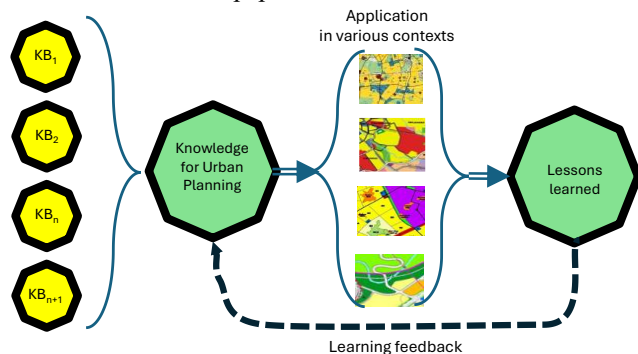


Figure 13. Learning feedback.

VI. CONCLUSION

In essence, urban planning is a multidisciplinary task [14]. So, the scope of this paper was to examine the articulations of the bundles of knowledge that planners have to master by means of a visual vocabulary. Two levels of

knowledge bundles have been identified: expert knowledge and shallow knowledge. Naturally, many other models can also be developed.

The next step is to define appropriate models not only for storing all types of bundles, but also for enabling reasoning over them. By completing and validating this framework, a new form of knowledge-assisted urban planning can be envisioned [10].

REFERENCES

- [1] G. L. Pepler, "Geddes' contribution to town planning," *Town Planning Review*, vol. 26, no. 1, pp. 19–24, 1955, doi: 10.3828/tpv.26.1.u02775g40240m504.
- [2] R. Laurini, *Information Systems for Urban Planning: A Hypermedia Cooperative Approach*. London, U.K.: Taylor & Francis, Feb. 2001, p. 308.
- [3] R. L. Ackoff, "From data to wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3–9, 1989.
- [4] R. Laurini, "A mathematical language for the modeling of geospatial static rules," *Journal of Visual Languages and Computing*, vol. 51, pp. 1–13, 2019.
- [5] J. Teller, K. Keita, C. Roussey, and R. Laurini, "Urban ontologies for an improved communication in urban civil engineering projects: Presentation of the COST Urban Civil Engineering Action C21 TOWNTOLOGY," *Cybergeog: European Journal of Geography*, no. 386, 2007, doi: 10.4000/cybergeog.8322.
- [6] A. De Nicola and M. L. Villani, "Smart city ontologies and their applications: A systematic literature review," *Sustainability*, vol. 13, no. 10, Art. no. 5578, 2021, doi: 10.3390/su13105578.
- [7] A. Bennet and D. Bennet, "The depth of knowledge: Surface, shallow or deep?" *VINE Journal of Information and Knowledge Management Systems*, vol. 38, no. 4, pp. 405–420, 2008, doi: 10.1108/03055720810917679.
- [8] Y. Liu, J. Ding, Y. Fu, and Y. Li, "UrbanKG: An urban knowledge graph system," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 4, Art. no. 60, pp. 1–25, 2023, doi: 10.1145/3588577.
- [9] S. Wang, P. Qiu, Y. Zhu, J. Yang, P. Peng, Y. Bai, and Y. Qi, "Review, framework, and future perspectives of Geographic Knowledge Graph (GeoKG) quality assessment," *Geo-Spatial Information Science*, pp. 1–21, 2024, doi: 10.1080/10095020.2024.2403785.
- [10] R. Laurini, "Urban planning knowledge modeling for urban intelligence," in *Knowledge Systems and Human Intelligence: How We Organize, Share, and Apply Information*, O. Polishchuk, Ed. Hauppauge, NY, USA: Nova Science Publishers, 2026, in press.
- [11] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic Geography*, vol. 46, no. 2, pp. 234–240, 1970.
- [12] M. Hooli, "Digital twins and urban planning: Designing smarter, more inclusive cities," *European Journal of Computer Science and Information Technology*, vol. 13, no. 44, pp. 104–112, 2025, doi: 10.37745/ejcsit.2013/vol13n44104112.
- [13] M. Batty, "Digital twins in city planning," *Nature Computational Science*, vol. 4, pp. 192–199, 2024, doi: 10.1038/s43588-024-00606-7.
- [14] R. Laurini, "Groupware for urban planning: An introduction," *Computers, Environment and Urban Systems*, vol. 22, no. 4, pp. 317–333, 1998, doi: 10.1016/S0198-9715(98)00029-5.

