Ivan Chorbev • Dejan Spasov •
Biljana Risteska Stojkoska (Eds.)

AI AND THE DIGITAL FRONTIER: RESHAPING MODERN SOCIETY THROUGH TECHNOLOGY AND COMPUTER SCIENCE

WEB PROCEEDINGS

17TH ICT INNOVATIONS CONFERENCE

11-13 OCTOBER 2025
METROPOL LAKE RESORT, OHRID, MACEDONIA
ONLINE EDITION PUBLISHED ON HTTP://ICTINNOVATIONS.ORG

SKOPJE, 2026

Ivan Chorbev • Dejan Spasov • Biljana Risteska Stojkoska (Eds.)

17TH ICT INNOVATIONS CONFERENCE,

AI and the Digital Frontier: Reshaping Modern Society Through Technology and Computer Science

WEB PROCEEDINGS

11-13 October 2025

Metropol Lake Resort, Ohrid, Macedonia

ISBN 978-608-65468-5-4 © ICT ACT

Publisher: Society of Information and Communication Technologies (ICT-ACT)

Online edition published on http://ictinnovations.org

Skopje, 2026

Edited by: Ivan Chorbev, Dejan Spasov, Biljana Risteska Stojkoska

Technical support: Ilinka Ivanoska, Zorica Karapancheva, Mila Dodevska

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004:621.39(062)

ICT innovations conference (17 ; 2026 ; Ohrid)
17th ICT innovations conference [Електронски извор] : AL and the digital frontier: reshaping modern society through technology and computer science : web proceedings : 11-13 October 2025 Metropol Lake Resort, Ohrid, Macedonia / (eds.) Ivan Chorbev, Dejan Spasov, Biljana Risteska Stojkoska. - Skopje : Society of information and communication technologies ICT-ACT, 2026

Начин на пристапување (URL):
https://ictinnovations.org/wp-content/uploads/2026/01/WebProceedings2025.pdf.
- Текст во PDF формат, содржи XII, 264 стр., илустр. - Наслов преземен од екранот. - Опис на изворот на ден 29.01.2026. - Библиографија кон трудовите

ISBN 978-608-65468-5-4

а) Информациско-комуникациски технологии -- Примена -- Собири

COBISS.MK-ID 68005381

Preface

We are pleased to present the proceedings of the 17th International Conference ICT Innovations 2025, held in Ohrid, North Macedonia, from October 11–13, 2025. This edition of the conference continues the long-standing tradition of ICT Innovations as a premier international forum for the exchange of cutting-edge research and innovative ideas in the broad field of information and communication technologies.

The ICT Innovations conference series, organized by the Macedonian Society of Information and Communication Technologies (ICT-ACT) and supported by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, has built a strong reputation over the years as a platform that successfully bridges fundamental research, applied science, and real-world technological challenges. Since its inception, the conference has continuously evolved to reflect emerging trends and pressing societal needs shaped by rapid technological progress.

This year's theme, "AI and the Digital Frontier: Reshaping Modern Society Through Technology and Computer Science," highlights the transformative role of artificial intelligence and advanced computing technologies in redefining modern life. The conference encourages critical reflection on innovation while addressing the societal, ethical, and practical implications of intelligent systems, digital transformation, and secure computing infrastructures.

ICT Innovations 2025 was held jointly with the International Applied Cybersecurity Conference (IACyC 2025), creating a unique interdisciplinary environment that fostered collaboration between the ICT and cybersecurity research communities. This joint organization strengthened the dialogue between AI-driven technologies, system security, and digital sovereignty, reflecting the growing need for holistic and resilient digital ecosystems.

This volume contained 18 full papers (plus 19 papers in the Web Proceedings edition), which were carefully reviewed and selected from 80 high-quality submissions. These papers covered a wide range of topics, including machine learning, network science, digital transformation, natural language processing, and more. The review process was rigorous, with about 100 reviewers from 35 countries providing detailed feedback. Each submission was evaluated by at least three experts in the field, ensuring that the selected papers met the high standards of academic excellence and originality that this conference is known for. Together, they reflect the diversity and dynamism of contemporary ICT research and its growing influence on society and industry.

The scientific program featured two distinguished keynote speakers whose work exemplifies the convergence of advanced computing, security, and societal impact. Guy Gogniat, Professor of Electrical and Computer Engineering at the Université Bretagne Sud, Lorient, France, delivered a keynote on "A Comprehensive Approach to System-on-Chip Security," addressing hardware-level security

challenges in modern computing systems. Aleksandar Jevremović, Professor at the Faculty of Informatics and Computing, Singidunum University, Belgrade, Serbia, presented a keynote titled "Technological Sovereignty and Cybersecurity," offering critical insights into strategic, technological, and policy-driven aspects of digital independence.

In addition to the main conference sessions, the program was enriched by workshops, special sessions, and joint activities organized within the framework of IACyC 2025 and the CyberMACS Erasmus Mundus Joint Master Degree project, including Advances in AI for Cybersecurity supported by the NATO SPS MYP SENTAI, the Technology Transfer Workshop on Confidentiality of a Large Language Model for Legal Aid, the MKSafeNet Workshop, EIT Community Hub Role in the Emerging Innovation Ecosystem – the Macedonian Case Workshop, and A Review on Applications and Uses of an eIDAS / ETSI CEN Standards-Based SAM (NSSAM) in Artificial Intelligence Workshop. These activities provided participants with opportunities for in-depth discussion, hands-on learning, and interdisciplinary exchange, further strengthening the conference's educational and collaborative mission. The conference also offered a variety of social events aimed at fostering connections among participants, a tradition that has been highly valued since the conference's inception.

We extended our heartfelt thanks to all the authors who contributed their work to this year's proceedings, to the reviewers who ensured a fair and thorough evaluation process, and to all the participants who enriched the conference with their knowledge and expertise. Special thanks went to our generous sponsors, companies Netcetera, Nextsense and IWConnect; also to the organizing and program committees as the technical support team at FCSE, whose dedication and hard work were instrumental in making this conference a successful high-quality international event.

As we concluded this edition of ICT Innovations, we looked ahead with excitement to future conferences, where we would continue to explore the frontiers of ICT and foster innovation across industries and disciplines. We invite you to join us at the 18th ICT Innovations conference in 2026, where we would continue this journey of scientific discovery and collaboration.

Sincerely,

January 2026

ICT Innovations 2025 Conference Chairs,
Ivan Chorbev, Dejan Spasov and Biljana Risteska Stojkoska

Organization

General Chairs

Ivan Chorbev	Ss. Cyril and Methodius University in Skopje, MK
Dejan Spasov	Ss. Cyril and Methodius University in Skopje, MK
Biljana Risteska Stojkoska	Ss. Cyril and Methodius University in Skopje, MK

Program Commitee Chairs

Dejan Spasov	Ss. Cyril and Methodius University in Skopje, MK
Biljana Risteska Stojkoska	Ss. Cyril and Methodius University in Skopje, MK
Ilinka Ivanoska	Ss. Cyril and Methodius University in Skopje, MK

Program Committee

Aleksandar Bojchevski	University of Cologne, DE
Aleksandar Stojmenski	Ss. Cyril and Methodius University in Skopje, MK
Aleksandra Mileva	University Goce Delcev, MK
Alessandro Cantelli-Forti	Lab RaSS National Laboratory - CNIT, IT
Amelia Badica	University of Craiova, RO
Ana Madevska Bogdanova	Ss. Cyril and Methodius University in Skopje, MK
Andrea Kulakov	Ss. Cyril and Methodius University in Skopje, MK
Andrej Brodnik	University of Ljubljana, SI
Andreja Naumoski	Ss. Cyril and Methodius University in Skopje, MK
Antonio De Nicola	ENEA, IT
Antun Balaz	Institute of Physics Belgrade, RS
Arianit Kurti	Linnaeus University, SE
Betim Cico	EPOKA University, Tirana, AL
Biljana Risteska Stojkoska	Ss. Cyril and Methodius University in Skopje, MK
Biljana Mileva Boshkoska	Faculty of information studies, SI
Blagoj Ristevski	Faculty of Information and Communication Technologies Bitola, MK
Bojan Ilijoski	Ss. Cyril and Methodius University in Skopje, MK
Bojana Koteska	Ss. Cyril and Methodius University in Skopje, MK
Boris Delibašić	University of Belgrade - Faculty of Organizational Sciences, RS
Dejan Gjorgjevikj	Ss. Cyril and Methodius University in Skopje, MK
Dejan Spasov	Ss. Cyril and Methodius University in Skopje, MK
Dilip Patel	London South Bank University, UK
Dimitar Trajanov	Ss. Cyril and Methodius University in Skopje, MK
Edmond Jajaga	University for Business and Technology, XV
Eftim Zdravevski	Ss. Cyril and Methodius University in Skopje, MK

Matus Pleva	Technical University of Košice, SVK
Melanija Mitrović	University of Niš, RS
Mile Jovanov	Ss. Cyril and Methodius University in Skopje, MK
Milos Jovanovik	Ss. Cyril and Methodius University in Skopje, MK
Milos Stojanovic	Visoka tehnicka skola Nis, RS
Miroslav Mirchev	Ss. Cyril and Methodius University in Skopje, MK
Monika Simjanoska	Ss. Cyril and Methodius University in Skopje, MK
Natasha Ilievska	Ss. Cyril and Methodius University in Skopje, MK
Natasha Stojkovikj	University Goce Delcev, MK
Nevena Ackovska	Ss. Cyril and Methodius University in Skopje, MK
Novica Nosović	Faculty of Electrical Engineering, University of Sarajevo, BiH
Özge Büyükdağlı	International University of Sarajevo, BiH
Pance Ribarski	Ss. Cyril and Methodius University in Skopje, MK
Pece Mitrevski	University St. Kliment Ohridski, Faculty of ICT - Bitola, MK
Periklis Chatzimisios	International Hellenic University, GR
Petar Sokoloski	Ss. Cyril and Methodius University in Skopje, MK
Petre Lameski	Ss. Cyril and Methodius University in Skopje, MK
Riste Stojanov	Ss. Cyril and Methodius University in Skopje, MK
Rossitza Goleva	New Bulgarian University, BG
Sashko Ristov	University of Innsbruck, AT
Sasho Gramatikov	Ss. Cyril and Methodius University in Skopje, MK
Sergio Ilarri	University of Zaragoza, ES
Shuxiang Xu	University of Tasmania, AU
Simona Samardjiska	Radboud University, NL
Slobodan Kalajdziski	Ss. Cyril and Methodius University in Skopje, MK
Smilka Janeska Sarkanjac	Ss. Cyril and Methodius University in Skopje, MK
Snezana Savoska	Faculty of Information and Communication Technologies, Bitola, MK
Stanimir Stoyanov	University of Plovdiv "Paisii Hilendarski", BG
Suzana Loshkovska	Ss. Cyril and Methodius University in Skopje, MK
Tarik Namas	International University of Sarajevo, BiH
Ustijana Rechkoska-Shikoska	UIST - Ohrid, MK
Vacius Jusas	Kaunas University of Technology, LT
Verica Bakeva	Ss. Cyril and Methodius University in Skopje, MK
Vesna Dimitrievska Ristovska	Ss. Cyril and Methodius University in Skopje, MK
Vesna Dimitrova	Ss. Cyril and Methodius University in Skopje, MK
Vesna Dimitrievska	Silicon Austria Labs, Villach, AT
Vladimir Trajkovik	Ss. Cyril and Methodius University in Skopje, MK
Vladimír Siládi	Matej Bel University, SVK
Zlatko Varbanov	Veliko Tarnovo University, BG

Technical Committee

Table of Contents

Session 1

Reinforcement Learning for Energy-Efficient Job Orchestration: A Lightweight Evaluation Framework 2
Enes Bajrami

Real-time Semantic Segmentation in Remote Sensing with PIDNet 20
Marko Petrov, Ivica Dimitrovski, Ema Pandilova, Vlatko Spasev, Ivan Kitanovski and Pance Ribarski

Stock Trading Recommendations Using Deep Q-learning and NLP 33
Kostadina Veljanovska, Simeon Nalovski, Blagoj Ristevski and Snezana Savoska

An exploratory paper into Mixture of Experts and their application in Autonomous Vehicles 48
Vladimir Djepovski and Petre Lameski

A Machine Learning Pipeline for Enhanced Speaker Diarization Using Segmentation, VAD, and GMM-Based Clustering 65
Dimitar Marenoski, Mile Pelivanov, Jovan Kalajdzieski and Kostadin Mishev

Session 2

Changes in startup companies brought by the application of blockchain . . 79
Nebojsha Todorovikj and Smilka Janeska Sarkanjac

From Gatekeeping to Empowerment: Redefining IT as a Strategic Enabler for Human-Centric AI Integration in Industry 5.0 93
Darko Poposki

Designing a Digital Architecture for Forensic Case and Evidence Management System 107
Slobodan Oklevski, Ivan Chorbev and Mario Loleski

TravelSage: A Database-Driven Platform for Personalized Travel Planning 122
Sandra Ilievska, Zorica Karapancheva, Jordancho Eftimov and Ivan Chorbev

Session 3

Extracting Knowledge from Time Series Data: Digital Trends in the Balkans 142
Teodora Siljanoska, Natasha Blazheska-Tabakovska and Snezana Savoska

Session 4

Empowering Educators with AI-Enhanced Media Literacy and Cybersecurity Education: A Methodology that utilizes Participatory Action Research Approach 158
Vladimir Trajkovikj and Maja Videnovik

Towards Privacy-Preserving AI in Educational Platforms: Backend Strategies for Secure Data Analytics and Threat Detection 173
Jovana Trajcheska, Ivan Chorbev, Dejan Gjorgjevikj and Boban Joksimoski

Session 5

Evaluating LLMs on the Extractive Question-Answering Task in Macedonian 190
Stefan Milev, Monika Simjanoska Misheva and Kostadin Mishev

Context-Aware Information Retrieval in Workplace Messaging Systems via Retrieval-Augmented Generation and Vector-Based Memory 202
Ema Pandilova, Marija Maneva, Andrej Petkovikj, Ana Markovska, Vesna Pop-Dimitrijoska Koteska, Pance Ribarski and Bojan Ilijoski

LLM Chatbot with SQL Database 217
Dean Nastevski, Lazo Nikoloski, Dimitar Kitanovski, Zorica Karapancheva, Aleksandar Stojmenski, Ivan Chorbev and Petre Lameski

Session 6

Artificial Intelligence in Medicine in Macedonia 231
Dragica Bliznakovska Stanchev, Smilka Janeska Sarkanjac and Sinisha Stanchev

Higuchi's Fractal Dimension in EEG signals of Children with Autism and Typical Development 240
Aleksandar Tenev, Silvana Markovska-Simoska and Igor Mishkovski

Identifying medical ontologies in MIMIC-IV dataset 250
Zorica Karapancheva, Mirjana Sadikovikj and Goran Velinov

Session 1

Reinforcement Learning for Energy-Efficient Job Orchestration: A Lightweight Evaluation Framework

Enes Bajrami[1][0009−0005−7960−3959], Boro Jakimovski[1][0000−0001−9434−7475], Andrea Kulakov[1][0000−0002−7075−0953], and Petre Lameski[1][0000−0002−5336−1796]

Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Ruger Boskovik 16, 1000, Skopje, Republic of North Macedonia
enes.bajrami@students.finki.ukim.mk, {boro.jakimovski, andrea.kulakov, petre.lameski}@finki.ukim.mk

Abstract. This paper presents a Deep Reinforcement Learning (DRL) framework for energy-aware job orchestration on real computing infrastructure. The framework integrates a Proximal Policy Optimization (PPO) agent with a software-based monitoring layer to learn adaptive scheduling policies under thermal, performance, and service-level constraints. A 500-job synthetic workload was executed on physical hardware using 5-fold cross-validation, enabling structured comparison between FIFO scheduling and the DRL-based approach. The reward function combines penalties for CPU temperature, SLA violations, and migration overhead, with component-level contributions logged throughout training for interpretability. Experimental results show that the DRL scheduler reduced total execution time from 16,180 to 3,236 seconds on average while stabilizing CPU thermal profiles and avoiding temperature spikes. All experiments were conducted on a controlled local testbed to ensure reproducibility and precise monitoring. The framework is designed for cloud deployment, with future work focusing on extended action spaces, integration of real SLA definitions, and evaluation under production-scale workloads.

Keywords: Deep Reinforcement Learning · Job Scheduling · Energy Efficiency · Proximal Policy Optimization · Resource Optimization.

1 Introduction

The increasing need for intelligent services in domains such as smart cities, transportation, healthcare, and industrial automation has accelerated the development of computing systems that support scalable and responsive execution. These systems are often required to handle dynamic workloads and optimize for multiple objectives, including energy efficiency and real-time responsiveness [1]. However, traditional scheduling and orchestration mechanisms, while effective under static conditions, often fall short in dynamic environments where system states fluctuate and user demands vary [2, 3].

In particular, First-In-First-Out (FIFO) and rule-based orchestration methods typically fail to account for runtime resource contention, thermal throttling, and Quality of Service (QoS) constraints. These limitations are further magnified in real-time execution scenarios where resource provisioning must adapt rapidly to current system metrics such as CPU usage, thermal state, and job concurrency [4, 5]. Although heuristic-based solutions, including Genetic Algorithms and Particle Swarm Optimization, offer some flexibility, they often lack adaptability during execution and do not generalize well across workload types [6, 7].

Recent research has turned to Deep Reinforcement Learning (DRL) to address these challenges by enabling agents to learn optimal policies through interaction with the environment. DRL-based orchestration systems have the potential to dynamically adjust job placement, scaling, and migration decisions in response to real-time conditions. When carefully designed, these systems can reduce execution latency, balance resource usage, and minimize violations of Service-Level Agreements (SLAs) [8, 9].

This paper introduces a Deep Reinforcement Learning (DRL)-based orchestration framework tailored for energy-aware job scheduling in dynamic computing environments. The system leverages runtime software-level metrics, including CPU utilization, memory usage, and a software-proxied CPU temperature, to inform scheduling decisions made by a Proximal Policy Optimization (PPO) agent. The framework is implemented entirely in Python and evaluated on a local hardware testbed using a synthetic dataset of 500 jobs, with reproducibility ensured through a 5-fold cross-validation procedure.

The reward formulation integrates penalties for thermal behavior, service-level agreement (SLA) violations, and migration overhead into a unified signal. This design enables the agent to learn context-sensitive orchestration strategies that balance multiple objectives simultaneously. To enhance interpretability, the learning process is analyzed through per-episode reward curves and component-wise tracking of λ-weighted contributions.

Experimental results demonstrate that PPO-based orchestration achieves substantial improvements in both execution time and thermal regulation compared to FIFO scheduling, even on resource-constrained hardware without physical sensors. The framework is intended for deployment in cloud environments, with future extensions focusing on real workloads, distributed infrastructures, and integration of production-level SLA definitions.

2 Related Work

Recent advances in cloud and edge computing have intensified the demand for intelligent orchestration mechanisms capable of managing low-resource, dynamic environments. Traditional cloud infrastructure, while effective for general-purpose computation and storage, often struggles to support latency-sensitive or real-time applications. To address this, serverless computing models have emerged, particularly Function as a Service (FaaS), which decouples execution

logic from infrastructure provisioning and enables elastic and cost-efficient deployment. However, orchestrating microservices across distributed and heterogeneous edge environments remains challenging due to limited observability and unpredictable resource availability [10].

Several Deep Reinforcement Learning (DRL) based orchestration frameworks have been proposed to address these issues. One example is a DRL model for serverless FaaS environments that dynamically assigns event-driven functions across edge nodes. This method improves responsiveness while reducing latency and resource consumption, even under partial observability. It has demonstrated an 18 percent improvement over standard strategies such as Shortest Path and Load Balancing for function composition efficiency [11]. Another line of work focuses on mobility-aware orchestration. The DeepEdge model applies DRL in edge computing environments with user mobility and varying network conditions. It uses Double Deep Q-Networks (DDQN) to optimize task offloading. Tested on applications like augmented reality and healthcare, the system demonstrated improved task satisfaction and adaptability [12].

Further research has explored DRL applications in Extreme Edge Computing (EEC), where computation is moved to mobile or remote edge devices to reduce energy consumption and environmental impact. The DeTOrch framework combines DRL with task partitioning and mobility-aware scheduling to coordinate interdependent tasks in constrained environments. Experimental results show reductions in task completion time and improved resource utilization across various mobility scenarios [13]. These systems highlight the potential of DRL to enhance orchestration in highly dynamic infrastructures.

Despite these advances, many DRL based orchestration systems remain narrow in scope. Most focus on isolated challenges such as mobility or serverless deployment and lack comprehensive integration of real-time telemetry, SLA-awareness, and energy efficiency. Furthermore, few are validated in reproducible and lightweight environments that reflect realistic single-node system constraints. The framework proposed in this paper addresses these gaps by emphasizing job scheduling efficiency and energy-aware orchestration. It is evaluated on a local computing system using real workload traces to ensure transparency, modularity, and practical applicability under constrained conditions.

3 Theoretical Background: Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) merges Reinforcement Learning (RL) with Deep Learning (DL), enabling agents to learn optimal decision-making strategies in high-dimensional, uncertain environments [14]. Unlike supervised learning, DRL agents interact directly with an environment to explore and exploit actions that maximize long-term rewards, even when the state-action space is continuous or partially observable [15].

3.1 Markov Decision Process and Agent Objective

DRL problems are framed as Markov Decision Processes (MDPs), denoted by the tuple (S, A, P, R, γ), where:

- S is the state space,
- A is the action space,
- $P(s'|s, a)$ is the transition probability function,
- $R(s, a)$ defines the reward for taking action a in state s,
- γ is the discount factor [14].

The agent's goal is to learn a stochastic policy $\pi(a|s)$ that maximizes the expected cumulative discounted return:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{1}$$

as defined in the canonical RL objective [15].

3.2 Value Functions and Bellman Equations

The state-value and action-value functions evaluate expected returns:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s\right], \quad Q^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s, a_0 = a\right] \tag{2}$$

These functions satisfy the Bellman equations:

$$Q^\pi(s, a) = R(s, a) + \gamma\sum_{s'} P(s'|s, a)V^\pi(s') \tag{3}$$

which are foundational to value-based approaches including Deep Q-Networks (DQN) [16].

3.3 Policy Gradient and Actor–Critic Methods

Policy gradient algorithms optimize a parameterized stochastic policy $\pi_\theta(a|s)$ by estimating the gradient of the expected return:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)Q^\pi(s_t, a_t)\right] \tag{4}$$

as formalized in the REINFORCE algorithm [17].

Actor–Critic frameworks introduce a critic to estimate the value function, enabling variance reduction. The advantage function used in training is defined as:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s) \tag{5}$$

3.4 Proximal Policy Optimization (PPO)

Proximal Policy Optimization improves training stability by limiting the size of policy updates. The clipped objective is:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right] \tag{6}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio between the new and old policy [18].

3.5 Soft Actor–Critic (SAC)

SAC augments the standard objective by maximizing expected return along with the policy entropy to improve robustness and exploration:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \left[r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right] \tag{7}$$

where $\mathcal{H}$ denotes the entropy of the policy and α is the temperature parameter controlling the entropy scale [19].

3.6 Architectural Overview

The practical implementation of DRL requires neural networks to approximate both policy and value functions. This results in an architecture where agents continuously interact with their environments to update parameters through backpropagation.

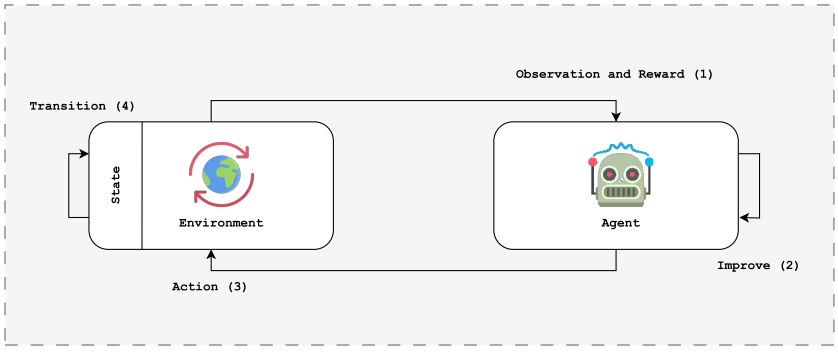

Fig. 1. Conceptual architecture of the DRL agent-environment interaction loop.

The interaction between agent and environment in Deep Reinforcement Learning (DRL) follows a standard loop in which the agent observes the system state s_t, selects an action a_t, receives a reward r_t, and updates its policy π_θ based

on the observed transition. This framework enables scalable learning and adaptive decision-making across domains such as microservice orchestration, robotic control, and energy-aware scheduling [14, 16, 19, 20].

However, challenges remain in improving sample efficiency, ensuring convergence stability, and reducing sensitivity to hyperparameter tuning [20].

4 Methodology

4.1 Research Questions

The study is guided by the following research questions:

- **RQ1:** Can a PPO-based DRL agent improve execution time and throughput compared to FIFO scheduling under dynamic workloads?
- **RQ2:** Can the agent regulate thermal behavior and reduce CPU temperature spikes while maintaining SLA adherence?
- **RQ3:** How interpretable are the contributions of different reward components (temperature, SLA, migration) to the overall policy optimization?
- **RQ4:** Does cross-validation across diverse workload folds confirm the generalization ability of the PPO agent?
- **RQ5:** How does the proposed framework compare to existing DRL-based orchestration approaches in terms of experimental realism, interpretability, and multi-objective optimization?

4.2 Overview of the Framework

This study presents a Deep Reinforcement Learning (DRL)-based orchestration framework designed to enhance energy efficiency and workload scheduling in real computing environments. The system is implemented entirely in Python and integrates three core modules: (i) a software-based monitoring layer for collecting real-time system metrics, (ii) a Proximal Policy Optimization (PPO) decision engine, and (iii) an execution layer that applies orchestration decisions at runtime. All tests were conducted on physical hardware to ensure reproducibility and precise monitoring of resource dynamics.

4.3 Testing Environment and Dataset

Experiments were carried out on a personal computing system with an Intel Core i3-7020U CPU at 2.30GHz. A workload of 500 synthetic jobs was constructed to simulate realistic orchestration tasks. Each job entry specifies an arrival timestamp, an execution duration, and a memory requirement. To ensure fair evaluation, job arrival times were uniformly scaled, and deadlines were defined as a fixed multiple of the job duration.

The experimental evaluation consisted of two phases:

1. Baseline Phase: Jobs were executed sequentially using a First-In-First-Out (FIFO) policy. Each job performed a 200×200 matrix multiplication to generate CPU load. Metrics such as CPU utilization, memory usage, and estimated temperature were logged. The temperature was approximated using a software proxy $T = 40 + 0.5 \times \text{CPU\_load}$ and further smoothed with an exponential moving average.
2. DRL Phase: The same workload was executed under the DRL-based orchestration framework. The PPO agent observed system states in real time and selected among four actions: allocate, migrate, scale, or idle. Each job performed a 1300×1300 matrix multiplication to increase computational intensity and test the agent's ability to manage thermal and resource constraints under higher load.

4.4 DRL Formulation and Agent Design

The orchestration problem was modeled as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ):

- S denotes the system state, including CPU utilization, estimated CPU temperature, and memory usage,
- A represents the discrete action space: *allocate*, *migrate*, *scale*, or *idle*,
- P is the (unknown) state transition probability,
- R is the immediate reward signal,
- γ is the discount factor prioritizing long-term gains.

The reward function was formulated as a weighted sum of three penalties:

$$r_t = -\lambda_1 T_t - \lambda_2 V_t - \lambda_3 C_t \tag{8}$$

where T_t is the estimated CPU temperature, V_t is an SLA violation indicator (deadline miss), and C_t is the migration overhead proportional to requested memory. The scalar weights λ_1, λ_2, and λ_3 were fixed during experiments to emphasize thermal stability, service quality, and responsiveness. To preserve interpretability, the contribution of each component was logged separately at every step, enabling detailed analysis of the agent's optimization behavior.