Proactive Link Failure Early Warning for Ultra-Long-Range Maritime 5G via Edge Intelligence and Random Forest

Kun Yang, Yu Xu, and Li Qin
School of Information Engineering
Zhejiang Ocean University
Zhoushan, China

e-mail: yangkun@zjou.edu.cn, xuyu2025@zjou.edu.cn, ql_qinli@zjou.edu.cn

Abstract—In the Maritime Internet of Things (MIoT), the link stability of 5G communications is a critical bottleneck for the reliable execution of mission-critical tasks by autonomous ships and maritime mobile robots. As an integral component of intelligent sensing infrastructures, proactive link failure early warning serves as a cross-layer perception mechanism, providing essential connectivity for robust navigation and situational awareness in extreme environments. However, traditional physical-layer hardware enhancement strategies exhibit diminishing returns when addressing nonlinear, stochastic signal fading in Beyond-Line-of-Sight (B-LoS) scenarios. To address this challenge, this paper proposes an embedded communication-aware framework for robot perception. By treating Channel State Information (CSI) as virtual sensing data, temporal features are dynamically extracted via a sliding window mechanism, and an intelligent sensor data interpretation model is constructed using a Random Forest (RF) algorithm to achieve proactive awareness of interruption risks. Validation performed on field-test data collected from a 341 km voyage of the “Sansha No. 2” supply ship demonstrates that the model achieves a prediction accuracy of 94.95%. Results indicate that the algorithm effectively balances millisecond-level inference latency with predictive precision, thereby fulfilling the strict real-time processing requirements of embedded platforms under resource-constrained conditions. This research provides a methodological reference for integrating adaptive sensing strategies into maritime autonomous systems to enhance navigation resilience.

Keywords—Maritime Internet of Things; robot perception; Random Forest; link failure prediction; edge intelligence.

I. INTRODUCTION

The rapid evolution of the Maritime Internet of Things (MIoT) has made autonomous ships and maritime mobile robots essential to the global blue economy. In typical maritime scenarios, such as multi-vessel coordinated scheduling, marine scientific observation, and autonomous navigation, constructing high-reliability and high-bandwidth wireless communication networks is a primary objective in both academia and industry [1]. Although terrestrial cellular networks have achieved robust coverage in coastal regions, ensuring end-to-end Quality of Service (QoS) under long-distance and highly dynamic sea conditions remains a major technical challenge in maritime communications [2].

Figure 1 illustrates the maritime 5G topology and communication scenarios. The figure depicts terrestrial base

stations, experimental vessels, and signal propagation paths influenced by the Earth’s curvature and sea-surface reflection.

These propagation challenges become particularly pronounced as the offshore distance exceeds 60 km. Due to Line-of-Sight (LoS) blockage caused by the Earth’s curvature and multipath reflection effects induced by complex sea-surface fluctuations, signal amplitudes exhibit deep-fading characteristics [3]. As shown in Figure 1, the sea surface acts as an irregular reflector where reflected and direct waves superimpose coherently at the receiver, leading to severe nonlinear signal fluctuations. Related 5G channel measurement studies indicate that these complex propagation characteristics render traditional terrestrial propagation models inapplicable in maritime scenarios, resulting in a significant stability deficit for communication links in the open sea [4].

Consequently, both academia and industry often adopt enhanced hardware deployment strategies. However, empirical analysis in this study reveals that while pure hardware gain enhancements can improve the Reference Signal Received Power (RSRP), they cannot fundamentally suppress the stochastic drops in the Signal-to-Interference-plus-Noise Ratio (SINR). During an ultra-long-distance test of 341 km, enhanced terminals still experienced frequent instantaneous link crashes, demonstrating a clear stability bottleneck in physical-layer-only enhancement schemes. Existing research suggests that anomaly behavior recognition and tracking for dynamic spectrum environments can provide new technical pathways for securing and stabilizing communication links [5]. Deploying lightweight intelligent algorithms at edge nodes to proactively predict link risks offers a viable alternative to redundant hardware design, thereby improving the robustness of maritime communications.