The PPO agent architecture comprised two fully connected hidden layers with ReLU activation and a softmax output layer producing action probabilities. The policy was optimized using the Adam optimizer. All system metrics were sampled directly via the `psutil` Python library without simulation or emulation layers, ensuring that training and testing were performed in real time.

4.5 Cross-Validation and Logging

To ensure statistically robust evaluation, a 5-fold cross-validation scheme was employed. The 500-job dataset was partitioned into five sequential subsets of 100 jobs. In each fold, one subset served as the test set (100 jobs), while the remaining 400 jobs were used for training:

- Fold 1: Test [1–100], Train [101–500]
- Fold 2: Test [101–200], Train [1–100, 201–500]
- Fold 3: Test [201–300], Train [1–200, 301–500]
- Fold 4: Test [301–400], Train [1–300, 401–500]
- Fold 5: Test [401–500], Train [1–400]

This procedure ensured that each job was included once in testing and four times in training, providing diverse workload exposure across folds. During training, total reward per episode was logged to assess policy convergence, while the per-step contributions of $\lambda_1 T_t$, $\lambda_2 V_t$, and $\lambda_3 C_t$ were recorded to illustrate how the agent balanced competing orchestration goals. In the testing phase, detailed job logs were collected, including start time, execution duration, CPU usage, thermal proxy, action taken, SLA violation status, and reward decomposition.

4.6 Proposed Framework Architecture

Figure 2 presents the architecture of the proposed DRL-based orchestration framework, which is structured into three main layers. The Monitoring Layer collects runtime metrics including CPU utilization, memory consumption, job arrival times, and software-based thermal estimates. These values represent the current system state and provide the necessary input for decision making.

The Decision Layer contains the Proximal Policy Optimization (PPO) agent, which maps observed states to scheduling actions. The action space consists of allocate, migrate, scale, and idle. The agent selects actions according to a reward function that incorporates thermal penalties, SLA violations, and migration costs. This formulation enables the policy to balance execution efficiency with system stability and service quality.

The Execution Layer applies the chosen actions on the local infrastructure in real time. Jobs are executed, migrated, or scaled depending on the PPO policy, while updated system metrics are fed back into the Monitoring Layer. This closed-loop process ensures that orchestration decisions continuously adapt to workload dynamics. The framework is implemented entirely in Python, allowing reproducible evaluation on lightweight hardware and offering extensibility for future deployment in distributed cloud environments.

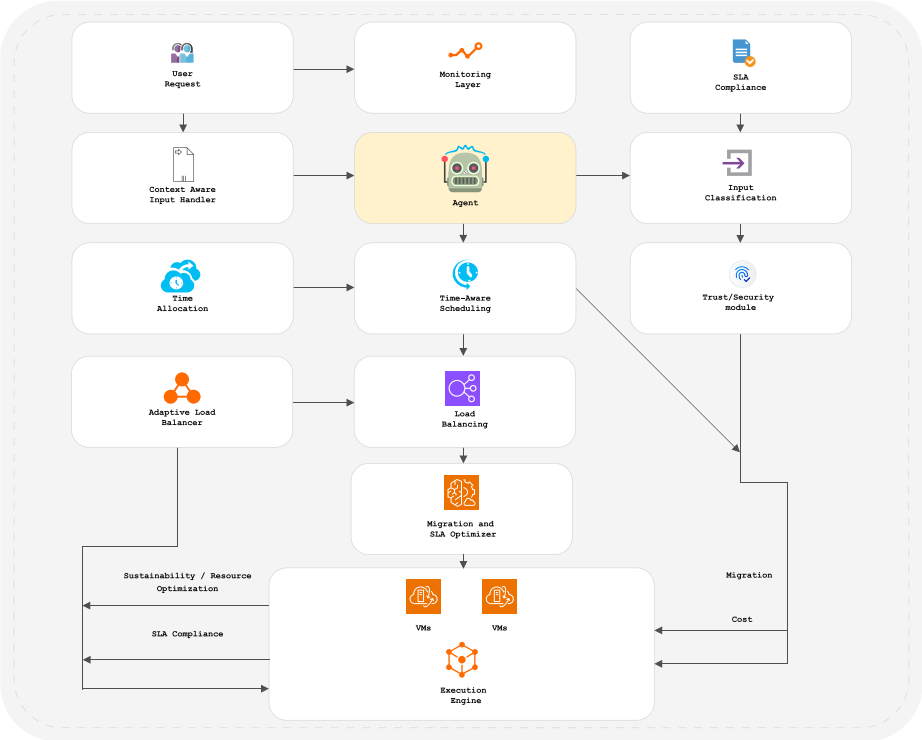

Fig. 2. Proposed DRL-Orchestrated Framework for Energy and SLA-Aware Job Scheduling on Local Infrastructure.

5 Results

This section presents the evaluation of the proposed PPO-based orchestration strategy using real hardware experiments on a synthetic workload of 500 jobs. The experiments were conducted using a structured 5-fold cross-validation procedure to ensure statistical robustness and reproducibility. In each fold, the PPO agent was trained on 400 jobs and tested on the remaining 100, so that every job in the dataset was used once for testing and four times for training. This procedure ensured a balanced distribution of workload and temporal behavior across folds, allowing the framework to be evaluated under diverse job sequences.

The evaluation focuses on multiple aspects of system performance, including reward progression during training, the influence of individual reward components, thermal regulation, execution time, and job-level behavior. Comparisons are drawn with the FIFO baseline to highlight the effectiveness of the DRL approach in improving throughput and thermal stability while reducing SLA violations. Results are reported both in terms of aggregate statistics across folds and detailed per-job analyses to illustrate the consistency and reliability of the proposed framework.

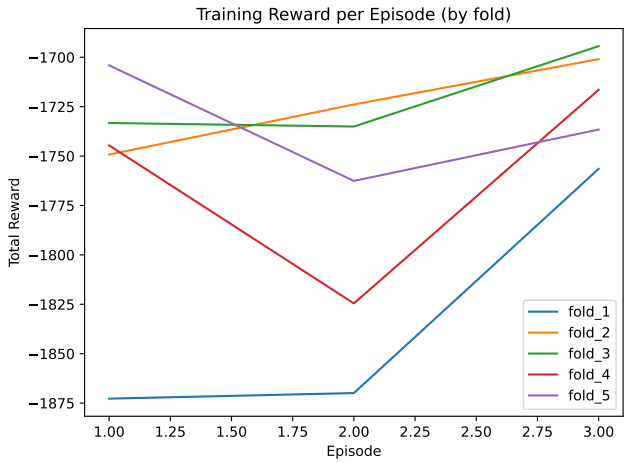

Fig. 3. Training reward per episode across five folds.

Figure 3 shows the cumulative reward progression during training. Each curve corresponds to a fold, highlighting differences in workload distribution. Despite fold-dependent fluctuations, the overall trend confirms that the PPO agent adapts and stabilizes its policy effectively. This indicates robust generalization across varying workload subsets.

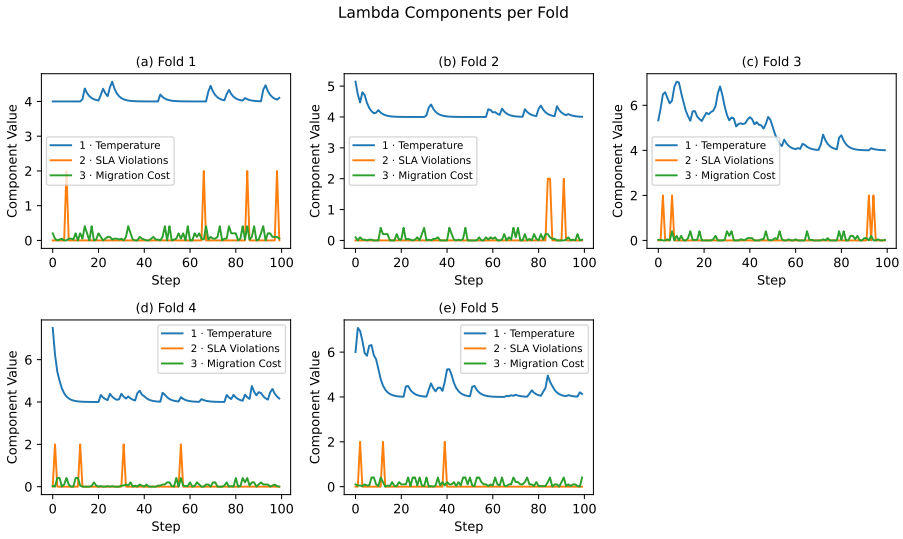

Fig. 4. Lambda component dynamics across folds 1–5.

Figure 4 illustrates the contributions of temperature, SLA, and migration penalties over training steps. Temperature is consistently the dominant factor, confirming its role as the primary optimization driver. SLA violations appear intermittently when deadlines are missed, while migration costs remain relatively small. These dynamics demonstrate how the PPO agent balances multiple objectives during orchestration.

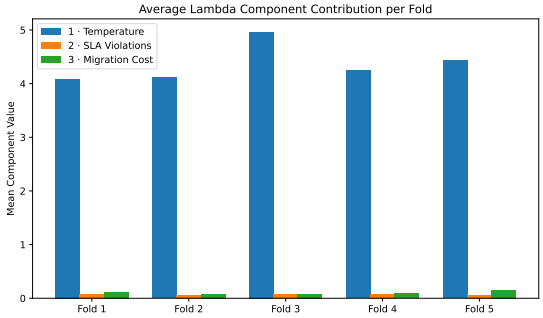

Fig. 5. Average reward component contributions per fold.

Figure 5 summarizes mean component contributions across all folds. The temperature penalty $\lambda_1 \cdot T$ dominates consistently, while SLA and migration penalties contribute smaller but steady effects. This distribution confirms the effectiveness of the multi-objective reward in prioritizing thermal stability without ignoring performance and cost.

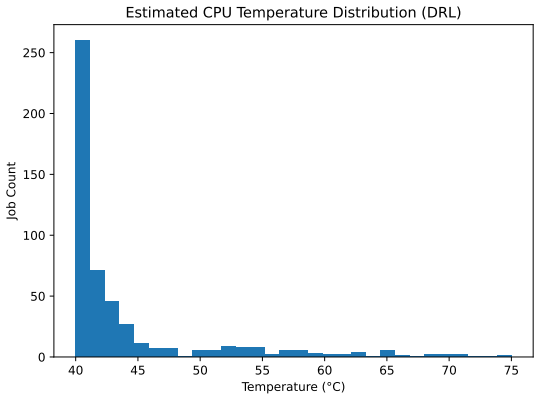

Fig. 6. Distribution of CPU temperatures under DRL orchestration.

Figure 6 shows the estimated CPU temperature distribution across 500 jobs orchestrated by the PPO agent. Most jobs remain below 45°C, indicating effective thermal control. Only a sparse tail reaches higher temperatures, associated with intensive workloads. The distribution demonstrates that overheating is rare and well managed by the proposed framework.

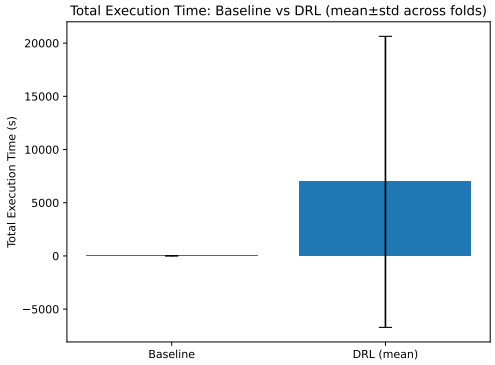

Fig. 7. Total execution time: Baseline vs. DRL.

Figure 7 compares cumulative execution time of 500 jobs under baseline FIFO scheduling and DRL orchestration. The baseline records significantly longer duration, while the PPO agent achieves a substantial reduction in total job completion time. Error bars show the standard deviation across folds, confirming that improvements are consistent despite workload variability. These results demonstrate the throughput advantages of DRL orchestration.

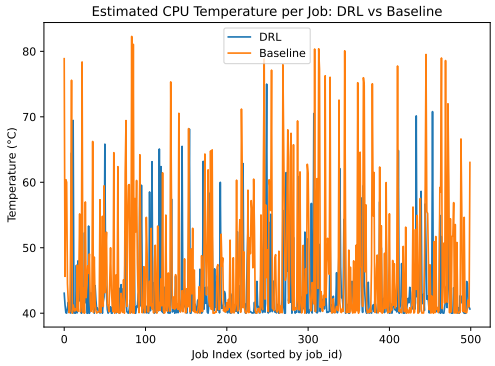

Fig. 8. Job-wise CPU temperature comparison between DRL and Baseline.

Figure 8 presents the evolution of job-level temperatures under both methods. Baseline scheduling exhibits frequent and high peaks, reflecting inefficient thermal control. In contrast, DRL orchestration maintains lower and smoother temperature trajectories across the workload. This behavior reduces thermal stress and supports energy-efficient operation.

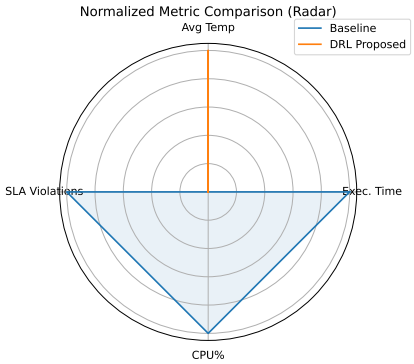

Fig. 9. Normalized radar comparison of DRL and Baseline across key metrics.

Figure 9 compares DRL and baseline across execution time, average temperature, SLA violations, and CPU utilization. The PPO-based approach outperforms the baseline in every dimension, highlighting improvements in performance, efficiency, and reliability. Lower execution time and reduced SLA violations validate scheduling efficiency, while reduced temperatures confirm effective thermal management. This multi-metric analysis underscores the robustness of the proposed orchestration framework.

6 Discussion

The proposed DRL-based orchestration framework is designed for deployment in cloud environments where dynamic workloads, multi-objective optimization, and heterogeneous resources are common. In this study, the evaluation was conducted on a controlled local infrastructure to ensure reproducibility and detailed monitoring of system metrics. This setup preserved the modularity of the framework and allowed workload variability to be tested, creating a foundation for eventual deployment in distributed cloud platforms.

The evaluation results confirm that reinforcement learning improves job orchestration compared to static FIFO scheduling. The PPO agent continuously adapts its policy in response to changes in workload characteristics and system state. This adaptability leads to more efficient resource utilization, reduced thermal stress, and higher throughput. Unlike static schedulers, the agent demon-

strates the ability to balance trade-offs between competing objectives through experience-driven learning.

Figure 3 presents the training progression across five folds. Despite differences in workload distribution, the reward trajectory improves over episodes, showing the ability of the PPO agent to derive stable orchestration strategies from policy gradient updates. This stability is critical for production systems that must manage unpredictable workloads.

A key feature of the framework is its multi-component reward design. Figure 4 illustrates how temperature, SLA violations, and migration cost penalties evolve across all folds. Temperature remains the dominant factor influencing learning, while SLA and migration costs are activated under specific scheduling decisions such as delayed placement or aggressive scaling. The mean values in Figure 5 confirm that the thermal component ($\lambda_1 \cdot T$) is the most influential term, which aligns with the goal of prioritizing thermal stability as a primary orchestration constraint.

Thermal regulation under DRL scheduling shows marked improvements. As shown in Figure 6, most job executions remain within the 40–45°C range, with only sparse instances of higher values. The job-level view in Figure 8 further demonstrates that DRL scheduling avoids extreme thermal spikes that appear under FIFO. This improvement supports sustained system reliability by reducing the likelihood of throttling and hardware degradation.

Execution time is also significantly improved. Figure 7 shows a reduction in cumulative execution time from 16,180 seconds under FIFO to approximately 3,236 seconds with DRL orchestration, averaged across folds. The addition of error bars highlights the consistency of these improvements despite variability in job distributions. These gains arise from reduced idle time, adaptive load balancing, and efficient overlap of jobs when resources are available.

The radar chart in Figure 9 consolidates improvements across four key dimensions: execution time, temperature, SLA violations, and CPU utilization. DRL scheduling achieves better performance across all criteria, reinforcing its suitability for multi-objective orchestration in real computing environments. By integrating thermal awareness with performance efficiency, the framework offers a practical and interpretable solution for future energy-aware scheduling.

6.1 Answers to Research Questions

RQ1: PPO-based orchestration reduced total execution time from 16,180 seconds under FIFO to an average of 3,236 seconds across folds. This demonstrates a clear improvement in throughput and scheduling efficiency.

RQ2: DRL scheduling maintained CPU temperatures within a narrow 40–45°C range and reduced thermal spikes compared to FIFO. SLA violations were fewer and occurred only in specific high-load cases, confirming the framework's ability to regulate thermal behavior while respecting deadlines.

RQ3: Logging of reward components showed that temperature consistently dominated the learning signal, while SLA and migration penalties were triggered

selectively. This decomposition provided interpretability, allowing the effect of each component on policy learning to be observed directly.

RQ4: The use of 5-fold cross-validation confirmed generalization. Across all folds, the PPO agent improved training rewards and delivered consistent reductions in execution time and thermal stress, even though workload distributions varied.

RQ5: Compared to existing DRL orchestration approaches, the proposed framework demonstrates novelty by combining multi-objective reward design with interpretable component logging, while conducting evaluation directly on real hardware. This contrasts with prior studies that rely primarily on simulated environments or optimize single objectives.

6.2 Comparison with Related Works

Several recent studies have explored DRL-based approaches for resource scheduling and energy optimization. In mobile edge computing, [21] investigated joint computation offloading and resource allocation using Q-learning and Double Deep Q-Networks (DDQN). Their approach effectively reduced energy consumption in dynamic MEC systems but was limited to simulation-based environments and did not address multi-objective orchestration including thermal constraints.

Similarly, [22] proposed a two-stage DRL framework to optimize energy consumption and server load balancing under latency constraints. Their work achieved significant energy savings compared to heuristic methods such as Round Robin, but the evaluation focused primarily on server utilization and energy trade-offs without incorporating detailed SLA penalties or migration costs.

Another direction was presented in [23], which addressed energy optimization in building energy management systems through cloud-based IoT integration. By leveraging PPO for specific tasks such as electric vehicle charging, they achieved notable cost savings and improved demand-side management. However, their focus was on energy-efficient appliance control rather than real-time orchestration of heterogeneous jobs on computing infrastructure.

Compared to these works, the present framework introduces a unified and interpretable reward formulation that simultaneously considers thermal stability, SLA adherence, and migration cost. Moreover, the experiments were performed on real hardware rather than simulations, providing a reproducible and transparent evaluation environment. This combination highlights the originality of the framework and its potential for extension to cloud and edge computing platforms.

7 Conclusion and Future Work

7.1 Conclusion

This study introduced a Deep Reinforcement Learning (DRL)-based orchestration framework for improving job scheduling efficiency and energy-aware resource

allocation. The framework was evaluated against a traditional FIFO scheduler using a Proximal Policy Optimization (PPO) agent, tested on real hardware with 5-fold cross-validation over a 500-job workload. Results show that the DRL agent substantially reduced total workload execution time, from approximately 16,180 seconds under FIFO to an average of 3,236 seconds, while simultaneously maintaining thermal stability and avoiding CPU temperature spikes. These improvements were achieved without predefined heuristics; instead, the PPO agent adapted dynamically through feedback from a unified reward function that combined thermal, SLA, and migration cost penalties. A key contribution of this work is the transparent reward design, which enabled detailed interpretability of learning dynamics across folds. The PPO agent demonstrated robust generalization, effectively balancing performance with system health in dynamic environments. Overall, the results validate DRL as a powerful tool for adaptive orchestration in scenarios where static scheduling strategies are inadequate, particularly in heterogeneous or resource-constrained systems, and provide a strong foundation for extending orchestration intelligence to cloud and edge environments.

7.2 Future Work

While the current evaluation was performed locally to ensure controlled testing and reproducibility, future efforts will focus on deploying the framework in distributed cloud environments with realistic production workloads. Planned extensions include expanding the action space to enable scheduling across multiple nodes, container-level migration, and energy-aware scaling policies. Comparative studies with alternative DRL algorithms such as Soft Actor-Critic (SAC) and Deep Deterministic Policy Gradient (DDPG) will be conducted to assess policy robustness under diverse workload patterns. Another key direction involves incorporating real-world SLA definitions, capturing stricter deadlines, latency objectives, and QoS constraints, to better reflect operational conditions. Finally, long-term experiments on non-stationary workloads will be used to evaluate policy adaptability and stability over time. Together, these developments aim to advance autonomous, interpretable, and energy-efficient orchestration agents for deployment in live cloud platforms.

Declaration of generative AI in scientific writing

During the preparation of this work, the author(s) used AI tools to improve readability and language. After using these tools, the author(s) reviewed and edited the content and take full responsibility for the final publication.

Data Availability

The source code, trained model logs, and all chart generation scripts used in this study are publicly available at: https://github.com/bajramienes/drl-based-

orchestration. The synthetic workload dataset and evaluation environment are also included to ensure full reproducibility.

References

1. Marinescu, D.C.: *Cloud Computing: Theory and Practice*. 3rd edn. Morgan Kaufmann, Waltham, MA (2022). https://doi.org/10.1016/C2020-0-02233-4
2. Duan, S., Wang, D., Ren, J., Lyu, F., Zhang, Y., Wu, H., Shen, X.: Distributed artificial intelligence empowered by end-edge-cloud computing: A survey. *IEEE Commun. Surv. Tutor.* **25**(1), 591–624 (2023). https://doi.org/10.1109/COMST.2022.3213237
3. Luo, Q., Hu, S., Li, C., Li, G., Shi, W.: Resource scheduling in edge computing: A survey. *IEEE Commun. Surv. Tutor.* **23**(4), 2131–2165 (2021). https://doi.org/10.1109/COMST.2021.3106401
4. Yan, J., Huang, Y., Gupta, A., Liu, C., Li, J., Cheng, L., Chung, W.: Energy-aware systems for real-time job scheduling in cloud data centers: A deep reinforcement learning approach. *Comput. Electr. Eng.* **99**, 107688 (2022). https://doi.org/10.1016/j.compeleceng.2022.107688
5. Lin, Z., Bi, S., Zhang, Y., Zhang, Y.-J.A.: Optimizing AI service placement and resource allocation in mobile edge intelligence systems. *IEEE Trans. Wireless Commun.* **20**(11), 7257–7271 (2021). https://doi.org/10.1109/TWC.2021.3081991
6. Houssein, E.H., Gad, A.G., Wazery, Y.M., Suganthan, P.N.: Task scheduling in cloud computing based on meta-heuristics: Review, taxonomy, open challenges, and future trends. *Swarm Evol. Comput.* **62**, 100841 (2021). https://doi.org/10.1016/j.swevo.2021.100841
7. Abadi, Z.J.K., Mansouri, N., Javidi, M.M.: Deep reinforcement learning-based scheduling in distributed systems: A critical review. *Knowl. Inf. Syst.* (2024). https://doi.org/10.1007/s10115-024-01965-9
8. Cheng, F., Huang, Y., Tanpure, B., Sawalani, P., Cheng, L., Liu, C.: Cost-aware job scheduling for cloud instances using deep reinforcement learning. *Cluster Comput.* pp. 1–13 (2022). https://doi.org/10.1007/s10586-022-03567-y
9. Hao, Y., Chen, M., Gharavi, H., Zhang, Y., Hwang, K.: Deep reinforcement learning for edge service placement in softwarized industrial cyber-physical systems. *IEEE Trans. Ind. Inform.* **17**(8), 5552–5561 (2021). https://doi.org/10.1109/TII.2020.3041713
10. Gao, Y., Zhang, C., Xie, Z., Qi, Z., Zhou, J.: Cost-efficient and quality-of-experience-aware player request scheduling and rendering server allocation for edge-computing-assisted multiplayer cloud gaming. *IEEE Internet Things J.* **9**(14), 12029–12040 (2022). https://doi.org/10.1109/JIOT.2021.3132849
11. Khansari, M.E., Sharifian, S.: A deep reinforcement learning approach towards distributed function as a service (FaaS) based edge application orchestration in cloud-edge continuum. *J. Netw. Comput. Appl.* **233**, 104042 (2025). https://doi.org/10.1016/j.jnca.2024.104042
12. Yamansavascilar, B., Baktir, A.C., Sonmez, C., Ozgovde, A., Ersoy, C.: DeepEdge: A deep reinforcement learning based task orchestrator for edge computing. *IEEE Trans. Netw. Sci. Eng.* **10**(1), 538–552 (2023). https://doi.org/10.1109/TNSE.2022.3217311
13. Safavifar, Z., Machalikh, C., Xie, J., Golpayegani, F.: Sustainable dependent sub-tasks orchestration at extreme edge computing: A partitioning-based

deep reinforcement learning approach. *ACM Trans. Internet Technol.* (2025). https://doi.org/10.1145/3723037
14. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. 2nd edn. MIT Press, Cambridge, MA (2018)
15. Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**(6), 26–38 (2017). https://doi.org/10.1109/MSP.2017.2743240
16. Mnih, V. et al.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015). https://doi.org/10.1038/nature14236
17. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**(3–4), 229–256 (1992). https://doi.org/10.1007/BF00992696
18. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017)
19. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *Proc. Int. Conf. Mach. Learn.* (2018)
20. Li, Y.: Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2018)
21. Zhou, H., Jiang, K., Liu, X., Li, X., Leung, V.C.M.: Deep reinforcement learning for energy-efficient computation offloading in mobile-edge computing. *IEEE Internet Things J.* **9**(2), 1517–1530 (2022). https://doi.org/10.1109/JIOT.2021.3091142
22. Zhou, X., Yang, J., Li, Y., Li, S., Su, Z.: Deep reinforcement learning-based resource scheduling for energy optimization and load balancing in SDN-driven edge computing. *Comput. Commun.* **226–227**, 107925 (2024). https://doi.org/10.1016/j.comcom.2024.107925
23. Islam, F., Ahmed, I., Mihet-Popa, L.: Development and testing of an IoT platform with smart algorithms for building energy management systems. *Energy Build.* **344**, 115970 (2025). https://doi.org/10.1016/j.enbuild.2025.115970

Real-time Semantic Segmentation in Remote Sensing with PIDNet

Marko Petrov[1], Ivica Dimitrovski[1], Ema Pandilova[1], Vlatko Spasev[1], Ivan Kitanovski[1], and Pance Ribarski[1]

Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University in Skopje, North Macedonia
marko.petrov@finki.ukim.mk, ivica.dimitrovski@finki.ukim.mk,
ema.pandilova@finki.ukim.mk, vlatko.spasev@finki.ukim.mk,
ivan.kitanovski@finki.ukim.mk, pance.ribarski@finki.ukim.mk

Abstract. Real-time semantic segmentation is a critical capability for Unmanned Aerial Vehicle (UAV)-based applications that require fast and accurate scene understanding in dynamic environments. By capturing fine-grained visual details from low altitudes, UAVs are particularly well suited for applications in infrastructure monitoring, traffic analysis, and environmental assessment. This paper explores the application of PIDNet, a real-time semantic segmentation architecture, to UAV-based urban scene understanding using the UAVid dataset. UAVid poses distinct challenges such as class imbalance, the presence of partially visible or obscured objects, and the need to accurately detect small dynamic objects from side-view perspectives. We evaluate three PIDNet variants - PIDNet-S, PIDNet-M, and PIDNet-L, each representing a trade-off between speed and accuracy: from the lightweight, real-time PIDNet-S to the more accurate but heavier PIDNet-L. The experimental results show that all three PIDNet variants achieve high segmentation accuracy across diverse urban scenes, capturing both dominant structures and fine details, highlighting PIDNet's reliability for practical real-time semantic segmentation in UAV-based applications.

Keywords: semantic segmentation, remote sensing, real time, UAV, deep learning, PIDNet

1 Introduction

Unmanned Aerial Vehicles (UAVs) have become essential tools in modern remote sensing, enabling fine-grained visual data acquisition for tasks such as environmental monitoring, infrastructure inspection, urban planning, and disaster response. Operating at low altitudes and varying angles, UAVs offer high-resolution imagery with diverse perspectives that are well-suited for detailed scene analysis. Among the most critical capabilities for UAV-based applications is semantic segmentation[1][2], the task of assigning a class label to each pixel in an image - allowing systems to distinguish between roads, buildings, vegetation, vehicles, and other scene elements in a structured and interpretable manner.

However, semantic segmentation of UAV imagery presents several unique challenges[3]. These include significant variation in object scale, class imbalance, occlusions from side-view perspectives, and complex backgrounds. Moreover, real-time performance is often required for time-sensitive applications such as autonomous navigation or live environmental surveillance. Consequently, there is a growing demand for segmentation models that are not only accurate but also computationally efficient enough to operate under resource-constrained conditions, including deployment on edge devices or UAV onboard systems.

Recent progress in efficient deep networks has improved the feasibility of real-time semantic segmentation under limited compute budgets [4]. In this context, PIDNet proposes a parallel-branch design that explicitly separates complementary feature processing for detail preservation, context aggregation, and boundary refinement, yielding competitive speed–accuracy trade-offs in real-time segmentation benchmarks [5]. Although PIDNet is commonly reported on street-view urban datasets such as Cityscapes [6], its behavior on side-view, low-altitude UAV imagery remains less studied.

In this work, we evaluate three PIDNet variants (PIDNet-S, PIDNet-M, and PIDNet-L) for real-time semantic segmentation of UAV urban scenes on the UAVid dataset [7]. We compare accuracy and efficiency, reporting class-wise IoU and mean IoU, and we additionally assess the effect of test-time augmentation (TTA) on performance. The remainder of the paper is structured as follows: Section 2 summarizes related work on UAV semantic segmentation and efficient segmentation networks. Section 3 describes the UAVid dataset. Section 4 details the training and evaluation protocol. Section 5 presents the results and discussion, and Section 6 concludes the paper.

2 Related work

Semantic segmentation involves assigning a class label to each individual pixel in an image, enabling detailed scene understanding far beyond conventional image classification. While early approaches relied on traditional machine learning techniques, such as pixel-based and region-based classifiers, they required handcrafted features and struggled with variations in lighting, texture, and object scale. These limitations became especially pronounced in remote sensing imagery, which is characterized by high resolution, complex spatial patterns, and large data volumes.

The advent of deep learning transformed semantic segmentation through the use of Convolutional Neural Networks (CNNs)[8] and Fully Convolutional Networks (FCNs)[9]. Semantic segmentation of UAV imagery introduces additional challenges due to its highly variable resolutions, diverse object scales, side-view perspectives, and frequent class imbalance. UAV datasets such as UAVid [7] feature complex urban environments with small dynamic objects, cluttered backgrounds, and large structural variations within a single scene. Other publicly available UAV datasets include ISPRS Potsdam and Vaihingen[10], which focus on aerial photogrammetry and urban mapping; SkyScapes[11], aimed at real-time

aerial scene understanding; DroneDeploy[12] and Urban Drone Dataset [13], both providing annotated UAV images of urban areas for segmentation and object detection tasks. Compared to satellite imagery, UAV imagery introduces greater variation in viewing angles and object scales, and more pronounced occlusions due to the lower flight altitudes and side-views. Models must therefore generalize across scenes captured at varying altitudes, camera angles, lighting conditions, and environmental contexts, while maintaining precision in segmenting both dominant and rare classes.

The growing demand for real-time processing in UAV applications, such as disaster response, traffic monitoring, precision agriculture, and environmental surveillance has driven the development of lightweight and efficient semantic segmentation models. Real-time segmentation networks aim to deliver high-quality predictions under limited computational resources, often on embedded hardware aboard UAV platforms or in edge-computing scenarios. Early models such as ENet[14], ESPNet[15], and ICNet[16] introduced strategies for drastically reducing model complexity and inference time through aggressive downsampling, efficient convolutions, and lightweight decoder designs. Further improvements were proposed in BiSeNet[17], which decouples spatial and context pathways to maintain high-resolution features while accelerating computation. More recent works include transformer-based and hybrid models such as SegFormer[18], which demonstrated competitive performance with efficient architecture suitable for UAV imagery[19]. In addition, models like SwiftNet[20] and FCHarDNet [21] also focus on achieving an optimal trade-off between accuracy and speed for mobile and real-time segmentation tasks. While this paper focuses on semantic segmentation, it's worth noting that object detection remains a parallel line of research in UAV-based remote sensing. Detection-based approaches like YOLO have been successfully adapted to aerial imagery using transfer learning techniques, achieving reliable performance in identifying small-scale urban objects [22]. Such methods offer complementary capabilities, particularly when precise object localization is more critical than pixel-level delineation. However, for tasks requiring complete scene understanding, segmentation models provide a richer spatial representation. Collectively, these efforts illustrate the steady progress toward models that not only perform well across diverse aerial scenes but also meet the growing demands of real-time, resource-constrained deployment - goals that remain central to this study.

3 Dataset

4 Dataset

The UAVid dataset [7] is a benchmark collection developed for semantic segmentation of urban environments captured from Unmanned Aerial Vehicles (UAVs). Unlike nadir-view satellite imagery, UAVid consists of side-view recordings acquired at relatively low altitudes, which introduces strong perspective effects and large variations in object scale.

These properties make the dataset particularly suitable for evaluating segmentation models under realistic UAV operating conditions.
The dataset was originally released with 30 video sequences recorded in 4K resolution and later extended to a total of 42 sequences. Each sequence corresponds to a distinct urban location in order to promote scene diversity and reduce overfitting. From each sequence, 10 frames are densely annotated, resulting in a total of 420 labeled images. The images are provided at two spatial resolutions, 3840×2160 and 4096×2160 pixels, preserving fine structural details of the scene.
UAVid defines eight semantic classes that commonly appear in urban UAV imagery: building, road, static car, moving car, tree, low vegetation, human, and background clutter. Large static structures such as buildings and roads dominate the pixel distribution, while dynamic and small-scale objects, including cars and pedestrians, occupy a much smaller fraction of the image area. This imbalance reflects realistic urban environments and poses additional challenges for learning robust segmentation models.
The official dataset split consists of 200 images for training, 70 for validation, and 150 for testing. All sequences were captured under favorable lighting and weather conditions to ensure visual clarity and annotation consistency. Figure 1 illustrates the pixel-level class distribution across the dataset, while Figure 2 shows representative UAVid samples together with their corresponding ground truth annotations.

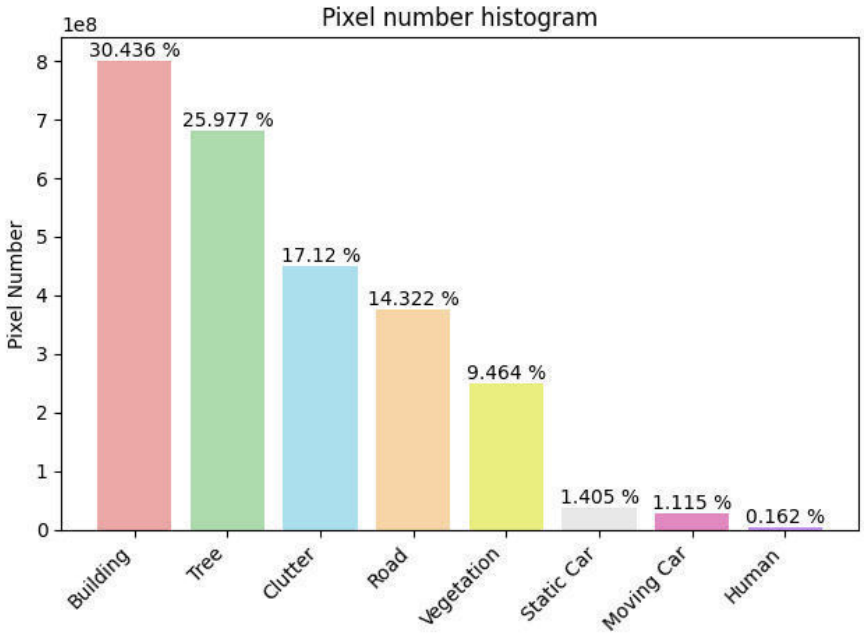

Fig. 1: Pixel-level class distribution in the UAVid dataset.

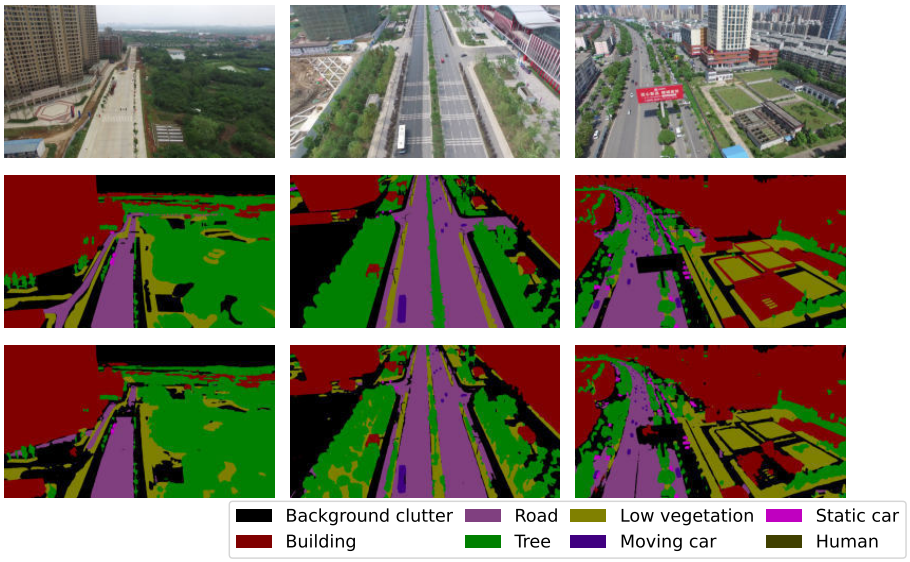

Fig. 2: Example images from the UAVid dataset, showing UAV-captured scenes (top row), their corresponding ground truth segmentation masks (middle row), and predicted outputs (bottom row).

5 Model architecture

PIDNet (Proportional-Integral-Derivative Network) [5] is a state-of-the-art real-time semantic segmentation architecture, inspired by classical PID control theory. It addresses a key limitation found in existing dual-branch networks used for semantic segmentation: the inadequate fusion of high-resolution detail features with low-resolution contextual information, which often leads to loss of important spatial details, particularly around object boundaries. This phenomenon, known as overshoot, can compromise segmentation accuracy in complex visual scenes.

To address this issue, PIDNet introduces a novel three-branch architecture, drawing conceptual parallels with full PID controllers in control systems.

As illustrated in Figure 3, the input image is first passed through several convolutional layers to extract low-level features and reduce its spatial dimensions. These features are then processed in parallel by the three branches of PIDNet:

- The **P-branch** focuses on capturing fine spatial details and small object structures. It preserves high-resolution features critical for segmenting narrow or intricate regions such as road lines or pedestrians.
- The **I-branch** performs aggressive downsampling (e.g., 1/16, 1/32, 1/64), aggregating global context necessary for understanding the

broader layout of the scene. This enables the model to disambiguate similar textures based on large-scale structure.
– The **D-branch** enhances edge awareness by isolating boundary-specific information. This branch supports the precise delineation of object boundaries, which is especially valuable in cluttered or overlapping regions.

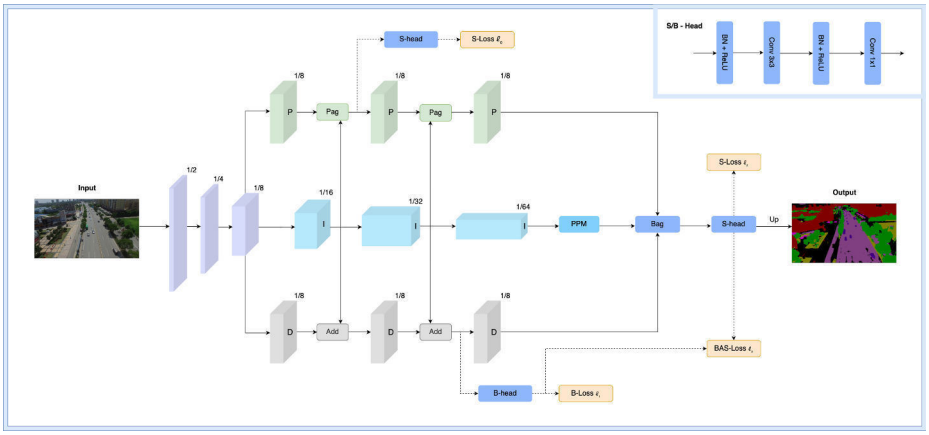

Fig. 3: Illustration of the PIDNet architecture showing the three branches: P (detail), I (context), and D (boundary), along with key modules and loss heads.

To combine information from the three branches effectively, PIDNet includes several critical modules. The **Pyramid Pooling Module (PPM)** aggregates contextual information at multiple spatial scales from the I-branch features, enabling the network to segment both large-scale structures and fine-grained details. This helps the model maintain consistent segmentation performance, even when object sizes vary within the image. The **Bag Fusion Module** serves to merge the high-resolution detail features from the P-branch with the low-resolution semantic features from the I-branch. It maintains a balanced integration, preventing the dominance of one feature type over the other and ensuring that detailed spatial information is not lost.
To further refine segmentation results, PIDNet utilizes multiple output heads and loss functions:
– The **Segmentation Head (S-head)** generates the final segmentation map and contributes to improving prediction accuracy during training.
– The **Boundary Head (B-head)** specializes in detecting object boundaries, supporting sharper transitions between adjacent classes.
– The network is trained using a combination of loss functions: **Segmentation Loss** (ℓ_0, ℓ_1), **Boundary Loss** (ℓ_2), and **Boundary Attention Supervision Loss** (ℓ_3). These losses are applied at different stages to enforce accurate region prediction and boundary localization.

6 Experimental setup

To enable effective training of the PIDNet models on the UAVid dataset, we adopted a simple pre-processing pipeline and training setup. Given the large native resolution of UAVid images, we partitioned them into fixed-size patches of 512x512 pixels using a stride of 256 pixels. This overlap ensured full scene coverage without losing contextual information at patch boundaries. The training set was augmented to 8000 patches and the validation set to 2800 patches after tiling. Images in the test split were preserved at their original resolution and not subjected to clipping. Our evaluation focused on the three PIDNet variants- PIDNet-S, PIDNet-M, and PIDNet-L, each initialized with pretrained weights on ImageNet to provide a strong starting point for training. Model performance was assessed on criteria such as number of trainable parameters, inference speed (FPS), and per-sample latency.

To improve model generalization, we applied a diverse set of data augmentations during training. Basic operations included random horizontal flipping and per-channel brightness and contrast adjustments. To further introduce variability, we incorporated complex spatial and color transformations such as Contrast-Limited Adaptive Histogram Equalization (CLAHE), grid distortion, and optical distortion. Finally, all images were normalized using channel-wise mean and standard deviation computed from the UAVid training set. No augmentations other than normalization were applied at validation or test time. Models were trained for up to 100 epochs using stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of $1e$ - 4. An initial learning rate of $1e$ - 3 was gradually reduced using a polynomial decay schedule to $1e$ - 7 by the final epoch. To balance region-level and pixelwise segmentation accuracy, we adopted a hybrid loss combining multiclass Dice loss and a pixelwise Cross-Entropy loss.

At test time, we used a sliding-window inference strategy with a clip size of 1024 pixels and a stride of 896 pixels to process high-resolution UAVid images without losing spatial detail. To improve prediction accuracy, we applied test-time augmentation (TTA) by horizontally flipping the input images and averaging the predictions from the original and flipped versions. Quantitative evaluation of all variants was conducted by reporting per-class Intersection-over-Union (IoU) and overall mean IoU. All experiments were conducted on a workstation running Debian Bookworm with the 6.1.0-31-amd64 Linux kernel. The system was equipped with an NVIDIA GeForce RTX 3090 Founders Edition GPU (24 GB VRAM), an AMD Ryzen 9 7950X CPU, and 64 GB of DDR5 RAM. The machine used NVIDIA Driver Version 535.216.01 and CUDA Version 12.2. The implementation was based on PyTorch Lightning and executed using a single GPU.

7 Results and discussion

The results from our experimental runs are presented in Table 1. The trained PIDNet model weights used in these experiments are publicly

available at our GitHub repository[1]. Among the baseline models, **PIDNet-M** achieves the highest overall mean IoU of **0.6521**, outperforming both PIDNet-S (0.6439) and PIDNet-L (0.6378). It consistently performs best across several key classes, including *Road* (0.7795) and *Tree* (0.7860), demonstrating strong capabilities in modeling elongated and vegetative structures. PIDNet-M also achieves the best score for the *Building* class with an IoU of **0.8617**, closely followed by PIDNet-L (0.8556), reflecting their effectiveness in capturing static vertical structures in urban scenes. The *Static Car* class appears to be consistently challenging, with IoUs ranging from 0.4916 (PIDNet-L) to 0.5528 (PIDNet-M), likely due to frequent visual confusion with *Moving Car* and occlusions common in UAV side-view imagery. The *Human* class yields the lowest IoUs across all variants, ranging from 0.2744 to 0.2803. These low scores are primarily caused by the small size, low frequency, and poor visibility in high-altitude aerial perspectives. As shown in Table 1, applying test-time augmentation (TTA) improves performance across all models. PIDNet-M benefits most, reaching an mIoU of **0.6659**, a gain of +1.38 points over its base result. Notable class-level improvements are observed in *Building* (0.8707), *Road* (0.7926), and *Tree* (0.7972). PIDNet-S also shows a meaningful improvement from 0.6439 to 0.6601, confirming its ability to leverage augmentation for better generalization. PIDNet-L improves as well, reaching 0.6513 mIoU, though its gains are more modest. Results, shown in Table 2, include latency in milliseconds, frames per second (FPS), and parameter counts. PIDNet-S demonstrated the best real-time performance with a latency of just 8.23 ms and throughput over 120 FPS, with a lightweight architecture of just 7.6 million parameters. In contrast, PIDNet-M and PIDNet-L had higher inference times of 13.93 ms and 18.21 ms, respectively, while maintaining solid accuracy.

While our benchmarks were conducted on an NVIDIA RTX 3090 to provide a controlled performance comparison across PIDNet variants, real-world UAV deployments typically rely on resource-constrained embedded devices such as NVIDIA Jetson Xavier/Orin or Qualcomm Snapdragon platforms. On such hardware, inference speeds are expected to be significantly lower due to limited GPU cores, lower memory bandwidth, and power consumption constraints. Nevertheless, the compact design of PIDNet-S (7.6M parameters) makes it a strong candidate for deployment in these scenarios, where real-time throughput above 20–30 FPS is often sufficient for UAV navigation and monitoring tasks.

Beyond intra-family comparisons, it is important to position PIDNet relative to other lightweight segmentation baselines on UAVid. EMNet[23], for example, achieves ∼71.5% mIoU, while UNetFormer[24] reports ∼67.8% mIoU with exceptionally high throughput (up to ∼322 FPS on 512×512 input). Similarly, STDC-CT [25] achieves ∼68.4% mIoU and has been validated on embedded hardware, running at ∼58 ms per frame on a Jetson TX2 device. Against this backdrop, PIDNet-M (66.6% mIoU with TTA) offers competitive accuracy while maintaining real-time capability, and PIDNet-S provides unmatched speed (121 FPS) with only a modest drop in segmentation quality. These results highlight that PIDNet

[1] `https://github.com/markopetrov1/real-time-segmentation-pidnet-paper`

remains well-aligned with the broader family of efficient semantic segmentation models, striking a favorable balance of speed and accuracy for UAV-based applications.

While our benchmarks were conducted on an NVIDIA RTX 3090 to provide a controlled performance comparison across PIDNet variants, real-world UAV deployments typically rely on resource-constrained embedded devices such as NVIDIA Jetson Xavier/Orin or Qualcomm Snapdragon platforms. On such hardware, inference speeds are expected to be significantly lower due to limited GPU cores, lower memory bandwidth, and power consumption constraints. Nevertheless, the compact design of PIDNet-S (7.6M parameters) makes it a strong candidate for deployment in these scenarios, where real-time throughput above 20–30 FPS is often sufficient for UAV navigation and monitoring tasks.

Table 1: Comparative performance of PIDNet variants on the UAVid dataset.

Model	Background	Building	Road	Tree	Low veg.	Moving car	Static car	Human	mIoU
PIDNet-S	0.628	0.852	0.770	0.776	0.601	0.696	0.552	0.274	0.643
PIDNet-M	0.650	0.861	0.779	0.786	0.610	0.695	0.552	0.280	0.652
PIDNet-L	0.638	0.855	0.772	0.783	0.610	0.670	0.491	0.279	0.637
PIDNet-S(tta)	0.644	0.861	0.786	0.785	0.617	0.716	**0.585**	0.284	0.660
PIDNet-M(tta)	**0.665**	**0.870**	**0.792**	**0.797**	**0.628**	**0.721**	0.566	0.285	**0.665**
PIDNet-L(tta)	0.653	0.864	0.784	0.794	0.626	0.693	0.503	**0.289**	0.651

Table 2: Latency and efficiency summary for each PIDNet variant (NVIDIA RTX 3090, 24GB VRAM, CUDA 12.2).

Model	**Parameters**	**FPS**	**Latency (ms)**
PIDNet-S	7.6M	121.58	8.23
PIDNet-M	34.4M	71.81	13.93
PIDNet-L	36.9M	54.90	18.21

The confusion matrix of PIDNet-M (Figure 4) reveals strong diagonal dominance for large, well-represented categories such as *Building* (93%) and *Road* (84%), confirming high classification confidence. Some confusion is observed between visually similar or spatially co-occurring classes. *Background clutter* is occasionally confused with *Tree* and *Low vegetation*, which share texture and color properties. Likewise, *Moving Car* frequently overlaps with *Static Car* and *Road*, a plausible result given spatial proximity in urban street views. *Human* is misclassified primarily

as *Low vegetation* or *Background*, which aligns with its low IoU score and reflects the difficulty of detecting such small instances.

Fig. 4: Confusion matrix for PIDNet-M predictions on the UAVid test set.

8 Conclusion

This work explored the segmentation performance of three PIDNet variants, PIDNet-S, PIDNet-M, and PIDNet-L, on UAV-based urban scenes using the UAVid dataset. Each model was tested with and without test-time augmentation (TTA) to evaluate both accuracy and efficiency. PIDNet-M delivered the highest mean IoU (0.6659 with TTA), outperforming the other two versions in most classes. It particularly excelled in segmenting large structures such as roads and buildings. While PIDNet-L showed competitive results, it came at a slightly higher computational cost. As expected, the larger the model, the slower the inference and the greater the memory usage. Interestingly, PIDNet-M offered a favorable balance - being more efficient than PIDNet-L while also achieving superior segmentation results. In contrast, PIDNet-S, though less accurate, demonstrated the fastest inference speeds and lowest resource consumption, making it suitable for real-time applications under constrained environments.

Despite their strong performance on major classes, all three models struggled with fine-grained segmentation of the *Human* class, which remained under 0.29 IoU even with TTA. This is likely due to the small size and sparse occurrence of humans in aerial imagery, as well as frequent occlusions and background confusion.
These findings confirm that PIDNet-S is well-suited for time-critical UAV deployments, whereas PIDNet-M provides the best trade-off when higher segmentation quality is required. Future work may explore deployment on embedded edge devices such as Jetson AGX Orin or Xavier NX, where trade-offs between latency, power consumption, and segmentation accuracy become critical. Evaluating PIDNet variants under such constraints will provide deeper insights into their suitability for real-world UAV integration, especially in autonomous navigation and time-sensitive monitoring applications. Additional directions include integration with downstream tasks such as tracking or planning, and training on additional UAV-specific datasets to enhance generalization and robustness.

Acknowledgement

The authors thank the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje for the provided finacial support under the SatTime ("Analysis of satellite image time-series") project.

References

1. Lucas Prado Osco, José Marcato Junior, Ana Paula Marques Ramos, Lúcio André de Castro Jorge, Sarah Narges Fatholahi, Jonathan de Andrade Silva, Edson Takashi Matsubara, Hemerson Pistori, Wesley Nunes Gonçalves, and Jonathan Li. A review on deep learning in uav remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102:102456, 2021.
2. Vlatko Spasev, Ivica Dimitrovski, Ivan Kitanovski, and Ivan Chorbev. Semantic segmentation of remote sensing images: Definition, methods, datasets and applications. In *International Conference on ICT Innovations*, pages 127–140. Springer, 2023.
3. Adrian Carrio, Carlos Sampedro, Alejandro Rodriguez-Ramos, and Pascual Campoy. A review of deep learning methods and applications for unmanned aerial vehicles. *Journal of Sensors*, 2017(1):3296874, 2017.
4. Fanghui Chen, Shouliang Li, Jiale Han, Fengyuan Ren, and Zhen Yang. Review of lightweight deep convolutional neural networks. *Archives of Computational Methods in Engineering*, 31(4):1915–1937, 2024.
5. Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19529–19539, 2023.

6. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
7. Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020.
8. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
9. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
10. ISPRS Commission II/WG II. 2d semantic labeling contest (isprs). `https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam/`, 2012. Accessed: 2024-06-29.
11. Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019.
12. J. M. Thornton. DroneDeploy. `https://github.com/dronedeploy`, 2013. Accessed September 2, 2022.
13. David Zhenyue Liu, Zechen Liu, Yuchen Zhuang, Song Bai, and Xiang Bai. Urban drone dataset: Object detection and tracking for uavs in urban scenes. *arXiv preprint arXiv:2009.08479*, 2020.
14. Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.
15. Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.
16. Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.
17. Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.
18. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.
19. Vlatko Spasev, Ivica Dimitrovski, Ivan Chorbev, and Ivan Kitanovski. Semantic segmentation of unmanned aerial vehicle remote

sensing images using segformer. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 108–122. Springer, 2024.

20. Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1296–1305, 2021.
21. Quanlong Chao, Zhaoqiang Li, Qing Ma, and Hao He. Fchardnet: A new backbone for efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 168–169, 2021.
22. Ema Pandilova, Marko Petrov, Vlatko Spasev, Ivica Dimitrovski, and Ivan Kitanovski. Transfer learning with yolo for object detection in remote sensing. In *International Conference on ICT Innovations*, pages 121–135. Springer, 2024.
23. Xiaolong Li, Yuyin Li, Jinquan Ai, Zhaohan Shu, Jing Xia, and Yuanping Xia. Semantic segmentation of uav remote sensing images based on edge feature fusing and multi-level upsampling integrated with deeplabv3+. *Plos one*, 18(1):e0279097, 2023.
24. Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.
25. Bo Jiang, Zhonghui Chen, Jintao Tan, Ruokun Qu, Chenglong Li, and Yandong Li. A real-time semantic segmentation method based on stdc-ct for recognizing uav emergency landing zones. *Sensors*, 23(14):6514, 2023.

Stock Trading Recommendations Using Deep Q-learning and NLP

Kostandina Veljanovska[1], Simeon Nalovski[1], Blagoj Ristevski[1] and Snezana Savoska[1]

[1] University "St. Kliment Ohridski" – Bitola, Faculty of Information and Communication Technologies, Partizanska bb, 7000 Bitola, RN Macedonia
kostandina.veljanovska@uklo.edu.mk, nalovski.simeon@uklo.edu.mk, blagoj.ristevski@uklo.edu.mk, snezana.savoska@uklo.edu.mk

Abstract. This paper explores the application of deep Q-learning and natural language processing (NLP) to the stock market trading process. Deep Q-learning, a method that is part of reinforcement learning, is used to create trading strategies based on historical data and technical indicators, aiming to optimize long-term returns through intelligent decision-making. Natural language processing techniques are applied to analyze financial news and social media content to extract sentiment and relevant stock market trends. By investigating the two approaches separately, the research aims to evaluate their effectiveness in predicting stock market trends and making informed decisions, offering a basis for future integration of the two approaches into a single financial system.