To meet the link maintenance requirements of ultra-long-range maritime 5G communication, the primary contributions of this paper are as follows:

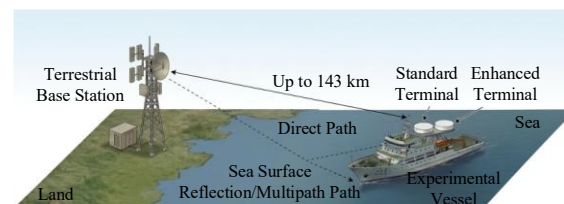


Figure 1. Maritime 5G topology and communication scenario.

- Construction of a scarce deep-sea field-test dataset: Based on the “Sansha No. 2” measurement platform, full-dimensional channel performance data were acquired for dual terminals at distances of up to 143 km from terrestrial base stations. These data provide empirical evidence for the Beyond-Line-of-Sight (B-LoS) coverage research of terrestrial 5G base stations in the open sea.
- Proposal and validation of a lightweight edge early-warning scheme: Addressing the computational resource constraints of MIoT routers, the effectiveness of a Random Forest (RF) algorithm for link failure early warning was verified. Field-test results indicate that this scheme achieves a real-time prediction accuracy of 94.95% for failure risks, providing empirical support for the construction of adaptive and intelligent MIoT.

The remainder of this paper is structured as follows: Section II reviews related work. Section III describes the methodology, including data preprocessing and classifier design. Section IV presents and analyzes the experimental results. Section VI provides a detailed discussion, and Section VI concludes the paper and discusses future work.

II. RELATED WORK

Accurate characterization of maritime channel features is a prerequisite for constructing a high-reliability MIoT. Although existing research has performed extensive measurements on 5G communication propagation characteristics across various scenarios [4], empirical studies in ultra-long-range maritime environments remain scarce. Wang et al. systematically reviewed maritime wireless channel models, identifying multipath fading and sea-surface reflection as the dominant factors affecting link stability [3]. However, most current maritime experiments are concentrated in near-shore areas within 20–40 km of the coast; consequently, there is a lack of authentic and continuous experimental data for long distance exceeding 100 km that cross the LoS boundary. Because signal propagation at long distance is further complicated by the combined effects of the Earth’s curvature and tropospheric refraction, existing near-shore empirical models exhibit significant limitations in describing the stochastic link failure behavior of deep-sea links.

Since traditional deterministic models struggle to characterize the nonlinear fluctuations of wireless channels, Machine Learning (ML) and Artificial Intelligence (AI) technologies are increasingly utilized for wireless network optimization [6]. Kibria et al. noted that the utilization of big data analytics and ML algorithms can significantly enhance resource scheduling and link management efficiency within complex heterogeneous networks [6]. Regarding the processing of highly dynamic channel variations, Liang et al. proposed an ML framework for vehicular network environments, demonstrating that reinforcement learning and ensemble learning possess superior generalization capabilities in addressing signal fading prediction caused by high-speed terminal mobility [7]. Applying ML algorithms to maritime networks, which share dynamic characteristics similar to

vehicular networks, is a promising approach to mitigate stochastic signal drops.

As the scale of Internet of Things (IoT) networks grows, the latency and bandwidth limitations of traditional cloud-centric processing become more apparent. This challenge is particularly evident in maritime operational environments: real-time early warning of maritime communication link quality imposes millisecond-level requirements on the inference latency of prediction algorithms; however, the satellite backhaul links between shipborne edge nodes and remote cloud servers often fail to meet these constraints. Consequently, migrating intelligent analysis functions from the cloud to edge nodes (e.g., shipborne routers) provides a practical solution.

For instance, Sudharsan et al. proposed adaptive communication strategies that rely on lightweight algorithms to locally optimize real-time communication quality for IoT edge devices [8]. Nevertheless, MIoT terminals are generally constrained by limited computational capacity and storage volume. Thus, further compressing model computational complexity and memory footprint while maintaining high prediction accuracy remains a key challenge for the practical deployment of edge intelligence in maritime scenarios.