Keywords: deep Q-learning, deep reinforcement learning, natural language processing, stock trading recommendations.

1 Introduction

In recent years, the application of artificial intelligence in financial markets has gained significant momentum, with the application of deep Q-learning and natural language processing representing a new approach to building intelligent trading systems. There are scientific efforts in this direction, such as integrating ANN, LSTM, and natural language processing (NLP) techniques with the deep Q network (DQN) in order to craft a novel architecture tailored specifically for stock market prediction [1, 2], or natural language processing (NLP) used to explore possibilities to advance the traditional approaches to stock price prediction [3], or applying an end-to-end double DQN model to financial time series analysis problem [4], or LSTM, CNN, and SVM utilized in predicting stock prices, volatility, and trends [5], or Deep Q-learning model with augmented sentiment analysis and stock trend labelling [6].

Many of the models proposed in the literature overcome the limitations of supervised learning approaches, Deep Reinforcement Learning (DRL) algorithms can scale to previously uncontrollable problems, i.e. DRL model are used to generate profitable trades in the stock market, effectively. By formulating the trading problem as a Partially Observed Markov Decision Process (POMDP) model, considering the constraints imposed

by the stock market, such as liquidity and transaction costs, scientists solved the formulated POMDP problem using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [7]. The theme of stock market forecast is interesting even for scientists that are developing reinforcement learning (RL) techniques typically in another areas [8, 9]. The power of RL can be utilized in different stochastic environments [10] and in this research deep RL excels over stochastic environment of stock trading proposals.

The paper covers the application of both segments of artificial intelligence to stock trading on the stock exchange, highlighting the way it works, data sources, details of their implementation and the corresponding challenges. Deep Q-learning allows an agent to learn optimal trading strategies through a process of trial and error in a simulated stock exchange environment. By interacting with historical data on the share price of a given company, the agent learns and later makes decisions, i.e. whether to buy or sell a share at a given point in time. In doing so, the agent aims to maximize the cumulative reward over time [11]. The model updates its policies based on feedback from the environment, further refining its strategy through a continuous process of exploitation and exploration. The application, named FinSmartRL, is implemented in the Python programming language, leveraging the flexibility of libraries such as TensorFlow and Keras. For training and testing, historical price data for 501 out of 503 companies in the US S&P 500 index fund were used, which were downloaded using the Yahoo Finance API and cover the period from 01.01.2019 to 13.12.2024.

On the other hand, NLP focuses on extracting meaningful information from unstructured text data. NLP techniques such as tokenization, sentiment analysis, and text generation were applied to capture and explain sentiment and highlight relevant events that the user would act on. The text data was retrieved from two sources. For real-time data access, NewsAPI was used, while for fine-tuning, a file of already rated news stories was retrieved from Kaggle. Also, the Python programming language and its libraries Transformers, Torch, Scikit-Learn, etc. were used for this project. Despite the potential of these two approaches, there were also certain challenges during development. One of the biggest challenges, especially in deep Q-learning, was overfitting. Namely, the agent sometimes memorized patterns specific to the training data set and performed poorly on unknown data (the testing data set). Meanwhile, the main challenges in NLP were irrelevant or noisy data, inconsistent headline lengths, and ambiguous sentiment signals. Ensuring data relevance and quality was a major challenge throughout the research. In the following, we will follow the project building process, the results of parameter research, optimal configurations, and a demonstration of how they work.

2 Deep Q-learning

Deep Q-learning is an extension of Q-learning, an algorithm within the framework of reinforcement learning, which is used to make decisions on complex problems where an agent learns to act optimally in its environment by optimizing the cumulative reward. In classical Q-learning, pairs of states and actions are stored in Q-tables [9, 10, 11].

However, in situations where the state space becomes large and/or continuous (such as the stock market), maintaining the Q-table becomes infeasible. To this end, deep Q-learning approximates (estimates) the value of the Q-function using a deep neural network [12]. As input, deep Q-learning takes the state of the environment and produces an output in the form of Q-values for each possible action. During training, the agent interacts with the environment in discrete time steps. At each step, the agent observes the state, takes an action based on its strategy, and transitions to a new state. The agent's state update is performed according to the Bellman equation. The stabilization of the training process is performed through the repetition of experiences, during which previous transitions (current state, action, reward and next state) are remembered. In doing so, random sampling of transitions is performed in order to break temporal correlations.

In the context of stock trading, we will present the stock market as a simulated environment, the state space will be represented by the historical price and technical indicators for each stock and the actions that the agent can perform are buying or selling a stock. The reward function is designed to reflect the performance, i.e. changes in the stock portfolio.

In this way, deep Q-learning offers a powerful framework for adapting trading policies from raw data, without having to explicitly model stock market dynamics. However, the effectiveness of this approach depends largely on the correct setting of the parameters, the reward and dealing with oversaturation, delayed reward and the constant change of the stock market.

2.1 Deep Q-learning Model Design

The data used for the FinSmartRL application are gained using the Yahoo Finance API. For this purpose, we previously defined the function used to create a list of symbols (in the files we encounter the English term ticker, taken from Wikipedia) of the stocks of the S&P500 index fund. Furthermore, we use the list of symbols in the call to the download method of the Yahoo Finance API. When calling the method, we use only a part of all parameters (5 out of a possible 20). For this purpose, a function has been created, which returns a data frame with the data from the method, with its parameters being a list of stocks, start date, end date, a grouping condition and an interval (1 day, indicating that data is downloaded for each day). At the end of this file, we perform preprocessing. With each new execution, the data is enriched with new stock price data for the period from the previous execution to the current execution.

The second part of the data preparation involves splitting the central file into smaller files, which we later call when training and testing the model. First, we load the central file into a data frame and extract the unique values (501 in total) from its Ticker column. Next, we go through each symbol in the list, take the data and discard the column in which the symbol is located. The obtained data (only numeric values) were divided into a training set and a testing set, with time order preserved, maintaining the ratio 80-20. This way allows easier implementation of the training and testing phases and less time spent extracting data. But on the other side, with this technique of splitting comes a risk of inflating the performance estimations due to temporal dependencies. Knowing this,

implementing a rolling-window evaluation could have given a more rigorous test of generalization and reduce the possibility of overfitting.

In the model, agent has been created, with the parameters for memory, size of states, size of actions, gamma (the rate of inclusion of the future reward), epsilon (the initial exploration rate for the epsilon-greedy policy), epsilon min (to prevent total exploration), epsilon decay (to promote exploration) and learning rate (to update the weights).

For model creation, the Keras Layer API is used (parts of Keras API 3 and Keras API 2, were used to create layers, apply layer activation, initialize layer weights, calculate loss, and optimize the model). For optimization, Adam optimizer was used. It uses the learning rate obtained with the decay according to Inverse Time Decay.

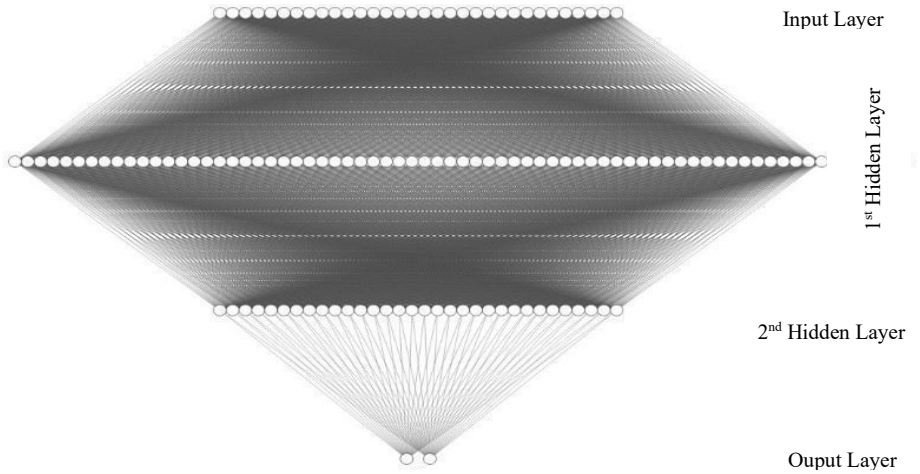

Fig. 1. Neural Network Architecture in the Model

From the visualization of the NN architecture (Fig. 1): NN has 4 layers (all neurons from the previous layer are connected to all neurons from the next layer). The input layer has 32 neurons. The 1st hidden layer (64 neurons) expands and enriches the feature space, allowing the model to "capture" the complex interactions and dependencies between the various indicators for a given stock. The 2ndhidden layer (32 neurons) compresses the learned representation from the first hidden layer, through filtering the extracted patterns and filtering noise. An output layer of two neurons has task to choose whether to buy or sell a stock at a given time point (day), by applying SoftMax activation to estimate probabilities (eq. 1).

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \quad (1)$$

where $\sigma(z_i)$ is the probability that action i is chosen, z_i is the output of the NN [13] for action i.

Within the replay function, we calculate the function Q(s',a) to obtain the target, predict the value of the function Q(s,a) and possibly reduce epsilon. Within the replay function, we calculate the Q value (instead of storing it in a Q table) using the Bellman equation, which in mathematical form is represented in eq. 2 [9, 10, 11]

$$Q(s,a)=r+\gamma a' \max Q(s',a') \quad (2)$$

This calculation is stored in the target variable, i.e. the calculated Q value. The target value contains the sum of the current reward and the product of gamma the best possible future state. After this calculation, the current Q value is updated, and the model is trained with the newly obtained Q value.

2.2 Research Experiments: FinSmartRL

The results from the application of FinSmartRL are discussed in two separate parts of this chapter. The first part is oriented towards exploring the parameters of the model and the parameters of the agents in order to obtain a model configuration that could simultaneously have a solid behavior during training and continue the same in testing, while showing an appropriate degree of generalization. The second part is oriented towards invoking the training and testing processes, followed by a textual display of results and a visual display using the three visualizations (the movement of the cumulative profit across the episodes during training and during training, and the actions taken during testing).

Research on Model Parameters and Agent Parameters. In order to investigate the possible combinations of model and agent parameters and their impact in the training and testing phases, 26 different combinations were tested, with changes in epsilon, minimum epsilon, epsilon decay rate, application or non-application of layer weight initialization, and application or non-application of learning rate decay techniques.

Table 1. Research Parameters and Results, organized into smaller group tables

Model	Gama	ε	ε_{min}	ε decay rate	Learning rate	Weights init	Avg profit (train)	Avg Reward (train)	Total profit (test)	Total Reward (test)	Imprvmnt with weight init(%)
model_NVDA25_01_21	0,95	1	0.1	0.995	0.0005	/	2.555635585524457 5	12.7033858847618 13	1.382356	104.930027	Referent model
model_ABT25_02_15	0,90	1	0.15	0.990	0.0001	/	1.259605858667594 7	94.2682025909423 5	1.073852	19.610039	/
model_ABNB25_02_15	0,99	1	0,1	0.996	0.001	/	1.319075506503 15	50.5167060852050 6	1.219876	-9.510056	/
model_ADM25_02_15	0,98	0.8	0,02	0.997	0,0001	/	1.366368394480042 4	45.9367244148254 06	1.680438	-13.720047	/
model_AFL25_02_16	0.97	0,85	0.05	0.992	0,005	/	1.276906944165138 2	-1.22380308151244 74	1.292364	22.490005	/
model_BA25_02_16	0.97	0,85	0.05	0.992	0,5	/	1.678535219762483 4	74.2060990905763	1.217801	38.389999	/
model_BEN25_02_16	0,98	0.8	0,02	0.997	0,6	/	1.406977661130264	2.17280010223388 3	2.300001	1.117433	/

model_BAX25_02_16	0,90	1	0.15	0.990	0.7	/	0.96148729216628 67	0.25720947265624 24	0.847144	-5.139999	/
model_BAC25_02_16	0,99	1	0,1	0.996	0.8	/	1.048398619204805	-1.63010095596313 05	0.646814	-15.68999	/
model_BDX25_02_16	0,95	1	0.1	0.995	0.9	/	1.076204216150238 8	4.90369644165 0415	1.042161 7	7.710006 1	/

A. Group of models without learning rate decay and without weight initialization

Model	Gama	ε	ε_{min}	ε decay rate	Learning rate	Weights init	Avg profit (train)	Avg Reward (train)	Total profit (test)	Total Reward (test)	Imprvmnt with weight init(%)
model_BG25_02_17	0,95	1	0.1	0.995	0.7	HeUniform + GlorotUniform	1.040985092181663 2	-4.642800083160385 5	1.281669	24.150002	-55,4%
model_BR25_02_16	0,99	1	0,1	0.996	0.8	HeUniform + GlorotUniform	1.157782776176905	4.606510086059552 5	1.253889	50.830002	- 48,9 %
model_CLX25_02_17	0,98	0.8	0,02	0.997	0,6	HeNormal + GlorotNormal	1.096078411383876 8	-4.305392227172873	1.221126	25.729996	-53,5%
model_CI25_02_17	0.97	0,85	0.05	0.992	0,5	HeNormal + GlorotNormal	0.974777794658152 1	-12.3490112304687 63	0.984234	-0.700012	-72,2%

B. Group of models without learning rate decay and with weight initialization

Model	Gama	ε	ε_{min}	ε decay rate	Learning rate	Weights init	Avg profit (train)	Avg Reward (train)	Total profit (test)	Total Reward (test)	Imprvmnt with weight init(%)
model_DD25_02_19	0,95	0.6	0.05	0.99	0.7 (Cosine Decay, decay steps 1800, alpha 0.2	/	0.841161465583307 8	11.684216403961182	1.120624	9.619995	/
model_DAY25_02_19	0,95	0,6	0,05	0,99	0.6(Polynomial Decay, steps 1500, end l.rate 0,0001 power 2.0	/	0.873221959654250 2	-7.629101371765135 5	0.871323	-7.879997	/
model_DAL25_02_19	0,99	1,0	0,05	0,996	0,5 (Exponential decay, 5000 steps, 0,99 decay rate)	/	1.461854920188170 5	0.498392372131341 7	1.639764	25.459999	/
model_CZR25_02_19	0,97	1,0	0,01	0,99	0,8 (Inverse time decay, 2500 steps, 0,005 decay rate)	/	2.156539127831802 5	-0.256247425079363 6	1.233180	8.900002	/

C. Group of models with learning rate decay and without weight initialization

Model	Gama	ε	ε_{min}	ε decay rate	Learning rate	Weights init	Avg profit (train)	Avg Reward (train)	Total profit (test)	Total Reward (test)	Imprvmnt with weight init(%)
model_CMI25_20_17	0.95	0,85	0.05	0.992	0,5 (Exponential decay, decay rate 0.96)	HeUniform + GlorotUniform	1.232576763386525 2	3.979206848144493 3	1.634086	142.010010	-16,7%
model_CME25_02_17	0,98	0.8	0,02	0.997	0,6 (Polynomial decay, end 0.0001,100	HeUniform + GlorotUniform	1.109541587971016 9	6.082489624023424	1.119823	23.770004	-61,4%

					0 steps, power 1						
model_C 25_02_18	0,99	1	0,1	0.996	0.8 (Inverse-TimeDecay, decay steps 1, decay rate 0.5	HeUni-form + Gloro-tUniform	1.237280 37397159 94	1.181797 25646972 88	0.629066	- 25.65000 2	-88,6%
model_D 25_02_18	0,95	1	0.1	0.995	0.7 (Cosine De-cay, decay steps 1000, alpha 0.1	HeUni-form + Gloro-tUniform	0.796056 68608877 17	- 14.89189 70108032 23	1.175945	7.280003	65,5%
model_C AT25_02 _19	0,95	0,9	0,05	0,995	0.5 (Expo-nential de-cay, 1000 steps, decay rate 0,96)	HeUni-form + Gloro-tUniform	1.248331 78097702 78	3.169808 27331541 23	1.505716	127.4400 02	-24,0%
model_A CN25_02 _19	0,97	0,85	0,05	0,992	0,6 (Poly-nomial de-cay, steps 1200, power 1, end lear. Rate 0,0001	HeUni-form + Gloro-tUniform	0.927830 25642469 94	- 32.18062 02697754 5	0.912568	- 30.85000 6	-83,7%
model_A CGl25_0 2_19	0,95	0,8	0,05	0,990	0,7 (Cosine decay, al-pha 0,05, steps 1500)	HeUni-form + Gloro-tUniform	1.718653 38595355 38	11.70680 44281005 86	1,096264	9,820000	-65,6%
model_F RT25_02 _19	0,94	0,74	0,05	0,985	0,8 (Inverse time decay, steps 1000, rate 0,01)	HeUni-form + Gloro-tUniform	0,793025 48990878 52	- 27,96940 47546386 67	1.238882	21.70999 9	-55,4%

D. Group of models with learning rate decay and with weight initialization

As shown in Table 1 (group tables), different parameters and techniques (shown in the 2nd to the 7th column) give different results (shown in the 8th to the 11th column) and significantly affect them. It is also worth noting that the Adam optimizer and 100 training episodes were used throughout the research. The model named model_NVDA25_01_21 was used as a reference model, and the comparison below will be with models in which we have applied techniques for initializing the weight by layers, taking into account the percentage change in performance (as a whole, taking into account the average profit per episode during training, the average reward per episode during training, the total profit during testing and the total reward during testing). The techniques used for decreasing the learning rate are Exponential Decay, Cosine Decay, Polynomial Decay and Inverse Time Decay. Each learning rate decay technique has its own advantages, from the point of view of application to our agent. Thus, Inverse Time Decay allows for gradual decay (the agent learns more slowly as episodes pass) and prevents overconfidence in the beginning of learning. Polynomial Decay allows for greater control over the decay, to obtain a more elegant and adaptive decay of the learning rate. Cosine Decay is useful when we are working with a longer training period and a fast-initial path would be useful. Whereas, Exponential Decay is useful when we have a rapidly converging environment, so stability of the agent's policies is necessary. From the table itself, we conclude that representatives of the He and Glorot initializations (in the form of Uniform and Normal) are used in the layers of the NN. He initialization is useful for the hidden layers of the NN, since they use ReLu activation. The ReLu activation itself is called with the parameter activation when defining the layer. The ReLu activation gives a result of 0 for negative inputs, which halves the number of neurons. This is where the He initialization comes into play, which scales the variance of the

weights, keeping its value at a healthy level, preventing the disappearance or explosion of the nuance (a measure of the change in all weights relative to the change in errors). On the other hand, the output layer uses a SoftMax activation function, which often gives too high or too low a variance, which can ultimately lead to an overly confident or underachieving probability distribution. For this purpose, Glorot uses an equation, in which the number of input and output units is taken into consideration, allowing for a stable shade. The choice of the versions of He and Glorot initializations (uniform or normal) comes down to the preferences of the developer or engineer themselves or the preferences for choosing the architecture of the NN itself. From the results obtained, we can draw the following conclusions: 1. The most stable is model_DAY25_02_19, because of the small differences between the average training results and the total testing results. 2. The most unstable (having the highest total reward in testing) is model_CMI25_20_17, due to a huge difference between the average reward during training and the total reward at the end of testing. 3. The highest average reward during training has model_ABT25_02_15. 4. The most profitable model during training/testing is model_NVDA25_01_21/model_BEN25_02_16. 5. The models with initialization showed overall worse performance compared to the reference model (the drop ranged between 16.7% and 88.6%). 6. The worse performance of the models with weight initialization is due to the large drop in performance during training. 7. The models model_CMI25_20_17 and model_CAT25_02_19 showed the smallest overall drop (-16.7%, respectively -24.0%). 8. The above models showed better performance during testing, model_CMI25_20_17 showed better total profit by 18.2% and better total reward by 35.4% during testing. On the other hand, model_CAT25_02_19 showed better total profit by 8.8% and better total reward by 21.6%. 9. The better performance of these models during testing indicates better generalization compared to the reference model, while sacrificing the quality of the training performance. 10. Despite the great diversity of the research, not all possible combinations of parameters and techniques have been exhausted.

Discussion of the Results - FinSmartRL Demonstration. The demonstration of the FinSmartRL application is done through a statically entered shortcut of the action, in our case - the company Nvidia (its shortcut is NVDA, with the agent parameters from Table 1). For the experiments setup, single stock evaluation has been done. This approach would have a natural extension, allowing to test the model's capabilities on stock data from multiple sectors and enabling stronger statistical claims.

After the implementation of the model (via a shortcut variable and a call to the function for creating "labels") the training and testing were conducted for the stock of the company Nvidia. The results show that the model delivered an average profit of 2.5556355855244575 across the 100 episodes and an average reward of 12.703385884761813 during the training. Fig. 2 shows the movement of the cumulative profit across episodes. The cumulative profit reached its peak around the 50th episode, reaching a value of 59.82624602317812.

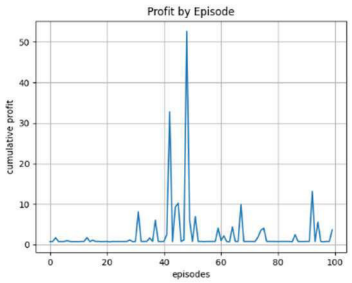

Fig. 2 Cumulative Profit across Episodes

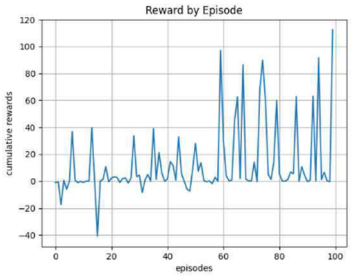

Fig. 3 Cumulative Reward

Fig. 3 shows the movement of the cumulative reward across episodes. The cumulative reward reached its peak in the 100th episode when it amounted to 112.57073807716378.

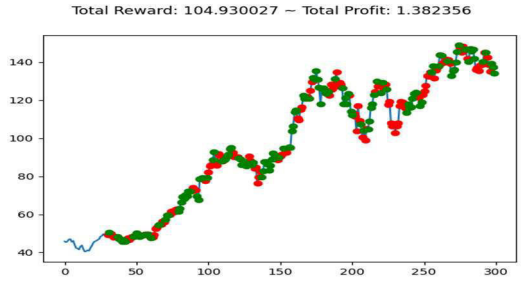

Fig. 4 Graph of actions taken

From the graph of actions taken (Fig.4), we conclude at which moments the model decided to buy (green dots) or sell (red dots) a stock. Taking into account the total profit and reward, the demonstration showed solid results, if we follow the model's actions (total profit of 38.2356% and high total reward) over the testing period of just over 14 months. The system has simulated a day trading strategy focused on short-horizon decisions instead of long-term investing. Given this context, comparing with a buy-and-hold baseline would have meant nothing, since the objective, risk exposure and evaluation horizons differ fundamentally. Our primary objective is to achieve adaptability to daily fluctuations and optimize sequential decision making. We have done a comparison of data for different companies. The behavior during testing with Nvidia's test data (Fig. 4), and it could be stated that, according to Nvidia's relationship with other companies, we have overfitting to the dynamics of data movement, as opposed to generating a universally applicable strategy. This alone indicates that the model needs to be trained with data from companies whose trends are not in Nvidia's sector.

3 Natural Language Processing (NLP)

NLP in the context of stock trading plays a key role, considering that it is used to convert vast amounts of unstructured text data (such as news, earnings reports, analyst comments, social media posts) into structured information for better decision-making.

One of the most common applications of NLP in the world of finance is sentiment analysis [14], which aims to classify text as positive, neutral, or negative based on its content. Methods for analyzing stock sentiment are numerous, starting with Naive Bayes classifiers and support vector machines, and ending with advanced transformer-based models (such as BERT and RoBertA). Thus, a news story containing data on large earnings in the past quarter can signal positive sentiment, which would further mean a signal to buy the stock for an automated trading system. These sentiment assessments should be expressed in numerical form and incorporated into predictive models. In order to adapt ready-made language models to financial data, pseudo fine-tune and fine-tune approaches are often used. Pseudo-fine-tuning involves "feeding" a ready-made language model with domain-specific data in order to adjust its vocabulary and embedding space, without changing the model weights. This technique is useful for certain tasks, for which we do not need intensive re-training of the model. On the other hand, fine-tuning involves continuing the training process of the ready-made model, using a data set that is both domain-specific and already analyzed sentiment (preferably by a human). In doing so, the model weights are adjusted and its adaptation to financial terminology, jargon, abbreviations and contextual meaning is enabled. Another important segment is text generation, which uses models such as GPT to generate a coherent and relevant financial response to a given request (prompt). In our case, it can be used to simulate the impact of news, generate scenario-based predictions or create a synthetic financial report for training and testing a model.

The integration of NLP brings with it a number of challenges, such as dealing with domain-specific jargon, handling sarcasm, noisy text (in the case of social media content), and real-time processing. In addition, careful selection of the data set and fine-tuning of parameters is critical to avoid overfitting and ensure that the extracted sentiment realistically corresponds to market behavior.

3.1 Objective and Organization of the NLP research

The objective of the research on the application of NLP for stock trading on the stock exchange is divided into two sub-objectives and they are: Finding an appropriate configuration for analyzing news for a given stock using FinBERT; Using the total sentiment, average sentiment and sentiment category obtained from the sentiment analysis, as parameters for generating a textual explanation of whether to buy, hold or sell a stock of a given company. For this purpose, 4 "notebooks" were created: NoFineTune, Pseudo-Fine-Tuned-Sentiment-Analysis, Real-Fine-Tune-Sentiment-Analysis and Text-Generation. Within the NoFineTune an evaluation of the finished FinBERT model according to its ability to accurately predict the sentiment of over 5800 headlines that have already been evaluated was done. Pseudo-Fine-Tuned-Sentiment-Analysis is used to perform pseudo fine-tuning, i.e. the news downloaded from the News API was analyzed by the finished FinBERT, followed by an automated evaluation in relation to the analysis of the Financial-RoBertA-large-sentiment model, fine-tuning according to those values and re-evaluation. In the Real-Fine-Tune-Sentiment-Analysis, a real fine-tuning was performed, whereby the finished FinBERT model was re-trained on the content of the file, after which it was used to analyze the news that were also analyzed

by Financial-RoBertA-large-sentiment. Finally, Text-Generation was used to generate a textual explanation for the next step, which has to be taken by the user, in the direction of whether to buy, hold or sell shares of a given company.

3.2 NLP Model and Data Used

In order to obtain text generation and prepare the parameters for its generation, three models were used (two of which for sentiment analysis and for text generation) and three data sources (a file with manually analyzed news, a file with sentiment category, average sentiment and aggregate sentiment and news downloaded from the News API). The models used are available on the Hugging Face platform and they are: 1. FinBERT by ProsusAI; 2. Financial RoBERTa Large Sentiment by soleimanian; 3. WiroAI-Finance-Qwen-1.5B by WiroAI. The FinBERT model for news analysis was used, taking into account that it is smaller and can be run on "more modest" configurations, compared to Financial RoBERTa Large Sentiment. This model is trained on a significantly larger corpus of data, which also includes news from the world of finance, and is therefore used for pseudo-fine-tuning and automatic evaluation. The last model, WiroAI-Finance-Qwen-1.5B [15], is one of the rare language models from the financial domain that is available through Hugging Face. This allows to use a model specific to the financial domain, not general models, such as GPT, DeepSeek, Gemini or Grog, meaning that the resulting text has greater precision, greater credibility and content.

The data used was divided into three parts: 5842 news items from Kaggle, manually analyzed and a negative, neutral or positive sentiment was assigned to each news item. The second part contains the total sentiment for the given set of news items, the average sentiment and the sentiment category based on the average sentiment, which will be used further for text generation and the third part contains news headlines downloaded using the News API. In this way, the finished news items are useful for testing the initial capability of FinBERT and the success of the fine-tuning process. Within the pseudo-fine-tuning and fine-tuning files, we have variables that will store a list of 100 news headlines from the last month for a given action. Namely, the number of news headlines is 100 and the period is one month. The research was done with the finalized FinBERT model from Hugging Face (using functions from the Transformers library) [14].

Pseudo-fine tuning of FinBERT. Considering the fact that most of the logic is repeated, pseudo-fine-tuning and fine-tuning were researched in terms of the differences in the approaches, through the prism of the defined functions, in relation to the initial testing of the "pure" FinBERT. A characteristic of pseudo-fine-tuning is that it does not rely on human-generated data (in our case, sentiment categories) to train the model. This approach is also known as semi-supervised learning. Pseudo-fine-tuning of FinBERT was done with the news retrieved from the News API. The Adam optimizer was used to optimize the model according to the learning rate (2e-5) and the weight decay rate (0.01) and the Cross-Entropy to calculate the average loss. Unlike the evaluation in the initial phase of the previous approach, here we have an evaluation of the predicted categories from RoBERTa (as true) and from the finished FinBERT (as predicted).

FinBERT Fine-Tuning. In order to properly perform and evaluate the further fine-tuning, the data are divided into sets for training and testing. The division is performed according to the 80/20 principle, where 80% will fall on the training set and 20% will fall on the testing set. Fine-tuning of the FinBERT model, in order to better respond to its task, i.e., news analysis was performed with Adam's optimizer, Cross Entropy as a loss function and a variable number of episodes. With this evaluation function, the quality of fine-tuning was evaluated, using the fine-tuned model for the same four measures.

Text generation. The essence of text generation begins by generating the variable which will serve to give the language model the context we want from it. First, the model must be informed what is known about the stock market sentiment for the given company and then asked for a recommendation whether the investor should buy, hold or sell a share of the given company. After that, a "pipeline" is generated for the given model, during which text is generated no longer than 500 words, without returning the previously sent request. To end this part of the function, we define conversation-like behavior, defining that the system (in our case the WiroAI-Finance-Qwen-1.5B model) should behave as a financial expert and respond to our request, contained in the third dataset. Within the second part, we call the corresponding text response to our request, followed by its purification and modification, in order to obtain a better, clearer output.

3.3 Results from the application of NLP

The NLP component builds directly on transformer-based architectures that have been widely applied in financial sentiment analysis. In particular, FinBERT and RoBERTa were employed to classify financial news and social media texts, enabling us to generate buy/hold/sell recommendations informed by market sentiment. Our findings are consistent with prior work demonstrating that transformer models capture domain-specific linguistic nuances in financial text. This situates our contribution within ongoing research that leverages pre-trained language models to enrich financial decision-making systems. The results of the application of NLP from two aspects show that the first aspect examines the four measures of success of the different configurations, when assessing the sentiment of the news from NewsAPI and the news from the from first dataset. The second aspect presents the process of obtaining the total sentiment, the average sentiment and the sentiment category for the shares of a given company and using the obtained data to generate a textual explanation for taking the next step, i.e. buying, selling and holding the shares of the given company.

3.4 NLP Model: Presentation of the Measures and Discussion of the Results

Table 2. Sentiment analysis configuration research measures

Configuration	Validation Loss	Accuracy dataset/ news	Precision dataset/ news	Recall dataset/ news	F1 dataset/ news
FinBERT	/	0,1761 / 0,21	0,2809/0,2513	0,1761/ 0,21	0,2006/0,2262

Pseudo fine-tune FinBERT (3/2e^{-5})	0.9210	0.15 / 0,68	0,1177/ 0,7652	0,15 /0,68	0,1304 /0,6399
Fine-tune FinBERT (3/2e^{-5})	0,2190	0,8024 / 0,63	0,8065/0,6587	0,8024/ 0,63	0,8039/0,6353
Fine-tune FinBERT (2/2e^{-5})	0,3144	0,8255/0,7273	0,811 / 0,7321	0,8255/0,7273	0,8125/ 0,7258
Fine-tune FinBERT (4/2e^{-5})	0,1780	0,7981 / 0,63	0,7889/0,6311	0,7981 / 0,63	0,7926/ 0,6150

In order to select a favorable configuration of FinBERT so that it can successfully analyze news, a study was conducted, comparing the application of the ready-made FinBERT model, pseudo-fine-tuning and tuning of the same model. In doing so, the impact of applying a different number of training episodes (the number to the left of the slash in the parentheses in Table 2.) at a constant learning rate of 2e-5, i.e., 0.00002, was tested. The measures validation loss, accuracy, precision, recall and F1 are used for the comparison. For each measure, with the exception of validation loss, their success is monitored when working with news from the first dataset and news from the News API (in the table marked with news). According to Table 2, it can be concluded: The ready-made FinBERT model gives poor results both when working with data from the CSV file and when working with news from the News API, which justifies further adjustment with pseudo-fine-tuning and fine-tuning; Pseudo-fine-tuning is a useful option if human-analyzed data is not available, although it comes at the cost of poor performance initially; The 2-episode configuration offers the best balance between performance and efficiency during training, which indicates good generalization when working with unknown data; Further increasing the number of episodes in our case does not bring significant improvements; The decrease in validation loss does not imply better performance with unknown data.

Demonstration of sentiment analysis and text generation. The demonstration of sentiment analysis and text generation was examined by examining file excerpts and the results they generate. Considering that the best configuration has already been chosen, sentiment analysis with both models was performed, after which total sentiment, average sentiment and sentiment category are calculated according to fine-tuned FinBERT. According to the fine-tuned FinBERT, the average sentiment of the news for META (based on news headlines in the timeframe from 16.05.2025 to 26.05.2025) is 1.1717, while the total sentiment is 116, which makes it neutral. Considering the values for average and total sentiment, the conclusion is that not all the news was downloaded. For text generation, the following output was received: Given the overall sentiment label of 'neutral' for the stock with ticker symbol META and the mean sentiment score of 1.1717, the stock does not exhibit strong support or resistance in recent market conditions. This suggests that the stock may not be attractive to buy at the current price. Considering the overall sentiment is neutral, the stock could be viewed as sideways or sideways in the short-term market. Therefore, an investor may choose to hold this stock

for the long-term, given its moderate support from the sentiment data. Taking into account the textual result obtained above, the user should hold on to the META shares they own (if they own any), in the hope that "better days" will come for buying or selling.

4 Conclusions and Further Research

In this research, the application of machine learning in the process of trading stocks on the stock exchange was investigated through two separate approaches, deep Q-learning and NLP. In the first approach, a deep Q-learning-based agent showed promising results, considering that it achieved an impressive profit of 38.2356% over a period of 14 months, which would mean an annual profit of 32.14% and thereby exceeding the current inflation rate in the United States, which for April 2025 was 2.4%. On the other hand, the application of technologies within the framework of NLP for sentiment analysis of financial news was investigated. Through the application of FinBERT, a series of experiments were performed with different configurations for fine-tuning (the configuration with 2 episodes and a learning rate of 2e-5 lead to the best results). It demonstrates a high value for F1 for already analyzed news (0.8125) and the most slightly decreased value for news downloaded from the News API (0.7258). The results themselves show that even without significant opportunities (in terms of quantity for already analyzed news and collected news with the News API), the application of sentiment analysis and text generation can serve for efficient and effective knowledge extraction from unstructured text and interpretation of the extracted knowledge, offering an additional layer of understanding of stock market trends.

Although, within the framework of this research, the application of deep Q-learning and NLP were considered separately, the results show that the future of intelligent stock trading systems lies in their integration into a single system. By combining them (in a near future, when pulling news from News API or other sources for free will cover a larger timeframe, to match the one of the data pulled with Yahoo Finance API), the best of both techniques could be used: the adaptability of the agent based on deep Q-learning and the real-time analysis, as the foundation of the application of NLP, would develop systems that can be significantly more intelligent than existing ones, resulting in systems that have the ability to react based on historical data, but can also predict stock market trends based on the latest news, reports or posts on social networks.

This will mean leveraging the fine-tuned sentiment analysis model, to evaluate the sentiment of historical news articles and generate a daily sentiment score (rather than an aggregate score over a certain timeframe, like the above mentioned 10 days), providing a powerful new lens for understanding market movement by quantifying the public perception and media tone. Ultimately, the generated daily sentiment score will enable more sophisticated trend analysis and improve the process of data-driven decision-making. The above mentioned integration would allow the agent based on deep Q-learning to learn policies that balance technical indicators with market sentiment, providing a unified framework for future work. Taking this into account, this research has laid a

serious foundation for further research and development of hybrid agents in the domain of stock trading proposals.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Awad AL, Elkaffas SM, Fakhr MW. Stock Market Prediction Using Deep Reinforcement Learning. *Applied System Innovation*. 2023; 6(6):106. https://doi.org/10.3390/asi6060106
2. G. Anjaneyulu, P. C. Shaker Reddy and P. Praveen, "A Hybrid Deep Reinforcement Learning Framework for Stock Market Prediction," *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkuru, India, 2024, pp. 1-5, doi: 10.1109/ICMNWC63764.2024.10872342.
3. Puh, K. and Bagić Babac, M. (2023), "Predicting stock market using natural language processing", *American Journal of Business*, Vol.38 No.2, pp. 41-61. https://doi.org/10.1108/AJB-08-2022-0124
4. Shi, Y., et al., Stock trading rule discovery with double deep Q-network, Applied Soft Computing,Vol 107,2021, 107320,ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021. 107320
5. Saberironaghi M, Ren J, Saberironaghi A. Stock Market Prediction Using Machine Learning and Deep Learning Techniques: A Review. *AppliedMath*. 2025; 5(3):76. https://doi.org/10.3390/appliedmath5030076
6. Li, X., et al., Stock Market Prediction Using Reinforcement Learning With Sentiment Analysis, International Journal on Cybernetics & Informatics (IJCI) Vol.12, No.1, 2023
7. Kalva, Sudhakar & Naganjaneyulu, s. (2023). Stock Market Investment Strategy Using Deep-Q-Learning Network. 10.1007/978-3-031-36402-0_45.
8. Papageorgiou G, Gkaimanis D, Tjortjis C. Enhancing Stock Market Forecasts with Double Deep Q-Network in Volatile Stock Market Environments. *Electronics*. 2024; 13(9):1629. https://doi.org/10.3390/electronics13091629
9. Veljanovska, K., Soft Computing for Adaptive Traffic Control, 12th International conference on Applied Internet and Information Technologies (AIIT2022), Zrenjanin, Serbia
10. Veljanovska, K,Gacovski, Z., Deskovski, S., Intelligent System for Freeway Ramp Metering Control, IEEE 6th International Conference Intelligent Systems Proceedings, Sofia, 2012
11. Sutton, R.S., and Barto, A.G., Reinforcement Learning - An Introduction second edition. MIT Press, Cambridge, Massachusetts, 2008
12. Yang, Z., Zhang, Y., Lin, D., A Theoretical Analysis of Deep Q-Learning, arXiv preprint, Washington D.C., 2019, No.1901.00137, https://arxiv.org/abs/1901.00137
13. Bishop, C.M, Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995
14. Araci, D., FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, Hugging Face, 2019, https://huggingface.co/ProsusAI/finbert
15. WiroAI Team, WiroAI-Finance-Qwen-1.5B, Hugging Face, 2024, https://huggingface.co/WiroAI/WiroAI-Finance-Qwen-1.5B

An exploratory paper into MoE and their application to AVs

Vladimir Djepovski[1] and Petre Lameski[1]

University of Ss. Cyril & Methodius,
Faculty of Computer Science and Engineering, Skopje
vladimir.djepovski@gmail.com
petre.lameski@finki.ukim.mk
https://www.finki.ukim.mk/en

Abstract. In this paper, we explored the principles behind the Mixture of Experts (MoE) architecture, but more importantly, its application to autonomous driving (AD) in general and autonomous vehicles (AVs) in particular. We identified the reasons why this architecture is so important, what makes it *special* but also what its current *pain points* and drawbacks are, especially in the context of AVs. While modestly exploited by academia in the field of AD/AV, we proved as expected that it is not completely neglected either and that it is mostly studied from a perspective of the Perception and Planning subtasks of an AV, all the while making continuous progress and coming up with some very interesting applications of MoE and LLMs affecting the very way we may interact with our vehicles in a not so distant future.

Keywords: Mixture of Experts · Autonomous Vehicles · Autnomous Driving.

1 Introduction

The main goal of this paper is to investigate to what extent Mixture of Experts (MoE) is used and applied to Autonomous Vehicles (AV). In doing so, we are simply *skimming the surface* of what the MoE architecture represents. In other words, we are not dealing with the details behind its very essence. In the context of autonomous driving technology, the same approach is even more pronounced due to the *commercialized aspect* of AVs.

MoE as an architecture is not new, but has regained in popularity as a result of recent events, specifically the launch of the *DeepSeek R1* LLM, which relies heavily on the MoE architectural design principles, thus achieving formidable computational efficiency, enabling the model to handle complex tasks without significant computational overhead or the need for a very high-end and SOTA (*state of the art*) hardware.

We believe that MoE makes sense, especially in resource-constrained systems such as AVs. Our expectation is that indeed, MoE is employed in autonomous driving (AD) technology, albeit it's applicability in AVs might not be sufficiently exploited by academia, partially due to the commercial nature of AVs.

In order to present a meaningful overview, this paper continues with Section 2 where the topic of *levels of autonomy* in AVs is briefly addressed before continuing with Section 3 where we attempt to present an idea about the architectural design of an AV's autopilot and the main components that go into it. We are fully aware of the speculative nature of the information presented in this section. It is our own rendering of the available information to what is arguably highly commercial and thus not so public and open in nature.

We then move on to Section 4 and briefly explore the topic of MoE and what it implies, without going into too much technical detail, before finally moving on to Section 5 where we address the main topic of this paper i.e. the specific application of MoE to AVs and how it is presently approached by academia.

2 Levels of autonomy

What does it actually mean for a vehicle to be *autonomous*? The *Society of Automotive Engineers* (SAE) defines 6 distinct levels of automation in their SAE J3016 standard, which technically is equivalent to the ISO/SAE DPAS 22736 standard (a joint publication between the ISO and SAE bodies). A free version of the J3016 standard available for download is one from UNECE [30]. However, the *SAE J3016 Levels of Driving Automation* summary chart provides enough detail and clarification [29].

The levels of autonomy as per the *SAE J3016* are: *Level 0* to *Level 5* [29], where *Level 0* implies no automation at all (basically all cars manufactured until the mid 90s) and *Level 5* is the highest degree of automation.

Table 1 summarizes all the different levels of autonomy in a vehicle as per the *SAE J3016* [29].

There are plenty of resources exploring the topic [39], [37], [38]. While most agree achieving *Level 5* is still ahead of us (albeit at a steady pace), there are those questioning how achieving mainstream *Level 3* will affect the rest of the automotive supply chain [35], specifically the effect over the growth and development of the silicon industry [36]. Arguably though, it is safe to say that most people will not accept any new developments towards autonomous vehicles until the technology is as safe as that implemented in aviation [39].

Still, an even more important question posed by some experts is: *do we really need Level 5 at all!?* [34] if *Level 4* implies the technology inside a vehicle will inherently decide if the conditions are deemed safe or otherwise.

3 Architectural overview of the autopilot in AVs

From a very broad and high-level perspective, the autopilot in an AV is actually very straightforward [3] (Figure 1).

We could argue the autopilot is composed of (arguably at least) 3 subsystems - a 1-*motion planner*, a 2-*vehicle controller* and a 3-*actuator controller* component.

Autonomy	Description	What it really implies
Level 0	No autonomy	There's absolutely no kind of autonomy present in the vehicle.
Level 1	Driver assistance	Some tech-enabled feature such as adaptive cruise control or anti-lock braking system where only one is active at any given moment.
Level 2	Partial autonomy	The vehicle can simultaneously control several functions such as e.g. steering and accelerating/braking but no self-driving is possible.
Level 3	Conditional autonomy	A technological leap over Level 2. A vehicle can handle most common scenarios but expects human intervention where deemed impossible to handle.
Level 4	High autonomy	Vehicles do not require human intervention in most cases and can fully autopilot the vehicle. Due to the current common legislative the technology is limited to certain zones e.g. where the speed of the vehicle is restricted to up to 30kmph - otherwise known as *geofencing*
Level 5	Full autonomy	No geofencing, no restrictions. In fact, as this technological milestone is achieved, there might not be any steering wheels in vehicles.

Table 1. SAE levels of vehicle autonomy

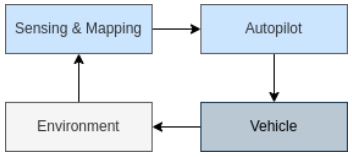

Fig. 1. Generalized architectural overview of an AV's control

The vehicle gets *a feel for the environment* from its sensors. We could also argue the environment affects the vehicle through the sensors which feed the motion planner, which in turn provides input to the vehicle controller. All subsystems play their role, but arguably it is the vehicle controller that decides how the vehicle should act within the environment and the context it is placed in, thus providing it as input to the actuator subsystem which drives the vehicle forward or backwards, steers it left or right and so on.

In reality of course, it is not that simple and if we start expanding even further on each subsystem thus adding more context and functionality, the entire diagram quickly gets more complex.

For example, the effects of the *Environment* on the vehicle are omnidirectional. The *Sensing & mapping* subsystem is actually an array of different types of sensors - from lidars to radars, from simple cameras to complex vision systems, ultrasonic sensors, GPS trackers and so on - each providing different types of output.

The vehicle does not experience the environment the same way humans do. To improve the reliability and enhance the accuracy, these outputs are then *synthesized* or *fused* [20] before further processed for either 1-*perception* or 2-*localization* purposes [12]. There could be some additional sensors involved that increase the awareness of the vehicle, such e.g. a microphone array that senses sirens from emergency vehicles - an approach that has been developed and employed by Waymo [32].

The *perception* part of the sensor network has probably the biggest impact on the performance of the AV [32], although the impact of the *localization* cannot be downplayed. An AV is equipped with multiple of these sensors [17] both in the front and the back but also on top of a vehicle and its sides. The part of the sensor network responsible for perception generates *point cloud* data i.e. a 3D representation of the world around the vehicle. Even camera sensors are being replaced with *depth* or *RGBD cameras* that are capable of generating point cloud data on top of a plain RGB image. The main point to keep in mind is that each sensor plays a role - especially in different and varying environmental conditions where one sensor outperforms the other [21].

Similar arguments apply for the localization part of the sensor network. Equipping the vehicle with GPS trackers only is not enough because they are generally lagging and depend on satellite coverage and exposure. Hence, the addition of IMUs or *Inertial Measurement Units* which provide immediate output such as e.g. orientation, current acceleration or deceleration of a vehicle.

If we take all of the above arguments and apply them to the generalization we made in Figure 1, we may expand the depiction of the *Sensing & Mapping* and *Autopilot* subsystems in Figure 2. Of course, there are other subsystems at play, such as *inter-vehicular communication* and *message exchange systems* which are basically used to exchange messages between vehicles in traffic (e.g. weather conditions, possible road blocks, etc.) [17], [32], [10], [26] and [1].

For the purposes of this paper and the statements made in it, we will remain on what is depicted in Figure 2.

Hence in summary, the *Sensor & Mapping* feeds the *Autopilot* with sensory input to get an *accurate* representation of the surrounding environment (*Perception*), while the rest is used to position the AV within that same environment (*Localization*) and some may be used to increase the AV's situational *Awareness*. From these subsystems, processed data is mostly fed to the *Motion Planning* while some may be fed directly to the *Vehicle Control* in case of an e.g. emergency stop.

Figure 2 is a subjective rendering of the interpreted referenced material, in particular the component represented as the *Context Provider*.

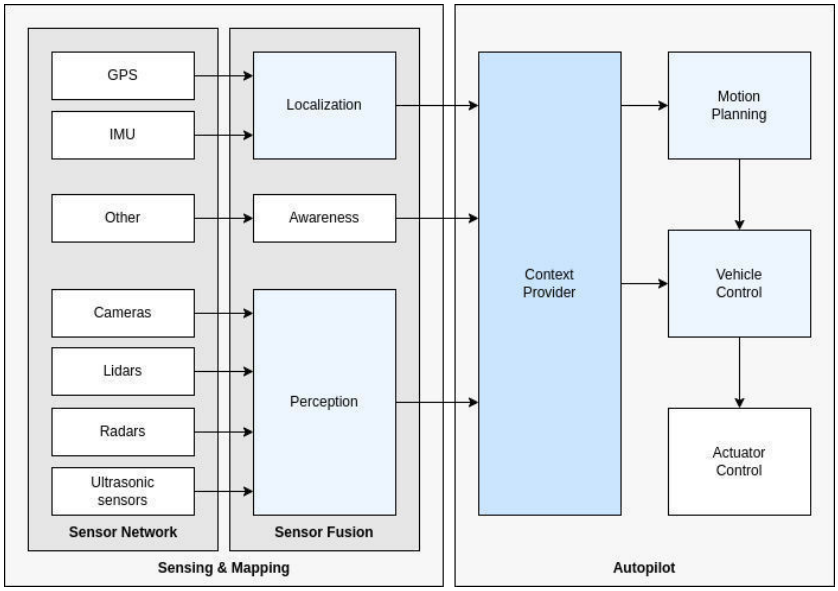

Fig. 2. An expanded view of the sensory input and the autopilot of an AV

If we look at any vehicle and consider the most basic functions it executes: 1-accelerate (speed up), 2-decelerate (slow down), 3-steer left or 4-steer right, we may argue it is the *Context Provider* that provides *the context* and whether the vehicle needs to take an immediate action (an abrupt stop?), slow down and make way (for an emergency vehicle?) or initiate a planned route to be taken.

In that regard, most of the subsequent research suggests a lot of emphasis has been given to the *Perception*, *Localization* and *Motion Planning* components. We argue that even though *Perception* is recognized as one the most (if not the most!?) important pillars of an AV, we cannot underestimate the significance of the remaining components and sub-systems.

4 An overview of MoE

The arguments made in the following section will not be an exhaustive elaboration on the topic of MoE but rather a gentle one so that the tone of the rest of the paper is established. For an excellent review on MoE and its main points, there are excellent online sources available, which capture the basics in a less formal tone [25].

We may argue that MoE is much like *context switching* if we apply a human brain analogy. Instead of one huge dedicated DNN (*Deep Neural Network*), the architecture of the derived network is to be broken into many smaller ones (often referred to as *experts*), each tailored to handle a very specific subject domain (*subtasks* within the context of a bigger problem) and on top of them all, one

that provides the context and decides which of the *specialized* neural networks does further processing (often referred to a *gating network* or *router* [2], which also learns over time which expert NN should be activated for a given input). This is often cited as *conditional computation* [23], [40].

To best depict what it is that MoE is to replace and improve, let us briefly look at the decoder-only transformer architecture in Figure 3 which is one of the basic building blocks of any LLM [11] [41]. In this representation, each input is *tokenized*, and each token is represented as a vector of *embeddings*), then each vector goes through several decoder blocks - and in each decoder block through a FFNN (*Feed Forward Neural Network*).

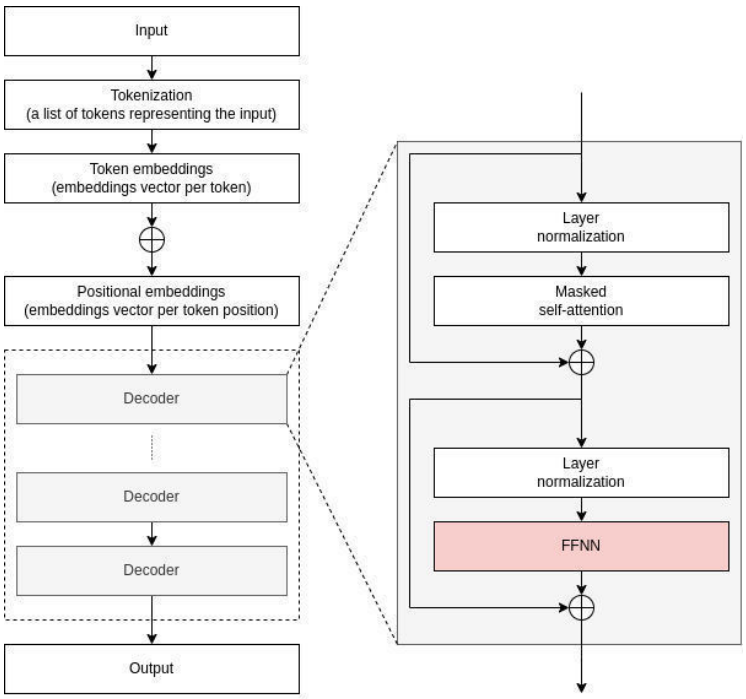

Fig. 3. A simplified view of the decoder architecture

This block or rather its neural network representation is very dense in the sense that each node is connected to all the rest of the nodes because inside the FFNN, each input is activated (Figure 4). In this example, each word, number symbol and punctuation sign represents one token and each token is considered to be represented by an embeddings vector with a dimension of 4. From this representation of the FFNN we may realize how this dense network could easily get complicated really fast.

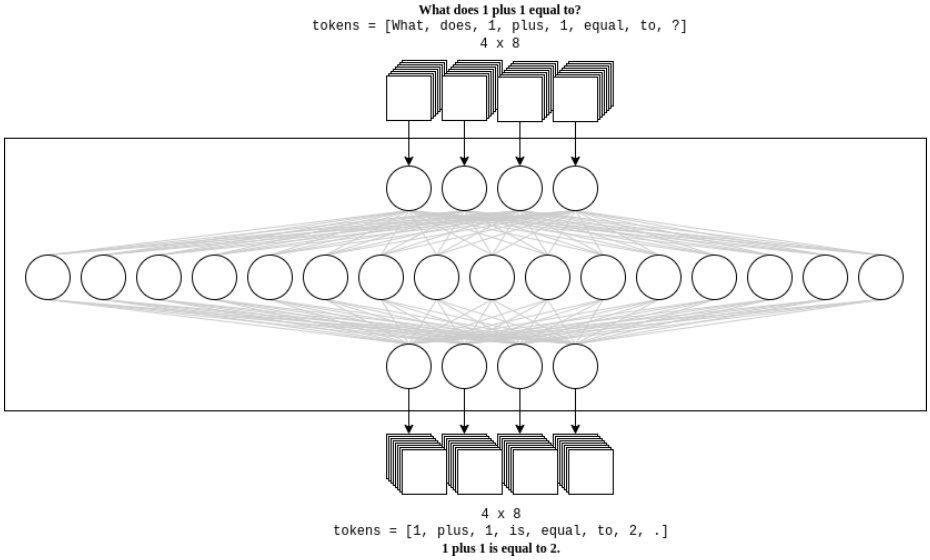

Fig. 4. A dense FFNN block

To go around this problem and activate only those nodes which are required, that dense FFNN block is replaced by several other - 'smaller' FFNNs which represent the so called *experts*! (Figure 5). The router is also a dedicated FFNN who's purpose is to: 1-calculate the probability relevance of an expert and 2-activate the *top K* experts which will form the output.

Common gating mechanisms usually are: *binary* or *sparse and continuous*, *stochastic* or *deterministic* [23], [42]. However, some authors have pointed to the notion that these commonly used gating mechanisms do not work any better than simply randomly routing inputs to experts and have come up with THOR (or *Transformer witH StOchastic ExpeRts*) - a mechanism where experts are randomly being activated for each of the inputs during both training and inference which according to the authors, outperform known transformer-based and MoE models [31].

The general consensus among researchers identified with this paper is that scaling up dense models (with more input parameters) does improve the performance of the model but at the expense of huge energy demands to power those models. Not only this is not technologically sustainable, but has also raised (and rightfully so!) environmental and societal concerns [2], [19], [45]. Furthermore, practically every source listed in this paper has pointed to migrating from dense to sparse models and MoE has a pivotal role to play.