III. METHODOLOGY

This section details the experimental design and the machine learning architecture, spanning from physical layer perception to edge-side decision-making.

A. Experimental Setup and Hardware Parameters

Field-test data were acquired using the “Sansha No. 2” supply ship during a 341 km voyage. The route reached a maximum offshore distance of 143 km, approaching the B-LoS coverage limits of terrestrial base stations. The complex meteorological and dynamic sea states provide an authentic environment for evaluating 5G stability. As shown in Figure 2, standard and enhanced terminals with distinct radio frequency gains were deployed on the main deck.

B. Data Preprocessing and Management of Missing Values

Regarding the non-stationary time-series data, this study did not employ conventional interpolation for filling gaps. All illegal observations containing non-numeric characters were removed to ensure the arithmetic validity of the input vectors. During large-scale disconnection intervals of the standard terminal, the state was labeled as “Out-of-Service” to preserve the authentic physical characteristics of channel collapse. To eliminate the influence of different physical dimensions on

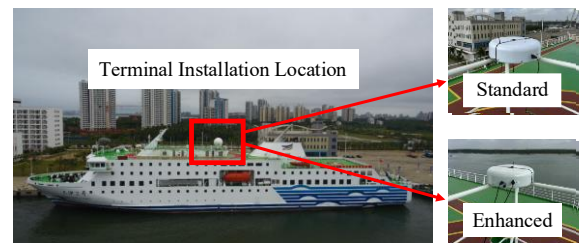


Figure 2. Terminal installation location.

model convergence speed, Min-Max normalization was applied to RSRP, Reference Signal Received Quality (RSRQ), SINR.

After preprocessing and sliding window transformation, the refined dataset comprises 2271 valid temporal samples, featuring a class distribution of 1590 (70.01%) normal states and 681 (29.99%) warning states (defined as SINR < 0 dB). To ensure rigorous evaluation and prevent data leakage, the samples were chronologically partitioned into a training set (80%, 1816 samples) and an independent testing set (20%, 455 samples).

C. Feature Engineering: Time-Series Sliding Window Mechanism

Wireless signal fading processes exhibit significant temporal dependency; deterioration in signal quality at the current moment is typically a precursor to imminent link interruption. To precisely capture these dynamic evolutionary patterns, this study introduces a sliding window mechanism to reconstruct the time-series signals and mine the temporal correlation features of channel states [9], as shown in Figure 3.

Let the raw feature vector at time t be defined as:

$$S_t = [RSRP_t, RSRQ_t, SINR_t]^T, \quad (1)$$

where the window length is set to n , the channel states of the past n consecutive moments are aggregated into the current model input vector. In this experiment, $n = 5$ was selected to integrate the observation sequence of the previous 5 seconds. The mathematical expression is:

$$X_t = \{S_{t-n}, S_{t-n+1}, \dots, S_t\}. \quad (2)$$

This methodology transforms the traditional static point-to-point classification problem into a pattern recognition problem based on temporal context. By incorporating temporal dimension observations, the model can capture the slope features of severe SINR fluctuations and the

instantaneous jitter trends of RSRQ, thereby identifying microscopic physical layer anomalies before a complete link interruption occurs.

D. Random Forest Classifier Design

This study employs the RF ensemble learning framework proposed by Breiman as the prediction algorithm [10], as shown in Figure 3. Compared to deep learning models with high computational overhead, such as Long Short-Term Memory (LSTM) networks, RF strikes a superior balance between algorithmic performance and feasibility for edge engineering deployment [11–13].

The RF algorithm constructs multiple decision trees via bootstrap sampling and selects only a random subset of features during the node-splitting process of each tree. This dual randomness enhances the model's generalization and mitigates overfitting when processing noisy and highly stochastic maritime channel data. Model training uses Gini impurity as the optimization criterion for node splitting; this index can effectively quantify the marginal contribution of different channel features (such as RSRQ fluctuation amplitude and SINR trend) to link failure time. Its mathematical expression is as follows:

$$Gini(D) = 1 - \sum_{k=1}^2 p_k^2, \quad (3)$$

where p_k denotes the probability that a sample belongs to the k^{th} category (Normal or Warning). In highly dynamic fluctuation scenarios of maritime channels, the advantage of the Gini criterion lies in its superior computational efficiency and robust tolerance to outliers. Specifically, the algorithm traverses all features within the sliding window (e.g., the variation slope of SINR) to identify the optimal splitting threshold that minimizes node impurity. By minimizing node impurity, this screening mechanism allows the model to decouple genuinely predictive SINR perturbation features from the power compensation effects of RSRP. While