In 'traditional' deep learning models and DNNs (*dense* models), the entire network is activated for each input sample it receives, resulting in (at least!) double the training costs for each subsequent increase in the model size and the size of the training set - a trend that is almost impossible to maintain [23]! On the

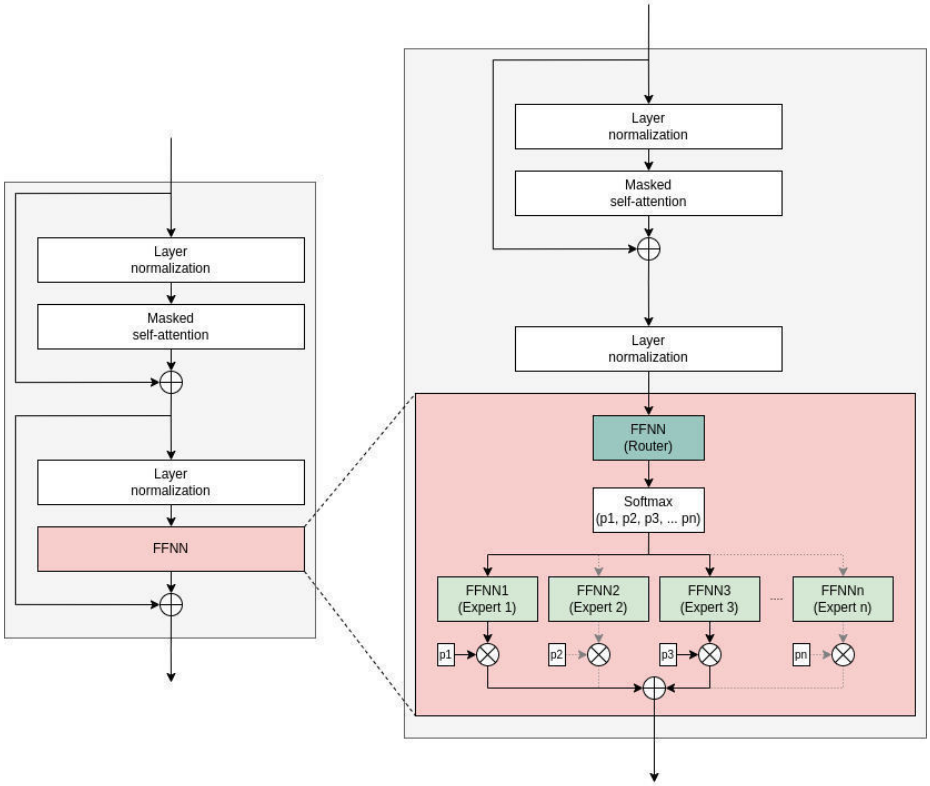

Fig. 5. A decoder with a MoE layer

contrary, in MoE enabled architectures, only a selected number of expert NNs are activated (sometimes even one!). This is what is known as *sparse activation*. In the training phase, each expert NN is individually trained but also jointly trained with the gating network, to ensure seamless co-operation. Over time this results in establishing balance between processing, training, performance and energy demands, making MoE an efficient tool in the AI domain.

As with most other concepts in AI, MoE is not a new paradigm and has been proposed by academia as early as the beginning of the early 90's of the last century [2], [4], [13], [24], [40]. Today, it is mostly used in NLP (*Natural Language Processing*) and LLMs (*Large Language Models*) but also in vision models (V-MoE or Vision MoE) by utilizing ViTs (*Vision Transformers*) [9], [14], [42], [44], where the dense feed-forward neural network (FFNN) in the ViT encoder is replaced with sparse MoE layers allowing for scaling up the vision models without affecting the overall efficiency and performance of a system with fewer resources. This is exactly why MoE (and its application) is of particular interest and importance to AVs, especially in the *Perception* sub-systems.

Therefore, if MoE as a paradigm is at least 3 decades old, why is it that only recently it is gaining so much attention?

According to the literature review, despite the promise it holds, it is not so straightforward to implement MoE in practice [23], [40]. Two of the most cited challenges are the choice of a routing strategy and training deficiencies, where the first often affects the latter.

The most common routing strategy employed in early MoE based models was to route the input tokens to the experts identified as the best match, according to their routing scores and hidden representations. This has since been recognized as leading towards *token clustering* where most of the input tokens center around the same expert *centroids* which in turn leads to *representation collapse* [45]. Employing an inadequate routing strategy not only leads to over-specializing or over-training some experts, but the opposite as well i.e. under-specializing or under-training others [42]. Some researchers have pointed out that merely activating one expert at a time limits a model's upper bound performance while activating too many of them is not affordable during training and inference runs. Furthermore, when the number of activated experts is too high, there might not even be any further performance improvement at all [9], [13], [43]!

The phenomenon of *training instabilities* has also been long observed and recognized [4] in that dense and sparse models behave differently across sample batches of different sizes where the learning rate also varies [2] and easily diverges [44].

Some authors have even identified that the *network bandwidth* leads to poor performance of models built around MoE. If we view the model as a network where the *model's router* and the individual model experts as all interconnected, that *interconnectivity* is the bottleneck i.e. it is slower than its individual components [23].

In order to address all of these issues and produce MoE based models whose architectures are easy to understand and which make efficient use of all available HW resources without immediately defaulting to scaling up, researchers have come up with different proposals: from converging experts into clusters and moving from MoE to MoEC (*Mixture-of-Expert Clusters*) where each cluster gets an equal chance of more diverse training samples [43] to a more heterogeneous *expert choice* approach where the experts select the top-K tokens as opposed to the other way around, resulting in each token-to-expert route having an equal chance of being selected. This ultimately leads to near performance load balancing [42]. Something similar has also been proposed by [4], advocating for a two-stage training process, where in the first stage learning is employed to get to a lightweight routing model and strategy that is decoupled from the backbone model. Then in the second stage, the *distilled* router is used to determine the token-to-expert assignment and freeze it.

While some researchers have opted for expert selection by hierarchical routing in what they refer to as a SAM or a (*Sparsely Activated Model*) where essentially a *switch router* selects a *device* within the larger model context from where a *mixture router* in turn selects the expert(s) within the selected device [13],

others have recognized the benefits of both mainstream approaches (i.e. dense vs. sparse) and are advocating for an *Efficient Ensemble of Experts* (E^3) [14].

DNNs (essentially dense models) are proven to show performance gains from further calibration even when under varying datasets, aggregating individual sub-model predictions, but in doing so at a cost of significant increase in computational cost. Sparse MoEs on the other hand, decouple training and growth by employing conditional computation.

5 The application of MoE in AVs

All *autonomous systems* in general (e.g. self-driving cars or AVs, robots) benefit from utilizing MoE on an architectural level because this enables a *modular decision-making* process and thus an *adaptive control strategy* [5]. In robotics in particular, a MoE-based controller may improve the adaptability of the robot to e.g. different terrains. Some researches have come up with an efficient adaption of MoE and combined multiple experts, addressing the terrain segmentation problem in robotics enabling autonomous and safe navigation in highly unstructured outdoor environments [27]. This arguably, equally applies to AVs as well as vehicles do not always drive over paved and marked roads.

Specifically to AVs however, a MoE-based architecture improves the perception capabilities of a vehicle and help adapt it to distinctively different driving scenarios [5]. The key problem however is whether such an architecture could effectively be applied to resource-constrained systems and in real time?

Unsurprisingly, *pedestrian detection* is identified as one of the key (pain)points of any kind of practical implementation of AI in AVs or any other autonomous system where *people as pedestrians* are key actors [8]. The authors of the cited paper have applied *a novel, multi-level approach* based around MoE to significantly improve pedestrian detection and classification by combining information from multiple cues and features. They also point to the fact that most current applications to this problem involve a 2-step approach: *feature extraction* and *pattern recognition* and emphasize that *depth* and *motion* are also key factors to consider that add and improve the entire process, thus discriminating from *pedestrians* and *non- pedestrians*. Their multilevel MoE framework is pose-specific and i.e. utilizes individual expert classifiers on pose, modality and feature levels, thus breaking down the problem of pedestrian classification into smaller yet better manageable problems. The individual experts classifiers in their framework are independent from each other, do not have to use the exact same training dataset and are thus less prone to over-fitting. Furthermore, if they are trained against different datasets and independently of each other, the resulting training times are much shorter.

The importance of accurately predicting human motion with low inference times, as being of extreme importance in real-time applications such as AVs has also been recognized by [6]. Their approach is somewhat different than [8]. Recognizing findings from prior research, in that traditional applications based on *Recurrent Neural Networks* (RNNs) do not cope well with processing sequential

information at low inference times, have turned their own attention to *Spatial-Temporal Transformers* and have integrated a MoE in the attention layer of the ST Transformer model, in what the authors believe is a novel approach that effectively predicts human motion while still within the realm of *real time* [6]. The approach the authors take, aims at optimizing the model's inference speed by selecting the most relevant components of the model during prediction, and allows for up-scaling without escalating computational demands.

As already outlined, perception modules or sub-systems play a crucial role in AVs because it is through these modules that AVs 'get a feel for the environment' and object detection is a critical task of any perception module [18]. To make it even more challenging, any detection task needs to happen in real time. But there is a trade-off to be made between high accuracy and low latency and existing systems and architectures tend to lean more towards one or the other but seldom both. Achieving both is the *Holy Grail* in perception module design and development. As the authors [18] have put it "an accurate but slow system fails to react in time, while a fast but inaccurate system makes unsafe decisions". The authors' contribution to this is EMC2 - or an *Edge-based Mixture-of-Experts Collaborative Computing* architecture that aims at striking a balance between accuracy and latency in AVs perception. Instead of a monolithic architectural model, the researches opted for a modular design, leveraging MoE to activate the most suitable/specialized expert depending on the characteristics and the context from the multimodal input layer which preprocesses data from all kinds of sensors.

Tracking a vehicle's steering angle is another important machine vision based problem in ADAS (*Automated Driving and Advanced Driver Assistance Systems*) with the task of accurately predicting the steering angle of a vehicle under all environmental and driving conditions using images captured by the vehicle's on-board camera(s) [15], [16]. At the same time, the accuracy of such systems should carefully be balanced with efficiency.

SOTA (*State-of-the-Art*) vision-based algorithms accomplish the task of predicting a vehicle's steering angle using end-to-end deep learning networks. The contribution of [15] to the already established body of knowledge is a *novel filtering algorithm* which is scene-based and employs the front camera of a vehicle and a LSTM (*Long Short Term Memory*) network of MoE - capable of robustly predicting and tracking a vehicle's steering angle under what the authors argue are - varying and challenging driving conditions. The initial idea as presented by [15] was further elaborated, extended and validated by [16].

Motion planners as previously stated in Section 3, are yet another key component of the AV's technology stack where as it currently stands, most of the decision making is executed [33]. At the same time, the authors point to the shortcomings of the motion planners, referring to them as the *weakest link* and the reason why AVs are currently still constrained and limited to specific operational domains (e.g. *zone fencing* in Level 4 autonomy).

A motion planner's task is to aid an AV in safely navigating all, but especially public roads [33] with high frequency of other actors such as e.g. vehicles

and pedestrians alike. Many of the established *conservative* methods to trajectory planning are based on specific or *handcrafted* rules while ML-based systems offer better generalization and scale rather well proportionally as the available training data scales up and are thus are more capable at learning more complex system behaviors [33]. Generalization put aside, these ML-based models on the other hand, when scaled up, render many of the incorporated design elements redundant as pointed out by [28] and are computationally inefficient and incapable of coping with unfamiliar driving scenarios [33].

One approach as suggested by [33] is their *SafePathNet* which represents an architectural design based on a DNN (*Deep Neural network*) which scales with data and essentially models and predicts at the same time a vehicle's possible trajectories as well as future locations of other actors/agents in a given scene. It then employs a MoE approach to select the best possible trajectory from a given distribution of learned i.e. predicted trajectories. A different approach as suggested by [28] is their *StateTransformer-2* (STR2), ultimately based on similar principles, but specifically benefiting from a ViT encoder (*Vision Transformer*) and a MoE *causal transformer* backbone thus representing a scalable and a decoder-only motion planner capable of modeling various driving rewards by various experts at each layer. The authors point out that most bottlenecks in systems such as motion planners happen as a result of trying to over-generalize situations and environments which otherwise are very complex, unpredictable and inconsistent such as the example we all can relate to, that in principle: "human drivers might cross solid white lines while overtaking slow traffic ahead" even if they're not supposed to according to any kind of a rule-based mechanism.

Certain flaws in both approaches exist and the authors are very transparent in exposing them. For example, as is in the case of the SafePathNet architecture, it is not guaranteed that all predicted trajectories from a learned distribution will be collision free. However, given that a collision is present and the so called *time-to-collision* is high enough, the model is capable of finding an alternate - collision-free trajectory in a subsequent re-planning cycle [33]. In the case of [28], even though the planning performance of the STR2 is ranked by the authors as *superior*, its inference performance is a different story and ultimately leads to slower inference as the model scales up.

An interesting application of MoE in AVs is in *Visual Odometry* or more specifically *Visual-inertial Odometry* (VO) i.e. determining the position and orientation of a vehicle [22] in unknown or unseen environments. We have already concluded that *Localization* and therefore VO as a subset of Localization much like *Perception*, is one more of the main components of an AV's technology stack. Most of the practical applications of VO are monocular or stereo-centric whereas SOTA AVs are equipped with multiple cameras, practically covering 360° *field of view* (FOV). As the authors further elaborate, it is already established that MoE-based techniques have successfully been deployed in various ML/CV tasks, ranging from: 1-*image classification*, 2-*object detection* to 3- *segmentation* and 4-*human pose estimation* but that probably their novel MIXO (*MIxture of eXperts Odometry*) approach is the first of its kind applied to a VO task in AVs,

combining odometry outputs from multiple sources (when available) to provide an optimal positional "vantage point" of the vehicle in any given driving scenario.

However, the precision that MIXO provides, comes at the cost of efficiency the authors warn i.e. getting the best possible odometry results from multiple sensors requires significantly more computational power thus putting the real-time efficiency of such a system in jeopardy. Another downside of the MIXO approach according to the authors [22] is that it essentially is a data-driven approach, and as with every data-driven approach, it depends on a large dataset. In fact, the bigger the dataset the better. Furthermore, all training is done in supervised mode which is known not to generalize well.

Last but not least, the use of VLMs (*Vision Language Models*) in autonomous driving applications, which through Q&A interactions, perform (or aid to) sub-tasks such as planning, prediction, perception and/or decision-making in a unified end-to-end model [7]. According to the authors of the study, such applications are not uncommon and already exist to some extent in AVs and in autonomous driving in general with the potential to become the new norm when it comes to how the AI agent in a car interacts with the driver.

The primary two components of any VLM are: 1-a vision encoder and 2-a LLM for any text generation, which implies that deploying them in practice would be computationally expensive. Furthermore, the majority of the known VLMs the authors of the afore-referenced paper argue, are being trained on single images (from monocular sources) but a modern vehicle today, with any sort of AI implemented in it, comes equipped as already suggested - with multiple cameras and this needs to be taken into a consideration in modern VLMs as well. This is where *MiniDrive* steps in, which according to [7] does not try to be a unified, transformer-based model but rather one which is derived from an "efficient backbone model" such as the *UniRepLKNet* and FE- MoE (*Feature Engineering MoE*) plus a DI-Adapter (*Dynamic Instruction Adapter*) which process *visual features* in order to obtain *visual tokens* which then serve as input to the LLM, which in turn generates natural language responses.

While the MiniDrive model addresses most of the challenges for an effective deployment of VLMs in AVs (robotics too!), such as demonstrating performance while still maintaining a real-time response, it certainly is not without any drawbacks, one of them being *hallucination issues* as well as the lack of generalization which the authors believe comes as a result of the limitations in the training set and have called for more open efforts and public datasets to address the anomalies in their model but we also would argue, for the sake of advancement of science and technology in general!

6 Conclusion

The topic of *Mixture of Experts* (MoE) has been well researched in the past few decades. It is still actively researched, as suggested by the sheer number of peer-reviewed papers on the subject. However, the same cannot be argued about the application of MoE to the domain of AVs (or autonomous driving, in general).

Scouring the available literature in an attempt to identify reliable sources on this very topic yielded comparatively less results than the more broader topic of MoE. A similar argument applies to many of the points and topics presented in Section 3. But as already argued, this was to be expected due to the commercial aspect of AVs.

In conclusion, the review of the literature confirms that the MoE architectural design is successfully being applied to AVs. This research suggests that MoE is mostly used for subtasks within the domain of an AV's *Perception* module (most notably the papers on Multilevel MoE [8], MoE with ST Transformer [6], EMC2 [18]) but also the *Motion Planning* module (e.g. SafePathNet [33], STR2 [28] and LSTM network of MoE [15], [16], albeit the latter is more inline with a subsequent AV module i.e. the *Vehicle Control* module, where the steering angle of a vehicle is determined from its visual sensor feeds).

MoE has also been successfully implemented in another important component of an AV's technology stack - *Localization* (MIXO [22]). However, a pleasant novelty that took the authors by surprise was the application of MoE in combination with LLMs in autonomous driving and robotic applications (MiniDrive [7]). It is these types of innovative solutions that emphasize the importance and the impact of academic research with the potential to push progress forward.

References

1. Ahangar, M. Nadeem and Ahmed, Qasim Z. and Khan, Fahd A. and Hafeez, Maryam. A Survey of Autonomous Vehicles: Enabling Communication Technologies and Challenges. *Sensors*, 21(3), 2021. URL: https://www.mdpi.com/1424-8220/21/3/706, doi:10.3390/s21030706.
2. Barret Zoph and Irwan Bello and Sameer Kumar and Nan Du and Yanping Huang and Jeff Dean and Noam Shazeer and William Fedus. ST-MoE: Designing Stable and Transferable Sparse Expert Models, 2022. URL: https://arxiv.org/abs/2202.08906, arXiv:2202.08906.
3. Berntorp, Karl and Hoang, Tru and Quirynen, Rien and Di Cairano, Stefano. Control Architecture Design for Autonomous Vehicles. *2018 IEEE Conference on Control Technology and Applications (CCTA)*, pages 404–411, 2018. doi:10.1109/CCTA.2018.8511371.
4. Damai Dai and Li Dong and Shuming Ma and Bo Zheng and Zhifang Sui and Baobao Chang and Furu Wei. StableMoE: Stable Routing Strategy for Mixture of Experts, 2022. URL: https://arxiv.org/abs/2204.08396, arXiv:2204.08396.
5. Dimitri, Vasily and Regina, Barbara and Alfonz, Magdolna. A Survey on Mixture of Experts: Advancements, Challenges, and Future Directions, 2025. URL: http://dx.doi.org/10.36227/techrxiv.173835706.69246194/v1, doi:10.36227/techrxiv.173835706.69246194/v1.
6. Edmund Shieh and Joshua Lee Franco and Kang Min Bae and Tej Lalvani. A Mixture of Experts Approach to 3D Human Motion Prediction, 2024. URL: https://arxiv.org/abs/2405.06088, arXiv:2405.06088.
7. Enming Zhang and Xingyuan Dai and Yisheng Lv and Qinghai Miao. MiniDrive: More Efficient Vision-Language Models with Multi-Level 2D Features as Text Tokens for Autonomous Driving, 2024. URL: https://arxiv.org/abs/2409.07267, arXiv:2409.07267.

8. Enzweiler, Markus and Gavrila, Dariu M. A Multilevel Mixture-of-Experts Framework for Pedestrian Classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, 2011. doi:10.1109/TIP.2011.2142006.
9. Fuzhao Xue and Ziji Shi and Futao Wei and Yuxuan Lou and Yong Liu and Yang You. Go Wider Instead of Deeper, 2021. URL: https://arxiv.org/abs/2107.11817, arXiv:2107.11817.
10. Garikapati, Divya and Shetiya, Sneha Sudhir. Autonomous Vehicles: Evolution of Artificial Intelligence and the Current Industry Landscape. *Big Data and Cognitive Computing*, 8(4), 2024. URL: https://www.mdpi.com/2504-2289/8/4/42, doi:10.3390/bdcc8040042.
11. Grootendorst, Maarten. A Visual Guide to Mixture of Experts (MoE), 2024. Accessed: April 25, 2025. URL: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts.
12. Guo, Qiaochu. Software System of Autonomous Vehicles: Architecture, Network and OS, 2020. Accessed: March 18, 2025. URL: https://fisher.wharton.upenn.edu/wp-content/uploads/2020/09/Thesis_Nova-Qiaochu-Guo.pdf.
13. Hao Jiang and Ke Zhan and Jianwei Qu and Yongkang Wu and Zhaoye Fei and Xinyu Zhang and Lei Chen and Zhicheng Dou and Xipeng Qiu and Zikai Guo and Ruofei Lai and Jiawen Wu and Enrui Hu and Yinxia Zhang and Yantao Jia and Fan Yu and Zhao Cao. Towards More Effective and Economic Sparsely-Activated Model, 2021. URL: https://arxiv.org/abs/2110.07431, arXiv:2110.07431.
14. James Urquhart Allingham and Florian Wenzel and Zelda E Mariet and Basil Mustafa and Joan Puigcerver and Neil Houlsby and Ghassen Jerfel and Vincent Fortuin and Balaji Lakshminarayanan and Jasper Snoek and Dustin Tran and Carlos Riquelme Ruiz and Rodolphe Jenatton. Sparse MoEs meet Efficient Ensembles, 2023. URL: https://arxiv.org/abs/2110.03360, arXiv:2110.03360.
15. John, V. and Mita, S. and Tehrani, H. and Ishimaru, K. Automated driving by monocular camera using deep mixture of experts. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 127–134, 2017. doi:10.1109/IVS.2017.7995709.
16. John, Vijay and Boyali, Ali and Tehrani, Hossein and Ishimaru, Kazuhisa and Konishi, Masataka and Liu, Zheng and Mita, Seiichi. Estimation of Steering Angle and Collision Avoidance for Automated Driving Using Deep Mixture of Experts. *IEEE Transactions on Intelligent Vehicles*, 3(4):571–584, 2018. doi:10.1109/TIV.2018.2874555.
17. Ján Ondruš and Eduard Kolla and Peter Vertaľ and Željko Šarić. How Do Autonomous Cars Work? *Transportation Research Procedia*, 44:226–233, 2020. LOGI 2019 - Horizons of Autonomous Mobility in Europe. URL: https://www.sciencedirect.com/science/article/pii/S2352146520300995, doi:10.1016/j.trpro.2020.02.049.
18. Liu, Linshen and Su, Boyan and Wu, Guanlin and Guo, Cong and Yang, Hao Frank. Multimodal Mixture-of-Experts Computing System on Edge for Efficient and Robust 3D Detection in Autonomous Driving. *SSRN Electronic Journal*, page 24, 03 2025. URL: https://ssrn.com/abstract=5167416, doi:http://dx.doi.org/10.2139/ssrn.5167416.
19. Mikel Artetxe and Shruti Bhosale and Naman Goyal and Todor Mihaylov and Myle Ott and Sam Shleifer and Xi Victoria Lin and Jingfei Du and Srinivasan Iyer and Ramakanth Pasunuru and Giri Anantharaman and Xian Li and Shuohui Chen and Halil Akin and Mandeep Baines and Louis Martin and Xing Zhou and Punit Singh Koura and Brian O'Horo and Jeff Wang and Luke Zettlemoyer and Mona Diab and Zornitsa Kozareva and Ves Stoyanov. Efficient Large Scale Language Modeling with

Mixtures of Experts, 2022. URL: https://arxiv.org/abs/2112.10684, arXiv:2112.10684.
20. Miller, Tymoteusz and Durlik, Irmina and Kostecka, Ewelina and Borkowski, Piotr and Łobodzińska, Adrianna. A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger. *Electronics*, 13(18), 2024. URL: https://www.mdpi.com/2079-9292/13/18/3660, doi:10.3390/electronics13183660.
21. Miller, Tymoteusz and Durlik, Irmina and Kostecka, Ewelina and Borkowski, Piotr and Łobodzińska, Adrianna. A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger. *Electronics*, 13(18), 2024. URL: https://www.mdpi.com/2079-9292/13/18/3660, doi:10.3390/electronics13183660.
22. Morra, Lia and Biondo, Andrea and Poerio, Nicola and Lamberti, Fabrizio. MIXO: Mixture Of Experts-Based Visual Odometry for Multicamera Autonomous Systems. *IEEE Transactions on Consumer Electronics*, 69(3):261–270, 2023. doi:10.1109/TCE.2023.3238655.
23. Noam Shazeer and Azalia Mirhoseini and Krzysztof Maziarz and Andy Davis and Quoc Le and Geoffrey Hinton and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, 2017. URL: https://arxiv.org/abs/1701.06538, arXiv:1701.06538.
24. Noam Shazeer and Azalia Mirhoseini and Krzysztof Maziarz and Andy Davis and Quoc Le and Geoffrey Hinton and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, 2017. URL: https://arxiv.org/abs/1701.06538, arXiv:1701.06538.
25. Omar Sanseviero and Lewis Tunstall and Philipp Schmid and Sourab Mangrulkar and Younes Belkada and Pedro Cuenca. Mixture of Experts Explained, 2023. URL: https://huggingface.co/blog/moe.
26. Parekh, Darsh and Poddar, Nishi and Rajpurkar, Aakash and Chahal, Manisha and Kumar, Neeraj and Joshi, Gyanendra Prasad and Cho, Woong. A Review on Autonomous Vehicles: Progress, Methods and Challenges. *Electronics*, 11(14), 2022. URL: https://www.mdpi.com/2079-9292/11/14/2162, doi:10.3390/electronics11142162.
27. Procopio, Michael J. and Kegelmeyer, W. Philip and Grudic, Greg and Mulligan, Jane. "Terrain Segmentation with On-Line Mixtures of Experts for Autonomous Robot Navigation". In Benediktsson, Jón Atli and Kittler, Josef and Roli, Fabio, editor, *Multiple Classifier Systems*, pages 385–397, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
28. Qiao Sun and Huimin Wang and Jiahao Zhan and Fan Nie and Xin Wen and Leimeng Xu and Kun Zhan and Peng Jia and Xianpeng Lang and Hang Zhao. Generalizing Motion Planners with Mixture of Experts for Autonomous Driving, 2024. URL: https://arxiv.org/abs/2410.15774, arXiv:2410.15774.
29. SAE International. SAE J3016 Levels of Driving Automation, 2021. Accessed: March 14, 2025. URL: https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf.
30. SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, 2021. Accessed: March 14, 2025. URL: https://wiki.unece.org/download/attachments/128418539/SAE%20J3016_202104.pdf?api=v2.
31. Simiao Zuo and Xiaodong Liu and Jian Jiao and Young Jin Kim and Hany Hassan and Ruofei Zhang and Tuo Zhao and Jianfeng Gao. Taming Sparsely Activated Transformer with Stochastic Experts, 2022. URL: https://arxiv.org/abs/2110.04260, arXiv:2110.04260.