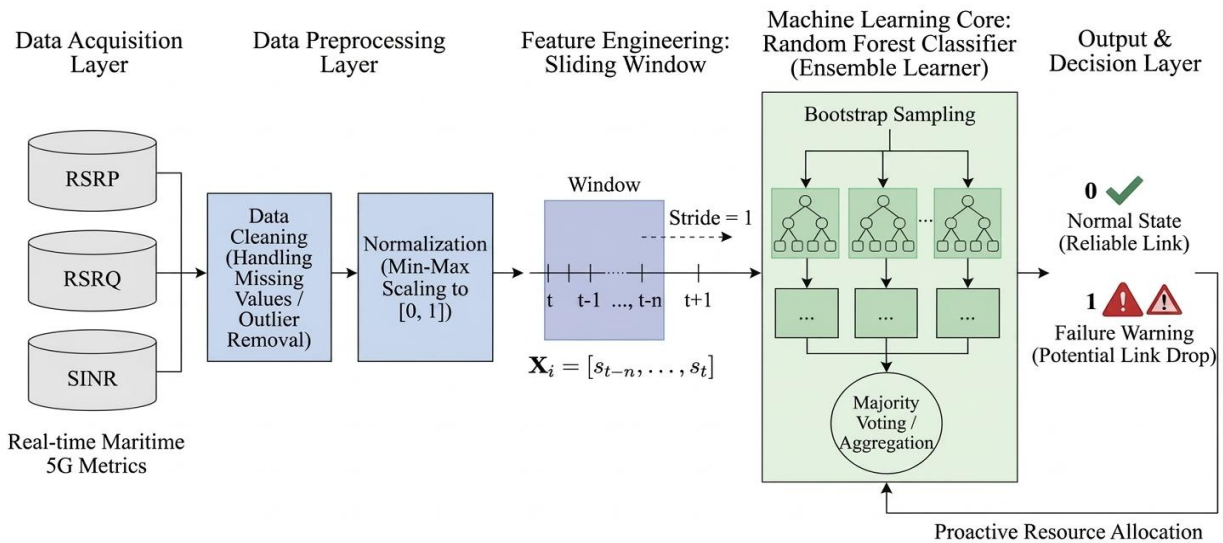


Figure 3. Proposed Edge-AI link prediction framework based on sliding window feature engineering and random forest classifier.

information entropy (Information Gain) is also a standard splitting criterion [14], the Gini index avoids complex logarithmic computations, making it more amenable for deployment on resource-constrained edge routers [14–15]. This low-complexity decision mechanism ensures the real-time inference performance of the algorithm on shipborne edge gateways.

Table I benchmarks the proposed RF model against Extreme Gradient Boosting (XGBoost), LSTM, and Gated Recurrent Unit (GRU) using the same sliding window configuration. While recurrent neural networks like LSTM and GRU exhibit high recall, their significantly lower precision would trigger excessive false alarms in autonomous maritime navigation. In contrast, the RF model achieves superior overall performance with an accuracy of 0.95. The selection of RF is further justified by its architectural suitability for edge intelligence. Unlike the complex tensor operations and high computational overhead associated with deep learning models, the RF algorithm relies on parallel logical decision trees that ensure millisecond-level inference latency on resource-constrained shipborne gateways without GPU support. Specifically, the model is configured with 100 estimators and a maximum tree depth of 10 to prevent overfitting while optimizing efficiency. By utilizing the Gini index, the algorithm avoids the complex logarithmic computations required by information gain, facilitating seamless deployment on System on Chip (SoC) based edge devices.

IV. EXPERIMENTAL RESULTS

The field-test dataset is quantitatively analyzed across two dimensions: physical-layer transmission characteristics and algorithmic prediction efficacy. This evaluation reveals the degradation patterns of 5G links in long-distance maritime environments and validates the proposed early-warning model.