32. Singh, Sehajbir and Saini, Baljit Singh. Autonomous cars: Recent developments, challenges, and possible solutions. *IOP Conference Series: Materials Science and Engineering*, 1022(1):012028, Jan 2021. URL: https://dx.doi.org/10.1088/1757-899X/1022/1/012028, doi:10.1088/1757-899X/1022/1/012028.
33. Stefano Pini and Christian S. Perone and Aayush Ahuja and Ana Sofia Rufino Ferreira and Moritz Niendorf and Sergey Zagoruyko. Safe Real-World Autonomous Driving by Learning to Predict and Plan with a Mixture of Experts, 2022. URL: https://arxiv.org/abs/2211.02131, arXiv:2211.02131.
34. The Arrow Editorial Staff. Autonomous driving levels explained, 2023. Accessed: March 14, 2025. URL: https://www.arrow.com/en/research-and-events/articles/vehicle-autonomy-levels-explained.
35. The Car ADAS Editorial Staff. SAE Autonomous Level 3, The Level of Driving Automation, 2023. Accessed: March 14, 2025. URL: https://caradas.com/sae-autonomous-level-3-levels-of-driving-automation/.
36. The IDTechEx Editorial Staff. AI and the Road to Full Autonomy in Autonomous Vehicles, 2023. Accessed: March 14, 2025. URL: https://www.idtechex.com/en/research-article/ai-and-the-road-to-full-autonomy-in-autonomous-vehicles/29902.
37. The Imagination Tech Editorial Staff. What Are The Six Levels Of Autonomous Driving Technology?, 2025. Accessed: March 14, 2025. URL: https://www.imaginationtech.com/future-of-automotive/when-will-autonomous-cars-be-available/what-are-the-levels-of-autonomy-in-self-driving-cars/.
38. The Neonode Editorial Staff. Driving the Future – Levels of Autonomous Vehicles Explained, 2025. Accessed: March 14, 2025. URL: https://neonode.com/newsroom/article/driving-the-future-levels-of-autonomous-vehicles-explained.
39. The Synopsys Editorial Staff. The 6 Levels of Vehicle Autonomy Explained, 2025. Accessed: March 14, 2025. URL: https://www.synopsys.com/blogs/chip-design/autonomous-driving-levels.html.
40. William Fedus and Barret Zoph and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022. URL: https://arxiv.org/abs/2101.03961, arXiv:2101.03961.
41. Wolfe, Cameron R. Ph.D. Decoder-Only Transformers: The Workhorse of Generative LLMs, 2024. Accessed: April 25, 2025. URL: https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse.
42. Yanqi Zhou and Tao Lei and Hanxiao Liu and Nan Du and Yanping Huang and Vincent Zhao and Andrew Dai and Zhifeng Chen and Quoc Le and James Laudon. Mixture-of-Experts with Expert Choice Routing, 2022. URL: https://arxiv.org/abs/2202.09368, arXiv:2202.09368.
43. Yuan Xie and Shaohan Huang and Tianyu Chen and Furu Wei. MoEC: Mixture of Expert Clusters, 2022. URL: https://arxiv.org/abs/2207.09094, arXiv:2207.09094.
44. Yuxuan Lou and Fuzhao Xue and Zangwei Zheng and Yang You. Cross-token Modeling with Conditional Computation, 2022. URL: https://arxiv.org/abs/2109.02008, arXiv:2109.02008.
45. Zewen Chi and Li Dong and Shaohan Huang and Damai Dai and Shuming Ma and Barun Patra and Saksham Singhal and Payal Bajaj and Xia Song and Xian-Ling Mao and Heyan Huang and Furu Wei. On the Representation Collapse of Sparse Mixture of Experts, 2022. URL: https://arxiv.org/abs/2204.09179, arXiv:2204.09179.

A Machine Learning Pipeline for Enhanced Speaker Diarization Using Segmentation, VAD, and GMM-Based Clustering

Dimitar Marenoski, Mile Pelivanov, Jovan Kalajdjieski, and Kostadin Mishev

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
dimitar.marenoski@students.finki.ukim.mk,
mile.pelivanov@students.finki.ukim.mk, jovan.kalajdzieski@finki.ukim.mk,
kostadin.mishev@finki.ukim.mk

Abstract. In this paper, we present a novel methodology for speaker diarization utilizing a Machine Learning approach. Unlike conventional methods that rely on manually crafted spectral features for speaker embeddings, our method employs a pipeline of algorithms to enhance speaker segmentation and clustering. Initially, we develop a batching algorithm to serialize the dataset, followed by the implementation of a multi-class classification model to estimate the number of speakers in an audio file, which informs the subsequent segmentation stage. We then introduce a voice activity detection (VAD) module to distinguish between human speech and silence. To refine this process, a Gaussian Mixture Model (GMM) ensures that only segments identified as human speech by the VAD are processed further. Finally, we apply a clustering algorithm to group speech segments by speaker, based on similarity criteria. Experimental results show a significant improvement in accuracy compared to other well-known diarization techniques. The code implementation is given at the following link [1].

Keywords: Speaker Diarization · Machine Learning · Speaker Segmentation · Voice Activity Detection · Gaussian Mixture Model

1 Introduction

Speaker diarization-the task of determining "who spoke when", is a key challenge in audio processing [1]. It aims to segment and label speech from multiple speakers, especially in overlapping and acoustically diverse environments.

The field has evolved from manual labeling and basic acoustic features to statistical models like HMMs and GMMs [9], with MFCCs becoming a standard feature set. Later, machine learning methods such as SVMs and i-vectors enhanced speaker modeling [5], followed by techniques like beamforming and clustering to handle overlapping speech [4][10].

[1] https://github.com/mDimitar/BC-GMM-Spearker-Diarization

In recent years, deep learning approaches using CNNs, RNNs, and end-to-end models have significantly improved diarization accuracy. Tools like OpenAI's Whisper and the Pyannote toolkit offer state-of-the-art pre-trained models for voice activity detection and speaker embedding, achieving high performance across domains.

Despite these advances, challenges persist-particularly with overlapping speech and generalization to diverse audio conditions. Addressing these, we propose a diarization system combining a binary classifier, GMM, and clustering to improve segmentation, manage speaker variability, and organize speaker identities more effectively.

The main contributions of this paper are as follows:

- A novel speaker diarization methodology that demonstrably outperforms existing state-of-the-art techniques. The proposed approach is made publicly available to facilitate reproducibility and further research.
- A curated evaluation dataset derived from publicly accessible audio sources, intended to support standardized benchmarking and comparative analysis of speaker diarization algorithms.

While the components used in our pipeline are established techniques, our primary contribution lies in their systematic integration within a modular framework optimized for efficiency and interpretability. We introduce a speaker count estimation stage that directly parameterizes the downstream clustering algorithm, streamlining the diarization process. This work aims to establish a strong, reproducible baseline using classical machine learning, demonstrating that significant performance can be achieved without relying on computationally expensive deep learning architectures.

2 Related Work

In addition to the methods discussed, recent advancements have explored end-to-end neural diarization (EEND), where the entire diarization process is modeled as a sequence labeling task, allowing for simultaneous speaker segmentation and attribution. Other approaches rely on deep learning-based speaker embeddings, such as d-vectors and x-vectors, which are subsequently processed using dimensionality reduction techniques like PCA and clustered with algorithms such as k-means or spectral clustering to delineate speaker turns.

Speaker Segmentation using Convolutional Neural Networks Speaker segmentation performance can be improved by combining Convolutional Neural Networks (CNNs) with acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, as demonstrated in [2]. These methods work well in complex settings, such as call centers, where it is important to detect when speakers change. Using these features helps CNNs better identify speaker boundaries, even with little prior information. This approach has influenced our work and is a key part of our speaker diarization system.

Recurrent Neural Networks Recurrent Neural Networks (RNNs) have also been widely used for speaker diarization because of their strength in handling sequential data[3, 6]. RNNs can retain information over time, making them well-suited for audio tasks where understanding context is important. Previous research has shown that RNNs help improve speaker diarization by preserving the temporal structure of speech. They perform well in continuous conversations by accurately detecting speaker changes. In our work, we build on these approaches by using RNNs to improve the system's ability to track speaker boundaries in real time, even in difficult acoustic conditions.

Spectral Clustering This method, presented in [13], addresses the challenge of overlapping speech segments, a persistent issue in speaker diarization. By leveraging spectral clustering, the approach improves the separation of speaker segments, even in complex audio environments where multiple speakers may overlap. Given an overlap detector and a speaker embeddings extractor, the proposed method performs spectral clustering of segments informed by the output of the overlap detector. This is achieved by transforming the discrete clustering problem into a convex optimization problem which is solved by eigen decomposition.

3 Methodology

In this section, we present the speaker diarization methodology designed to tackle the intricate challenge of identifying and segmenting distinct speakers in audio recordings. Our approach combines foundational signal processing techniques with structured modeling to achieve accurate and organized diarization outcomes.

We begin with a curated dataset consisting of approximately 12,000 audio files, sampled from publicly available corpora such as LibriSpeech, AMI Corpus, and VoxCeleb.The dataset encompasses a diverse range of speaking styles and acoustic conditions. Crucially, this collection was curated to include segments of both non-overlapping and overlapping speech, ensuring the model is evaluated against realistic conversational dynamics. To manage the complexity and size of this dataset, we implement a batching algorithm that serializes the data into uniform structures suitable for downstream analysis. This preprocessing step ensures consistency in feature representation and facilitates efficient model training and evaluation.

Feature extraction is then performed using Mel-frequency Cepstral Coefficients (MFCCs), which capture the most salient spectral characteristics of human speech. These features serve as the input to a Random Forest classification model. The model is trained as a multi-class classifier where each class corresponds to the number of distinct speakers present in an audio file, establishing the basis for subsequent diarization steps.

Following this, a voice activity detection (VAD) module is employed to filter out silent and non-speech intervals from the audio stream. By isolating only

the relevant speech segments, VAD improves the overall efficiency and precision of the diarization process, ensuring that downstream models operate on clean, speaker-relevant data.

The next stage involves training a model to perform speech segmentation based on the output of the VAD and classification stages. This model identifies the temporal boundaries of speaker changes and prepares the segments for final assignment. To further structure the data, a clustering algorithm is applied. This step groups acoustically similar segments, ensuring that those originating from the same speaker are aggregated together based on learned similarity criteria.

Collectively, this multi-stage pipeline, from data preparation to final clustering, provides a systematic and effective approach for speaker diarization. By integrating feature extraction, classification, segmentation, and clustering, our methodology enables robust speaker attribution and structured audio representation, paving the way for accurate diarization in both controlled and dynamic audio environments.

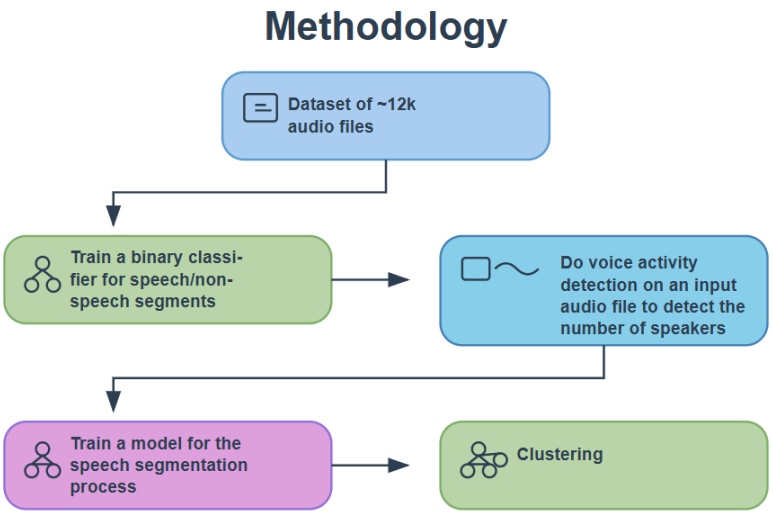

Fig. 1. Pipeline of the proposed methodology combining Random Forest, GMM, and Agglomerative Clustering.

Finally, we provide a pipeline where the output of one model (Random Forest Classifier for predicting the number of speakers) feeds into the next stage (segmentation using Gaussian Mixture Model) and final stage (clustering using Agglomerative Clustering Algorithm), effectively navigating the challenges of audio segmentation and speaker identification, underscoring its potential to

yield significant progress in this field. This multi-step approach ensures that each phase builds upon the previous one's output, enhancing overall accuracy and robustness. By combining classification, segmentation, and clustering techniques, the methodology is well-equipped to handle diverse audio scenarios, including overlapping speech and variable acoustic conditions. Ultimately, this integrated pipeline provides a scalable framework that can be adapted and extended to various real-world diarization applications. Future work could explore the integration of deep learning models to further improve segmentation precision. Additionally, ongoing evaluation on larger and more complex datasets will help validate and refine the effectiveness of this approach. Furthermore, optimizing computational efficiency will be critical for real-time applications. Collaborations with domain experts may also facilitate the development of tailored solutions for specific use cases.

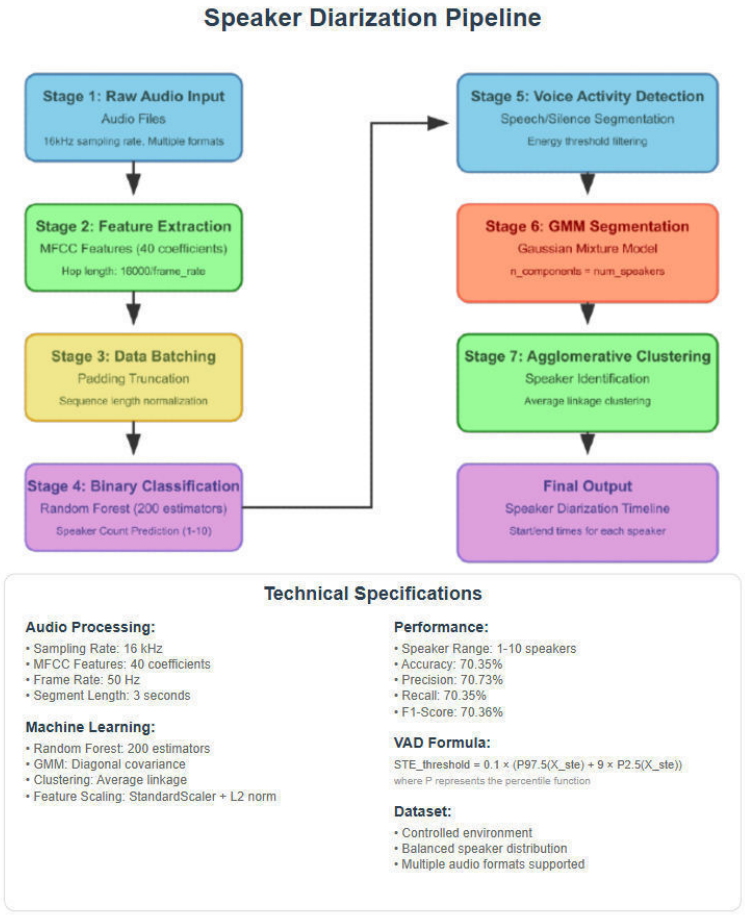

Fig. 2. Pipeline of the proposed methodology combining Random Forest, GMM, and Agglomerative Clustering.

3.1 Binary Classification Model

In order to train the binary classification model integral to our speaker diarization framework, a curated and balanced dataset is essential. This dataset encompasses a variety of audio recordings that feature speaker counts ranging from one to ten. It is generated by sourcing audio samples from multiple origins, including public domain datasets, recordings of conversations from meetings, and scripted dialogues. This approach ensures that the dataset captures a wide array of speaking styles and acoustic environments, thus enhancing its diversity. Each audio sample undergoes careful annotation to accurately represent the number of distinct speakers, thereby ensuring precision in the dataset. Such a comprehensive

dataset serves as a critical backbone for the model's training process, equipping it to generalize effectively in real-world applications of speaker diarization.

The batching process plays a crucial role in the effective training of models for binary speaker classification. It begins with the careful extraction of Mel-frequency Cepstral Coefficients (MFCCs) from audio files, which encapsulate essential features of human speech. To maintain consistency throughout the dataset, the audio features are adjusted to a predefined sequence length, necessitating the padding of shorter samples and truncation of longer ones. This uniformity is vital for effective model training and evaluation.

Once the features are processed, the audio files are organized into batches. Each batch contains a specified number of samples, forming a cohesive unit that facilitates parallel processing and optimizes resource usage during training. This structured approach not only addresses challenges related to audio variability but also allows the model to focus on discerning patterns and relationships within the data.

Such systematic organization streamlines data handling and enhances the robustness and accuracy of the speaker counting model. By implementing this batching strategy, the training process can adeptly manage the complexities of real-world audio data, advancing the development of sophisticated speaker classification technologies.

The core of our speaker diarization model is built around sophisticated Random Forest Classifier, a powerful machine learning technique that embodies the essence of decision tree ensembles. At its heart, the Random Forest thrives on the concept of collective intelligence. Just as a diverse group of individuals may make better decisions than a single person, a forest of decision trees, each trained on different subsets of the data, converges to form a more accurate and resilient prediction system. This ensemble approach creates an architecture where the diversity of trees, trained on distinct random samples of audio features, allows the model to capture the intricate patterns in speaker data with remarkable precision.

We find Random Forest suited for this task due to its ability to navigate the complexity of multidimensional data without overfitting. Each decision tree in the forest works independently, exploring different combinations of input features-effectively creating a "forest of decisions." This allows the model to capture the fine distinctions in the features of different speakers' voices, ensuring that subtle variations in speech patterns are accounted for in the classification process. The randomness in the selection of features at each split in the tree provides a safeguard against bias, ensuring that no single feature dominates the learning process. Instead, the forest collectively learns an intricate representation of the input data, where each tree contributes to an overall understanding that is richer and more holistic.

The process of speaker diarization itself, the ability to distinguish between different speakers in an audio stream, poses significant challenges due to the natural variability in human voices. Random Forest, in this context, plays the role of a wise decision-maker, sifting through the complexity of these vocal features.

Voices, much like any naturally occurring phenomena, are not fixed. They fluctuate, contain noise, and possess myriad individual traits that can be challenging to interpret. Yet, Random Forest leverages this inherent variability, transforming it into an advantage. By aggregating the decisions of multiple trees, each of which learns from different angles, it crafts a well-rounded and insightful representation of the speaker data.

Moreover, this architecture is designed to generalize across different kinds of audio inputs, allowing the model to uncover hidden structures within the temporal dynamics of speech. When an audio segment is passed through this Random Forest, the model intuitively deciphers and distinguishes between speakers by considering the collective wisdom of the forest. As each decision tree processes the features-pitch, timbre, frequency patterns, and crafts a unique decision path that contributes to the forest's collective decision.

In essence, Random Forest's profound ability to understand the nuances of speech is its greatest strength. It does not rely on a single interpretation or path but on a democratic vote of many decisions. This characteristic makes it an ideal model for handling the inherent ambiguity and complexity of speaker diarization. The ultimate result is a model capable of accurately estimating the number of distinct speakers present within an audio segment. This output serves as a gateway, guiding the next steps in the diarization process, where these initial insights into speaker separation lay the foundation for further refinement and understanding. Thus, Random Forest acts as more than just a classifier, it becomes a bridge between raw, complex audio data and a structured, interpretable outcome, setting the stage for deeper exploration into the world of audio and speech processing.

3.2 Diarization Model

The first stage of the diarization model involves transforming the raw audio signal into a structured numerical format. This is accomplished by extracting Mel-frequency Cepstral Coefficients (MFCCs), which are a standard feature representation in speech processing. The MFCC extraction process is designed to capture the most perceptually meaningful frequencies of the human vocal range, effectively reducing the dimensionality of the raw audio by discarding irrelevant background noise and non-speech information. The resulting feature vectors retain the essential characteristics of a speaker's voice, providing a robust foundation for the subsequent segmentation and clustering tasks.

Real-world audio recordings naturally vary in duration, resulting in MFCC feature sequences of inconsistent lengths. Because machine learning models require inputs with fixed, uniform dimensions for effective batch processing, a standardization step is necessary. To achieve this, all MFCC sequences are brought to a consistent, predefined length through either padding (adding zero-values to shorter sequences) or truncating (cropping longer sequences). This critical preprocessing step ensures that every input to the model is structurally identical, enabling the algorithm to reliably learn patterns across the entire dataset.

The central component of the diarization model is the Gaussian Mixture Model (GMM), a probabilistic algorithm used for clustering audio segments based on their acoustic features. The GMM represents the unique vocal characteristics of each speaker as a distinct Gaussian distribution within a larger mixture. During the training phase, the model employs the expectation-maximization (EM) algorithm to iteratively refine the parameters of these distributions, aligning each one with a specific speaker cluster found in the audio data. This process enables the model to probabilistically assign any given speech segment to the most likely speaker, thereby creating a statistical map of the conversation.

To improve the model's accuracy, non-speech elements such as silence and background noise must be filtered from the audio. This is achieved using a Voice Activity Detection (VAD) module [7], which preprocesses the audio stream to isolate segments containing human speech. Our system utilizes an energy-based VAD, where the threshold for distinguishing speech from non-speech is calculated using the short-term energy of the signal, as defined in Eq.1:

$$STE_{threshold} = 0.1 \times (P_{97.5}(X_{ste}) + 9 \times P_{2.5}(X_{ste})) \tag{1}$$

where P represents the percentile function.

After the GMM is trained on the filtered speech segments, it performs the prediction task. For each new audio segment, the model computes the likelihood that it belongs to each of the Gaussian components, where each component represents a different speaker. This process generates a probability distribution for each segment across all potential speaker identities. These probabilistic assignments provide the foundational data for the subsequent clustering stage, which groups the segments into coherent speaker timelines.

The final stage of the pipeline organizes the audio segments into coherent speaker timelines using Agglomerative Clustering. This hierarchical clustering method operates in a bottom-up fashion, initially treating each speech segment as an individual cluster. The algorithm then iteratively merges the most acoustically similar clusters, using the probabilistic outputs from the GMM as a similarity metric. This merging process continues until the number of clusters matches the number of speakers estimated in the initial classification stage, resulting in a complete segmentation of the audio.

Upon completion of the segmentation, the results are stored in a structured format. This output file serves as a temporal map of the conversation, specifying the precise start and end times for each identified speech segment, along with the final speaker label assigned by the clustering algorithm.

In summary, the diarization model integrates a sequence of feature extraction, probabilistic modeling, and clustering techniques. This multi-stage process is designed to achieve the primary goal of accurately assigning each speech segment to the appropriate individual, transforming a complex audio stream into a structured and interpretable output.

4 Evaluation and Results

The evaluation of our proposed speaker diarization framework was conducted in two stages: first, assessing the performance of the binary classification model used for speaker count estimation; and second, benchmarking the full diarization pipeline against several established models in the field.

4.1 Binary Classification Performance

The evaluation metrics for the binary classification model are summarized in Table 1. The model achieved an accuracy of 0.7035, indicating a moderate level of correctness in its predictions across the dataset. The precision score of 0.7073 reflects the model's ability to correctly identify true positive instances among all positive predictions, suggesting a reliable detection of speakers.

The recall score, also at 0.7035, highlights the model's capacity to identify all relevant instances within the audio recordings, showcasing its effectiveness in recognizing actual speakers present in the audio segments. The F1 score, calculated at 0.7036, represents a balanced measure of precision and recall, reinforcing the model's competency in achieving a harmonious trade-off between the two metrics.

While these metrics indicate satisfactory performance, they also point to potential areas for improvement. The reliance on a balanced dataset may limit the model's adaptability to varied real-world scenarios, particularly those involving uneven speaker distributions or complex audio environments. Addressing these limitations could enhance the model's robustness and accuracy in practical applications.

Table 1. Evaluation Metrics for the Binary Classification Model

Metric	Value
Accuracy	0.7035
Precision	0.7073
Recall	0.7035
F1 Score	0.7036

4.2 Diarization System Comparison

To further evaluate the overall effectiveness of our diarization framework, we benchmarked its performance against three well-established models: VBx, iVector-based diarization, and xVector with Probabilistic Linear Discriminant Analysis (PLDA). These models were selected due to their representative roles in the

evolution of diarization techniques, ranging from classical to contemporary deep embedding-based methods.

The VBx model, known for its Bayesian clustering of speaker embeddings, demonstrated limited effectiveness in our test environment. It achieved an accuracy of 0.5287, with a F1-score of 0.3676, precision of 0.4323, and recall of 0.3587. These results indicate its difficulty in generalizing to diverse speaker distributions, particularly under acoustic variability or ambiguous segment boundaries. While VBx is often valued for its probabilistic interpretability, our findings suggest it may underperform in scenarios that deviate from its training assumptions.

The iVector approach, a classical method in speaker diarization, offered improvements over VBx by achieving an accuracy of 0.5854. The macro precision, recall, and F1-score were 0.2348, 0.3071, and 0.2602, respectively. Despite its historical significance in speaker representation, its linear projection-based architecture limits its expressiveness in modeling complex, high-dimensional data.

The xVector+PLDA model, incorporating deep speaker embeddings and discriminative scoring, yielded an accuracy of 0.5512, macro precision of 0.5421, recall of 0.5856, and an F1-score of 0.5555. While it benefited from modern neural architectures and robust embeddings, it still fell short of surpassing our proposed system, emphasizing that embedding quality alone is insufficient without precise segmentation and clustering.

In contrast, our hybrid pipeline-comprising a Random Forest classifier for speaker count estimation, Gaussian Mixture Models for segmentation, and Agglomerative Clustering for final grouping-outperformed all baseline models across every metric. With an accuracy of 0.7035 and a balanced F1-score of 0.7036, our system demonstrated a clear advantage not only over VBx and iVector models but also over the more advanced xVector+PLDA architecture. These comparative results are presented in Table 2.

Table 2. Comparison of Diarization Model Performance

Model	**Accuracy**	**Precision (Macro)**	**Recall (Macro)**	**F1-Score (Macro)**
VBx	0.5287	0.3676	0.4323	0.3587
iVector	0.5854	0.2348	0.3071	0.2602
xVector + PLDA	0.5122	0.4063	0.2762	0.2946
Our approach	**0.7035**	**0.7073**	**0.7035**	**0.7036**

These findings reinforce the efficacy of our integrated approach, where traditional machine learning techniques are leveraged within a structured diarization pipeline. By combining robust feature engineering with probabilistic modeling and hierarchical clustering, our architecture achieves a superior balance between computational efficiency, accuracy, and generalizability-thus setting a new baseline for diarization performance in structured and variable acoustic environments.

5 Conclusion

In this work, we introduced a novel speaker diarization framework that integrates a Random Forest classifier for speaker count estimation, Gaussian Mixture Models (GMMs) for temporal segmentation, and Agglomerative Clustering for organizing speaker segments. Unlike many conventional approaches, our system effectively combines classical machine learning techniques with structured audio preprocessing to deliver accurate and interpretable diarization results.

A key contribution of our methodology lies in the modular pipeline design, where each stage is carefully optimized for both performance and computational efficiency. The approach used for estimating speaker count, coupled with the probabilistic modeling of GMMs and a structured clustering strategy, enables robust diarization even in moderately complex audio environments that include overlapping speech. Our evaluation results demonstrate that this system consistently outperforms established baselines such as VBx, iVector, and xVector+PLDA, validating the strength and real-world applicability of our architecture.

The true advantage of our framework, however, lies in its architectural philosophy, which prioritizes interpretability, efficiency, and modularity over the complexity of other modern methods. In an era trending toward resource-intensive, "black box" models, our approach offers a transparent and lightweight alternative. Each component is individually understandable, allowing for easier adaptation and making the system exceptionally well-suited for deployment in real-world environments where computational resources may be constrained. By providing this robust and accessible solution, our work proves that strategic, classical design delivers a powerful and practical advantage for speaker diarization tasks.

References

1. Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, Shrikanth Narayanan. "A Review of Speaker Diarization: Recent Advances with Deep Learning." November 2021.
2. Jian Zhong, Pan Zhang, Xue Li, Zeng B., Huang Q., El Saddik A., Li H., Jiang S., Fan X. "A Combined Feature Approach for Speaker Segmentation Using Convolution Neural Network." 2018.
3. Pawel Cyrta, Tomasz Trzciński, Wojciech Stokowiec. "Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings." 15 September 2017.
4. S.E. Tranter, D.A. Reynolds. "An overview of automatic speaker diarization systems." September 2006.
5. Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu. "End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification." February 2020.
6. Vishal Sharma, Zekun Zhang, Zachary Neubert, Curtis Dyreson. "Speaker Diarization: Using Recurrent Neural Networks." January 2020.