A. Physical Layer Performance Analysis and Hardware Bottleneck Observations

To further quantify the actual contribution of hardware enhancement to link performance, Cumulative Distribution Functions (CDFs) are used to statistically model the field-test dataset. The spatial distribution of link failures and the corresponding ship trajectories are visualized in Figure 4(a) for the standard terminal and Figure 4(b) for the enhanced terminal. As illustrated in the SINR and RSRP CDF curves in Figures 4(c)–(d), high-gain hardware effectively extends the coverage radius. Specifically, the enhanced terminal reduces the proportion of sampling points in the “weak coverage zone” (below -105 dBm) to less than 5%. However, as shown in the SINR CDF curve in Figure 4(c), a non-negligible proportion of samples for the enhanced terminal still fall within the failure interval where $\text{SINR} < 0$ dB due to stochastic maritime multipath fading. These statistical results reveal an inherent “power-quality paradox”: relying solely on hardware-level power gain cannot fundamentally resolve nonlinear signal fluctuations, necessitating the introduction of intelligent prediction algorithms for proactive link maintenance.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT PREDICTION MODELS

Model	Accuracy	Precision	Recall	F1-score
RF	0.95	0.93	0.90	0.94
XGBoost	0.92	0.87	0.88	0.91
LSTM	0.87	0.71	0.96	0.86
GRU	0.87	0.73	0.91	0.85

B. Quantitative Summary of Network Performance Indicators

We statistically analyzed the communication indicators collected by both terminals across the entire 341 km voyage. As presented in Table II, the average network throughput of the enhanced terminal is nearly four times higher than that of the standard terminal. However, its average communication latency remains constrained by the physical laws governing long-distance electromagnetic wave propagation and has not undergone significant improvement.

C. Performance Evaluation of the Random Forest Early Warning Model

Following the identification of the performance limitations of hardware enhancement schemes, this section evaluates the actual early-warning performance of the RF prediction model based on temporal sliding window feature engineering. The core of this model is to learn microscopic fluctuation precursors of signal quality to achieve proactive warning of link failure events (defined as $\text{SINR} < 0$ dB).

Figure 5(a) presents the confusion matrix of the RF classifier on the test set, achieving an overall prediction accuracy of 94.95%. Quantitative evaluation results show that the model recall and F1-score are 90% and 94%, respectively, indicating that the algorithm can effectively identify the vast majority of potential channel drop events and significantly reduce the risk of data transmission interruption caused by false negatives. Under edge computing constraints, the RF model exhibits robust performance, effectively filtering non-stationary channel noise induced by maritime multipath

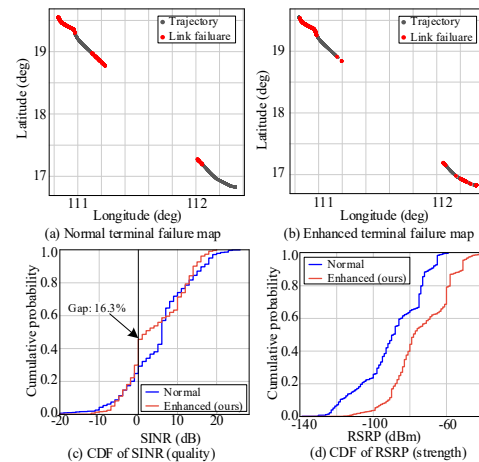


Figure 4. Comprehensive performance evaluation of maritime communication links.

TABLE II. SUMMARY OF NETWORK PERFORMANCE INDICATORS

Metric	Standard Terminal	Enhanced Terminal
Average RSRP	-103.5 dBm	-82.9 dBm
Average RSRQ	-13.5 dB	-11.9 dB
Average SINR	5.4	4.4
Average Network Latency	98.56 ms	86.93 ms
Average Downlink Rate	3.73 Mbps	13.11 Mbps
Average Uplink Rate	0.71 Mbps	2.76 Mbps
Link Failure Rate	48%	12%

effects and precisely capturing channel degradation trends with clear physical significance.

D. Feature Importance Analysis and Physical Interpretation

To interpret the model's decision-making mechanism, the Gini Importance index was employed to quantify the contribution weights of the input features.

Figure 5(b) illustrates the contribution weight distribution of each feature indicator to the prediction results. Analysis results indicate that the temporal variation rate of SINR and the fluctuation magnitude of RSRQ are the dominant factors determining link stability. From a physical mechanism perspective, while RSRP characterizes absolute signal strength, its sensitivity as an indicator for instantaneous link interruption is lower than that of SINR in maritime propagation environments where background noise and multipath interference effects are significant. Feature importance ranking further confirms that the first-order difference of SINR within the sliding window (i.e., the variation trend) accounts for the highest proportion, which corroborates the physical law that link failure follows a progressive degradation pattern. By extracting the fluctuation characteristics of RSRQ and SINR, the RF model can effectively identify complex nonlinear interference patterns, which is the primary reason its prediction performance surpasses traditional fixed-threshold decision methods.

V. DISCUSSION

Field test results across the 341 km voyage confirm that hardware-centric strategies encounter physical limitations in highly dynamic maritime environments. While enhanced terminals extend coverage, relying solely on antenna gain introduces additional multipath interference and creates “perception blind spots” for robot navigation. In contrast, the proposed RF-based early-warning scheme offers a more efficient alternative. By leveraging edge intelligence, this soft-hardware co-design improves link stability with minimal computational overhead, providing a predictive diagnostic layer for resilient MIoT deployment.

Maritime communication imposes stringent real-time requirements where network latencies exceeding 1000 ms can lead to complete interruption. Edge computing meets these ultra-low latency demands by offloading tasks [16]. Deploying a lightweight RF model within maritime gateways embeds the prediction logic directly into edge processing units.

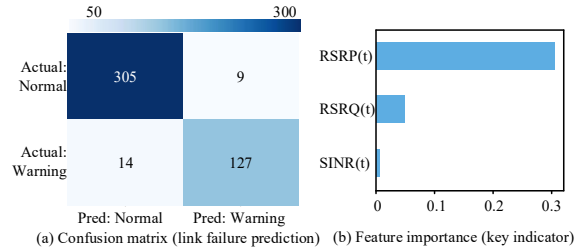


Figure 5. Classification Performance Evaluation of the Random Forest Classifier.

Since RF consists of logical condition judgments without complex tensor operations, it achieves true millisecond-level warning. This “algorithmic sinking” alleviates backhaul bandwidth pressure and provides a critical safety buffer, allowing robots to execute fail-safe maneuvers well before total link loss occurs.

The consistency of high performance in RF underscores its generalization potential. By capturing relative evolutionary trends through the sliding window rather than absolute thresholds, the model perceives degradation patterns that are universal across different sea states. While current results are based on a 341 km voyage, this methodology provides a baseline for future cross-scenario validation in extreme climates. Future research could integrate online or federated learning to perform adaptive parameter fine-tuning for specific deep-sea environments. Furthermore, this edge framework can be migrated to 6G Non-Terrestrial Networks (NTN) to compensate for atmospheric scintillation and deep fading, contributing to resilient global ocean sensing networks.

VI. CONCLUSION AND FUTURE WORK

Based on the “Sansha No. 2” field-test platform, this paper systematically investigated 5G link stability mechanisms in extreme maritime conditions across an ultra-long-distance route. Our findings indicate that while hardware-centric enhancement strategies extend coverage boundaries, they exhibit diminishing marginal returns in addressing nonlinear, stochastic signal fluctuations. This confirms that power compensation alone cannot fundamentally resolve deep-sea 5G interruptions, necessitating an active link-aware mechanism at the software level.

The proposed lightweight early-warning scheme, integrating sliding window feature extraction with RF, achieved a prediction accuracy of 94.95% with millisecond-level inference overhead on a real-world dataset. This approach transitions maritime communication from passive failure recovery to proactive predictive maintenance, demonstrating the practical value of edge intelligence for MIoT robustness. Ultimately, integrating edge intelligence with physical-layer perception is critical for high-reliability deep-sea connectivity. Future research will focus on adaptive federated learning mechanisms and migrating this framework to 6G NTN to support resilient global ocean sensing networks.

ACKNOWLEDGMENT

This research was supported by National Natural Science Foundation of China Under the Grant No. 62341127, “Pioneer

Leading Goose + X” Science and Technology Program of Zhejiang Province under the Grant No. 2025C02018, Zhoushan Science and Technology R&D Project Under the Grant No. 2024C03007, General Scientific Research Fund of Zhejiang Provincial Education Department under Grant No. Y202560032, and Fund Project of Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resource under Grant No. MESTA-2024-B005.

REFERENCES

- [1] J. Wang, Y. Li, X. Zhang, and Z. Chen, "Wireless channel models for maritime communications," *IEEE Access*, vol. 6, pp. 68070–68088, Nov. 2018, doi: 10.1109/ACCESS.2018.2879653.
- [2] R. He, X. Gao, L. Chen, Y. Li, and J. Zhang, "A survey of 5G channel measurements and models," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1588–1616, Aug. 2021, doi: 10.1109/COMST.2021.3090193.
- [3] T. Zhao, S. Sarkar, Y. Tian, and D. Cabric, "Anomaly transmitter recognition and tracking," in *Proc. IEEE Int. Symp. Dynamic Spectrum Access Networks (DySPAN)*, May 2024, pp. 357–364, doi: 10.1109/DySPAN61131.2024.10547477.
- [4] T. Yang, Z. Cui, R. Wang, J. Zhao, Z. Su, and R. Deng, "A multivessels cooperation scheduling for networked maritime fog-ran architecture leveraging SDN," *Peer-to-Peer Netw. Appl.*, vol. 11, no. 4, pp. 808–820, Jul. 2018, doi: 10.1007/s12083-017-0610-8.
- [5] Y. Xu, "Quality of service provisions for maritime communications based on cellular networks," *IEEE Access*, vol. 5, pp. 23881–23890, Nov. 2017, doi: 10.1109/ACCESS.2017.2764003.
- [6] M. G. Kibria, K. Nguyen, G. P. Villardi, O. A. Dobre, and S. J. Bae, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3238–3274, Fourth Quart. 2018, doi: 10.1109/COMST.2018.2845765.
- [7] L. Liang, H. Ye, and G. Y. Li, "Toward intelligent vehicular networks: A machine learning framework," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 124–135, Feb. 2019, doi: 10.1109/JIOT.2018.2846263.
- [8] B. Sudharsan, J. G. Breslin, and M. I. Ali, "Adaptive strategy to improve the quality of communication for IoT edge devices," in *Proc. IEEE 6th World Forum Internet Things (WF-IoT)*, Jun. 2020, pp. 1–6, doi: 10.1109/WF-IoT48130.2020.9221151.
- [9] Y. Yu, L. Huang, and X. Wang, "Time series outlier detection based on sliding window prediction," *Math. Probl. Eng.*, vol. 2014, pp. 1–10, Jul. 2014, doi: 10.1155/2014/879736.
- [10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [11] M. Krichen and A. Mihoub, "Long short-term memory networks: A comprehensive survey," *AI*, vol. 6, no. 9, Art. no. 215, pp. 1–18, Sep. 2025, doi: 10.3390/ai6090215.
- [12] J. P. P. M. Yucra, D. J. C. Quina, G. A. E. Espinoza, M. A. V. Solis, D. D. Y. A. Cardenas, and A. O. Salazar, "Design and implementation of an LSTM model with embeddings on MCUs for prediction of meteorological variables," *Sensors*, vol. 25, no. 12, Art. no. 3601, pp. 1–16, Jun. 2025, doi: 10.3390/s25123601.
- [13] G. Maor, A. M. Abdelrahim, and R. Berkovich, "An FPGA implementation of stochastic computing-based LSTM," in *Proc. IEEE 37th Int. Conf. Comput. Design (ICCD)*, Nov. 2019, pp. 1–6, doi: 10.1109/ICCD46524.2019.00012.
- [14] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
- [15] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the Gini index and information gain criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, May 2004, doi: 10.1023/B:AMAI.0000018580.96245.c6.
- [16] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.
- [17] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.