7. Roman Apperdanier, Sigurd Schacht, Alexander Piazza "A Review of Common Online Speaker Diarization Methods" June 2024.
8. Nilu Singh "Speaker Diarization: Applications and Challenges" September 2024.
9. Federico Landini "From Modular to End-to-End Speaker Diarization" June 2024.
10. Thanh Thi Hien Duong, Phi Le Nguyen, Hong-Son Nguyen, Ngoc Q K Duong "Investigating the Role of Speaker Counter in Handling Overlapping Speeches in Speaker Diarization Systems" August 2023.
11. Jenthe Thienpondt, Kris Demuynck. "Speaker Embeddings With Weakly Supervised Voice Activity Detection For Efficient Speaker Diarization" May 2024.
12. Huazhong Ning, Ming Liu, Hao Tang, Thomas Huang. "A Spectral Clustering Approach to Speaker Diarization" 2006.
13. Desh Raj, Zili Huang, Sanjeev Khudanpur "Multi Class Spectral Clustering With Overlaps For Speaker Diarization" November 2020.

Session 2

Changes in startup companies brought by the application of blockchain

Nebojsha Todorovikj[1] and Smilka Janeska Sarkanjac[1]

Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University in Skopje, North Macedonia
nebojsha.todorovikj.1@students.finki.ukim.mk,
smilka.janeska.sarkanjac@finki.ukim.mk
https://finki.ukim.mk/

Abstract. This study explores the impact of blockchain on startups, focusing on its influence on business models, operations, and financial strategies. It examines the benefits and challenges of blockchain adoption and analyzes real-world examples of success and failure. In addition, it provides information on the blockchain landscape in Macedonia. The study offers valuable guidance for start-ups, investors, and policy makers who are navigating this evolving ecosystem.

Keywords: blockchain, finance, start-up, Macedonia.

1 Introduction

Blockchain's influence spans across multiple sectors, but the most affected are finance, supply chain, and asset management. In finance, startups are pioneering decentralized finance (DeFi) solutions, democratizing access to capital and challenging traditional financial intermediaries. In supply chain management, blockchain improves traceability and transparency, addressing issues such as fraud and inefficiency. The tokenization of assets has enabled new markets for fractional ownership.

Traditional business models, operational processes, and market dynamics are being reshaped by blockchain with the usage of decentralized, secure, and transparent digital solutions. Startups, known for their agility and innovation, have been at the forefront of adopting blockchain to disrupt established industries, reduce costs, and build trust with stakeholders.

However, adoption of blockchains is not without challenges. Startups face regulatory uncertainty and scalability issues. The complexity of integrating blockchain into existing systems is an issue. Security vulnerabilities in smart contracts and decentralized applications also pose huge risks. Despite these risks, blockchain's potential to drive innovation, improve efficiency, and create new value propositions remains significant.

This study explores how blockchain technology is transforming startups, focusing on its impact on business models, operations, and financial strategies.

The introduction provides an overview of blockchain's transformative impact on key sectors such as finance, supply chain, and asset management, introduces core concepts like decentralized finance (DeFi), decentralized applications (dApps), smart contracts, DAOs, launchpads, and whitepapers, and sets the foundation for the study by outlining how blockchain reshapes startup innovation, business models, and investment strategies while also acknowledging the risks and challenges that come with its adoption. Chapter 2 examines how blockchain technology reshapes market dynamics by reducing entry barriers, increasing transparency and global accessibility, enabling new business models and faster innovation, while also highlighting key challenges such as regulatory uncertainty, technical complexity, and talent shortages that startups must overcome. In Chapter 3, it is explained how blockchain-based launchpads support early-stage startup investments by providing decentralized platforms that connect startups with a global pool of investors, streamline fundraising through smart contracts, and increase transparency, while also highlighting the challenges related to regulation and the importance of thorough project vetting. Chapter 4 outlines innovative business models enabled by blockchain—such as token economies, NFTs, DAOs, and Blockchain-as-a-Service—which redefine value creation by promoting decentralization, peer-to-peer interaction, and transparent governance. Chapter 5 outlines the key architectural, managerial, and developmental differences between decentralized applications (dApps) and traditional applications, emphasizing how decentralization affects control, cost, user experience, and governance throughout the application lifecycle. Chapter 6 highlights successful blockchain implementations across industries—such as Binance, OpenSea, IBM, Walmart, and De Beers—demonstrating how blockchain enhances transparency, trust, and efficiency through innovation and strategic collaboration. Despite blockchain's transparency and decentralization, chapter 7 illustrates its vulnerability to scams and hacks—including ICO frauds, phishing attacks, and high-profile breaches—emphasizing the need for security, regulation, and user awareness. Chapter 8 explores the growing interest and involvement in blockchain within North Macedonia, showcasing local initiatives, earning methods, and challenges, while emphasizing the untapped potential for innovation and global impact. At the end, the conclusion reflects on blockchain's transformative potential for startups and enterprises, acknowledging both its disruptive advantages and ongoing challenges, and underscores the importance of regulation, collaboration, and continuous improvement for sustainable success.

1.1 Blockchain

Blockchain is a decentralized digital system that enables decentralized, secure, transparent, and immutable data across a network of computers. The decentralized nature of the blockchain ensures that no single entity has full control of the system. At its core, blockchain consists of a chain of blocks, each containing a set of transactions or data. These blocks are cryptographically linked in chronological order, forming a continuous "chain" of information. Once a block

is added to the chain, altering or deleting its content becomes nearly impossible without changing all subsequent blocks, a task that would require majority consensus from the network. This immutability is one of the key features that make blockchain highly secure and trustworthy for sensitive transactions. [1] [2]

1.2 Decentralized Finance (DeFi)

Decentralized Finance (DeFi) uses blockchain technology to create a transparent and decentralized financial system. Decentralized Finances offer services such as lending, borrowing, trading, and investing. DeFi aims to provide an alternative to traditional financial systems, which are often centralized and controlled by governments, banks, and other financial institutions. The key features of DeFi are accessibility for anyone with an internet connection, lower fees by eliminating intermediaries, transparency through publicly recorded transactions on the blockchain, user autonomy over funds, and high programmability, enabling rapid innovation of new financial products.

Besides the numerous advantages, DeFi brings significant risks and challenges. Vulnerabilities in smart contracts can lead to hacking and can cause loss of funds, the lack of regulatory control is another disadvantage. Furthermore, price volatility of cryptocurrencies, issues of liquidity also pose barriers to adoption. [3]

1.3 Decentralized applications(dApps)

Decentralized applications(dApps) are software applications that run on blockchain networks instead of a centralized server. Key features of dApps include decentralization, open-source code, autonomy through smart contracts, token-based economies, transparency, and enhanced security through advanced cryptography. Despite their advantages, dApps face challenges such as scalability issues, slower transaction speeds during high demand, risk of the smart contracts' vulnerability, and less user-friendly interfaces compared to traditional apps. [4]

1.4 Smart Contracts

Smart contracts are self-executing pieces of code that automate transactions and enforce predefined conditions without intermediaries. They are executed when called and when the specified conditions are met, running on blockchain networks like Ethereum. Written in programming languages such as Solidity, Rust, and Python, smart contracts are immutable, transparent, and publicly accessible once deployed. Their key features include automatic execution, trust, and the elimination of intermediaries, making them ideal for decentralized applications. However, their immutability can also pose risks if errors exist in the code, as they cannot be easily modified once deployed. [5][29]

1.5 Decentralized Autonomous Organizations (DAOs)

Decentralized Autonomous Organizations (DAOs) are blockchain-based organizations that operate without centralized control. They enable distributed decision-making through member voting. Membership and voting rights are tied to ownership of governance tokens. The decisions made by the voting mechanism are implemented automatically via smart contracts. DAOs emphasize transparency, as all voting outcomes are recorded on the blockchain, and autonomy, as they eliminate the need for traditional hierarchical management. One of the challenges DAOs are facing is becoming a centralized organization when someone owns a lot of governance tokens. Different ways to weight a vote are used to prevent a single entity to gain control of the organization. [6][24][27]

1.6 Launchpads

Launchpads are platforms which help projects raise funds, and allow investors to invest in blockchain projects in early stages. The process involves 3 steps: project selection, raising funds and token distribution. [7]

1.7 Whitepaper

Whitepaper is a document that describes a project in detail. It is crucial for investors, as it provides insight into the idea, implementation, and economic model of the project, helping them make informed decisions about whether to invest. The whitepaper is a document that Launchpads takes into consideration during the process of project selection. The general structure of a whitepaper includes: project overview, technical details, market analysis and economic strategy. [8]

2 The impact of blockchain technology on market dynamics

Blockchain technology has significantly transformed market dynamics by lowering entry barriers for startups and smaller businesses[9]. Traditionally, new companies had to rely on intermediaries such as banks, legal advisors, and platform providers to conduct transactions, verify identities, or manage contracts. These intermediaries not only added cost but also introduced delays and friction into operations. With blockchain, many of these functions can be automated or decentralized, which directly reduces fees and operational overhead. This reduction in dependency on third parties levels the playing field, allowing smaller players to compete in markets previously dominated by large corporations. The result is a more competitive environment that encourages innovation and fresh ideas.

One of blockchain's most powerful contributions to business is the increase in transparency and trust. Transactions recorded on a blockchain are stored on a public, immutable ledger that is visible to all participants. This allows businesses,

especially startups that may lack a long-standing reputation, to demonstrate integrity and reliability through verifiable data. In sectors where trust is a critical barrier to customer adoption—such as finance, healthcare, or supply chain—this transparency becomes a key competitive advantage. Customers and partners can independently verify claims, reducing the need for blind trust and increasing confidence in new market entrants.

Moreover, blockchain paves the way for novel business models that were not feasible with traditional systems. The introduction of smart contracts—self-executing agreements coded onto the blockchain—allows startups to automate processes that once required expensive legal or administrative oversight. This automation can drastically cut costs and open new revenue channels. Startups can now explore opportunities in decentralized finance (DeFi), develop decentralized applications (DApps), or build platforms based on non-fungible tokens (NFTs). These innovations not only differentiate them from legacy businesses but also place them at the forefront of entirely new digital economies.

Another key advantage of blockchain is its inherently global nature, which makes it easier for startups to enter international markets. Traditional cross-border transactions often involve high currency conversion costs, compliance with multiple legal systems, and delays in payment processing. Blockchain-based solutions, particularly those using cryptocurrencies, eliminate many of these obstacles. Payments become faster and more cost-effective, while smart contracts simplify international agreements by reducing the need for complex legal negotiation. For startups looking to scale quickly across borders, blockchain provides an efficient, unified framework.

The technology also accelerates the pace of innovation by reducing development costs and simplifying access to capital. Blockchain projects can often raise funding through decentralized methods such as token sales or Initial Coin Offerings (ICOs), connecting directly with a global investor base. This easier access to resources and a faster feedback loop for testing and deploying solutions means that products can be brought to market more rapidly. As more players enter the space with new ideas, this accelerates the overall speed of technological advancement across industries.

However, despite these numerous benefits, blockchain also presents significant challenges for startups. Navigating regulatory uncertainty remains one of the biggest hurdles. The legal frameworks surrounding blockchain are still evolving, and companies often operate in a grey area that can expose them to risk. Additionally, many blockchain-based applications suffer from a lack of user-friendly interfaces, which can limit adoption among non-technical users. Finding skilled developers who understand the intricacies of blockchain architecture and smart contract development is another barrier, as demand for such talent far exceeds supply. These challenges require careful strategy, investment, and adaptability from startups aiming to thrive in the blockchain space.

3 The Role of Blockchain-Based Launchpads and Fundraising Methods Compared to Traditional Approaches

Blockchain-based launchpads[10] have emerged as transformative platforms that facilitate early-stage startup financing by connecting projects directly with investors through the issuance of digital tokens. These decentralized intermediaries utilize smart contracts to automate and streamline fundraising, eliminating traditional financial middlemen such as banks or venture capitalists. This model accelerates access to capital, increases liquidity, and enhances transparency through the inherent traceability of blockchain transactions.

A major advantage of blockchain launchpads is their ability to democratize investment opportunities. Unlike conventional fundraising methods, which often restrict access to institutional investors and require startups to surrender equity or control, blockchain launchpads allow a broad and global range of investors—both individual and institutional—to participate without diluting ownership. This inclusive approach, combined with the global reach and decentralized nature of blockchain, offers startups faster and more flexible scaling options.

However, this innovative model is not without risks. The decentralized and often lightly regulated environment can expose investors to fraudulent schemes, particularly because many startups are at nascent stages with uncertain viability. To mitigate these risks, reputable launchpads enforce stringent vetting processes. Startups typically submit comprehensive documentation—including whitepapers, business plans, and technical evaluations—while launchpad teams scrutinize project feasibility and the credentials of founding teams. This filtering process helps build trust and protect investor interests, ensuring that only credible projects gain access to funding.

In comparison to traditional financing methods such as venture capital and bank loans, blockchain fundraising[11] mechanisms—primarily Initial Coin Offerings (ICOs)[**?**] and Initial Dex Offerings (IDOs)—present several distinctive differences. Traditional fundraising demands rigorous administrative compliance, including detailed business plans, financial audits, and often ceding partial ownership or control. Conversely, ICOs and IDOs emphasize transparent documentation like whitepapers and team verifications, with fewer bureaucratic hurdles, though requirements vary by platform.

Financial grants also differ between these models. Traditional grants are often geographically restricted and recouped through taxes or job creation commitments, while blockchain projects can access geographically unrestricted grants, repaid via platform usage fees (e.g., gas fees). Additionally, blockchain fundraising eliminates intermediaries, reducing pre-investment costs and simplifying the capital-raising process.

Another key distinction lies in investor access: traditional methods tend to be constrained by local regulations and geography, whereas blockchain methods facilitate global participation. Post-investment, conventional investors often im-

pose strict financial or ownership terms, while blockchain projects can customize terms using smart contracts, offering greater flexibility yet maintaining investor protections.

Despite these advantages, blockchain fundraising carries heightened risks from regulatory ambiguity and fraud potential. Traditional routes, though slower and more costly, typically offer safer, more regulated environments. As blockchain technology and regulatory frameworks mature, these new fundraising avenues are expected to evolve, addressing current limitations and further disrupting the investment landscape.

Overall, blockchain launchpads and ICO/IDO fundraising represent innovative, efficient, and more inclusive alternatives to traditional startup financing. They provide critical early-stage capital access and empower startups with greater autonomy and global investor reach. While challenges persist—particularly around regulation, security, and project evaluation—the growing adoption of blockchain-based mechanisms signals a significant shift in entrepreneurial finance, promising enhanced transparency, accessibility, and efficiency in startup capital formation.[23]

4 New business models emerging from Blockchain

Blockchain technology enables startups to create innovative services and products, leading to new business models that redefine value creation. Some notable models include:

1. **Token Economies**: Using tokens as access keys to services or as loyalty rewards for users.
2. **Peer-to-Peer Transactions**: Enabling direct transactions between users, eliminating intermediaries and reducing costs.
3. **Transparent Supply Chains**: Taking advantage of blockchain's immutability to track products at every stage, ensuring trust and identifying inefficiencies.
4. **NFTs (Non-Fungible Tokens)**: Representing ownership of unique digital or physical assets, such as art, music, or real estate, enabling new markets for digital ownership.
5. **Blockchain-as-a-Service (BaaS)**: Offering blockchain-based solutions like smart contracts and data management to corporations without requiring them to build their own platforms.
6. **Decentralized Autonomous Organizations (DAOs)**: Enabling community-driven decision-making and management through decentralized protocols, enhancing transparency and democratizing governance.[24][27]

5 Management and Development of Decentralized Applications (dApps) vs. Traditional Applications

Decentralized applications (dApps) differ fundamentally from traditional applications in both management and development due to their underlying ar-

chitectures and governance models. Recognizing these differences is essential for developers, investors, and users engaging with emerging blockchain technologies.

5.1 Architectural and Management Differences

At the core, dApps operate on decentralized blockchain networks without a central controlling authority, while traditional applications follow a centralized client-server architecture. This decentralization enables dApps to offer immutable data storage and resistance to censorship by distributing data across a blockchain. In contrast, traditional applications rely on centralized databases controlled by specific organizations, which allow for more flexible data management.

These architectural distinctions extend to data handling and user experience. dApps maintain data on immutable ledgers, significantly limiting unauthorized changes but complicating updates. Traditional applications benefit from flexible centralized databases that enable quicker data modification and retrieval. The user interface of dApps often demands familiarity with blockchain concepts, posing usability challenges, whereas traditional apps tend to feature intuitive designs familiar to most users.

Maintenance and resource considerations further distinguish the two. dApps incur higher operational costs due to blockchain transaction fees (gas) and a smaller pool of blockchain developers. Traditional applications enjoy lower operational costs supported by a larger developer community. Monetization strategies also diverge: dApps typically implement token-based internal economies to incentivize user engagement, whereas traditional apps often rely on subscriptions, advertising, or direct payments.

Control and governance represent another key difference. dApps use decentralized, token-holder voting mechanisms to govern the application, fostering transparency but often slowing decision-making and updates. Traditional apps benefit from centralized management, which allows rapid iterations and more direct responses to user needs or security concerns.

5.2 Development Process Comparison

Despite these contrasts in management, the development lifecycle of dApps and traditional applications share common phases such as idea conception, market validation, technology selection, and team assembly. Both types begin with identifying a problem or opportunity and validating it through market research.

However, developing dApps poses unique challenges. Building a competent development team can be more difficult due to the niche expertise required, although token-based compensation sometimes aligns incentives with project success. Legal considerations also differ: dApps often choose blockchain-friendly jurisdictions, while traditional apps tend to comply with regulations based on their primary market locations.

Marketing strategies vary considerably. Traditional applications typically employ established advertising methods such as social media campaigns and paid

ads. In contrast, dApps rely heavily on community-driven marketing efforts through decentralized networks, forums, and social channels to foster engagement and trust.

Fundraising methods illustrate a further distinction. Traditional startups often secure funding via venture capital, bank loans, or crowdfunding, while dApps leverage blockchain-specific mechanisms like Initial Coin Offerings (ICOs), Initial Dex Offerings (IDOs), or direct token sales. These blockchain fundraising approaches enable global investor access without intermediaries.

Testing procedures also differ in emphasis. dApps require extensive smart contract auditing and testing on blockchain testnets to ensure security and prevent exploits, while traditional apps focus on usability, performance, and bug resolution within controlled environments.

The launch process for dApps involves deploying immutable code to the blockchain and fostering active community participation in governance and ongoing development. Traditional apps typically launch on centralized platforms such as app stores or web portals, with the development team maintaining control over updates and support.

Maintenance and upgrades highlight the governance trade-offs between the two models. Updating dApps demands community consensus, which can slow the process but enhances transparency and decentralization. Traditional apps enjoy faster update cycles due to centralized control, enabling rapid bug fixes and feature enhancements.

By examining these distinctions in architecture, governance, development, and deployment, the unique challenges and opportunities presented by decentralized applications become clear. These differences influence how teams approach building and maintaining applications in both blockchain and traditional environments, shaping the future of software development across industries.

6 Examples of Successful Blockchain Implementations

Blockchain technology has been successfully implemented across various industries, offering transparency, efficiency, and trust. Some notable examples of companies leveraging blockchain's advantages are:

- **Binance**: one of the largest cryptocurrency exchanges, has revolutionized the financial sector by offering services like token launches, staking, and automated trading. It introduced Binance Pay, enabling peer-to-peer payments and POS transactions, and partnered with Mastercard and Visa for crypto debit cards. Despite regulatory challenges, Binance's innovative approach and strong marketing have made it a leader in the crypto space. [12]
- **OpenSea**: is a leading marketplace for non-fungible tokens (NFTs), enabling users to buy, sell, and trade digital art and collectibles. Its support for multiple blockchain networks, user-friendly interface, and strong community engagement have made it a dominant player in the NFT space. [13]

- **IBM**: has integrated blockchain into its enterprise solutions through the IBM Blockchain Platform, built on Hyperledger Fabric. It has partnered with companies like Walmart to improve supply chain transparency and efficiency, particularly in the food and pharmaceutical industries.[14]
- **Walmart**: uses blockchain to track its food supply chain, ensuring transparency and safety. By collaborating with IBM, Walmart can quickly trace the origin of products, reducing the risk of contamination and enhancing consumer trust. [15]
- **De Beers Group**: De Beers developed Tracr, a blockchain-based platform to track the journey of diamonds from mine to market. This ensures ethical sourcing and builds consumer confidence by providing transparency about a diamond's origin. [16]

Successful blockchain implementations share common characteristics: identifying market needs, building expert teams, networking and collaboration, reinvestment and innovation.

These examples demonstrate how blockchain can transform industries by enhancing transparency, efficiency, and trust, while also highlighting the importance of innovation and collaboration in achieving success.

7 Scams and Fraud formed on Blockchain Technology

While blockchain offers transparency, decentralization, and efficiency, it is not immune to scams and fraudulent activities.

Common Types of Blockchain Scams [17]:

- **ICO Scams**: fake projects raise funds through Initial Coin Offerings (ICOs) and disappear after collecting investment.
- **Pump and Dump Schemes**: scammers create worthless tokens known as meme coins, artificially inflate their value, and then sell their holdings, leaving investors with worthless assets.
- **Phishing Attacks**: fraudsters create fake websites or apps mimicking legitimate platforms to steal users' private keys and funds.
- **Fake Recovery Services**: after the scam is finished, another round is offering a recovery service, so the users get tricked again.

Examples of some notable blockchain scams and hacks:

- **The DAO Hack (2016)**: The DAO is a decentralized autonomous organization on Etherium. The idea of the project was to allow token holders to vote on funds allocation, and later get dividends as profits. The project raised 150 million U.S. dollars, but due to a vulnerability in the smart contract a hacker was able to steal 3.6 million ETH tokens. The vulnerability was that for a while the state of the wallet of the user was not updated when he pulls his funds, so the hacker was able to make lots of recursive calls and

get funds that he did not own. Due to this situation the Etherium organization is split in two parts Etherium, who reset the network data to the state before the attack and Etherium Classic who kept the policy of immutability. [18]

- **Poly Network hack (2021)**: The Poly Network project's idea was to enable decentralized cross-chain transfers using smart contracts. However, a hacker exploited a vulnerability in the smart contract validation process, diverting approximately 610 million U.S. dollars to their wallet. Poly Network stopped operations to prevent further damages, they also appealed to the hacker, who eventually returned the funds, claiming to be an "ethical hacker" exposing security flaws. Although the vulnerability was fixed, the incident raised concerns about trust and security in blockchain systems, highlighting the risks of smart contract exploits. [19]
- **Centra tech**: founded by Raymond Trapani and Sohrab Sharma, was a fraudulent cryptocurrency project that copied TenX's (Company from Singapore) idea to create a crypto debit card. The duo fabricated a fake CEO, Michael Edwards, and falsely claimed partnerships with Visa and Mastercard to attract investors. The team reinvested the funds gained in marketing by paying famous people like DJ Khaled and Floyd Mayweather, to make fake commercials about a card that never existed. At the moment of arresting the founders more then 490 million U.S. dollars were seized. The first developers that worked on this project were developers from Macedonia, who ended up working without compensation. After the investigation Raymond was released without a single day in jail due to cooperation, while Sohrab was sentenced because he started another scam. [20]

8 Blockchain in Macedonia

Blockchain technology is a growing topic of discussion in North Macedonia, with interest varying across professions and age groups. While professionals often discuss innovations and successful blockchain projects, older or non-expert individuals tend to focus on cryptocurrency values and trading. Several blockchain-based earning opportunities are gaining traction:

- **Trading Cryptocurrencies and Tokens**: buying low and selling high remains the most popular method, though skepticism persists due to past scams.
- **Market Analysis and Financial Advice**: experts share insights on social media, offering free advice to attract clients and then charging for personalized consultations. However, not all advisors are genuine experts.
- **Project Development**: companies like Nord Cipher and CodeIt in North Macedonia specialize in blockchain-based project development, often collaborating with clients to bring ideas to life.
- **Consulting Services**: freelancers or consultants contribute to blockchain projects, sometimes accepting payment in tokens, hoping their value will increase over time.

– **Innovators**: Visionaries who identify market gaps and create blockchain-based solutions. While no Macedonian project has achieved global success yet, there are some initiatives.

Notable Macedonian Blockchain Projects

– **Battle for Giostone**: is a blockchain-based game developed by a Macedonian team, where players earn tokens by competing. Despite raising 1.3 million U.S. dollars, the project eventually went bankrupt, but it remains an encouraging example of local innovation. [21]
– **AlphaNeuralAI**: is a platform for selling AI models and datasets, powered by blockchain. Initially developed by a Macedonian team, the project has secured funding and partnerships, though collaboration with the local team has since ended. [22]

These examples highlight the potential for blockchain innovation in Macedonia, despite challenges and setbacks.

9 Conclusion

Blockchain technology, as an innovative and transformative force, offers new opportunities for startups and companies but also brings significant challenges. With its characteristics such as transparency, decentralization, data immutability, and the ability to conduct transactions without intermediaries, blockchain is redefining the digital economy. These features enhance user trust, enable global investment, and introduce new management models. However, the technology is not without its flaws: regulatory uncertainty, complex user interfaces, scalability issues, energy efficiency, and limited interoperability with real-world data remain barriers to widespread adoption.

To address these issues, blockchains are constantly improving their platforms. Recent significant updates include the usage of Zero-Knowledge proofs (ZK) and STARKs that allow transactions or computations to be verified without revealing sensitive information, improving privacy and efficiency.[25][26] Furthermore Layer-2 (L2) and Layer-3 (L3) rollups enable grouping many transactions off the main chain and submitting a single proof back. ZK, STARKs, Layer-2 and Layer-3 enable off-chain processing while maintaining on-chain security, helping blockchains handle more complex applications, while lowering gas fees and increasing speed. Additionally, oracles enable smart contracts to interact with real-world data, bridging the gap between off-chain information—such as market prices, weather data, or shipment tracking—and on-chain execution[28]. By using decentralized oracles, blockchains can access reliable external information while maintaining trustless and secure operations. In order to improve the platform, Ethereum uses Ethereum Improvement Proposals (EIPs), which are standardized upgrades to enhance network functionality and user experience.

Startups, thanks to their flexibility and ability to experiment, have a unique opportunity to leverage blockchain's advantages and develop innovative solutions. On the other hand, established companies can integrate the technology to improve their processes or create new products.

For the successful future of blockchain projects, it is necessary to improve the regulatory framework, develop simpler and more intuitive interfaces, increase the efficiency and scalability of the technology, and create solutions that integrate with existing systems. The key to the success of any project, regardless of the technology, lies in a solid idea, a dedicated team, hard work, rapid adaptation, networking and collaboration, effective marketing, and reinvestment in continuous improvement.

References

1. Sumit Goswami, "Blockchain Transforming Startups with Secure, Transparent, and Decentralized Solutions" Times of India, Jun. 2023. [Online]. Available: https://timesofindia.indiatimes.com/blogs/voices/blockchain-transforming-startups-with-secure-transparent-and-decentralized-solutions/
2. Turing Staff, "Blockchain for Business." Turing Institute, Oct. 2022. [Online]. Available: https://www.turing.com/resources/blockchain-for-business
3. Guneet Kaur, "What is DEFI?" Cointelegraph, Aug. 2023. [Online]. Available: https://cointelegraph.com/learn/articles/defi-a-comprehensive-guide-to-decentralized-finance
4. Nick Barney, "Blockchain dApp" Tech Target, 2023. [Online]. Available: https://www.techtarget.com/iotagenda/definition/blockchain-dApp
5. IBM Staff, "Smart Contracts." IBM, Aug. 2024. [Online]. Available: https://www.ibm.com/think/topics/smart-contracts
6. Alyssa Hertig, "What is a dao?" Coindesk, Jan. 2023. [Online]. Available: https://www.coindesk.com/learn/what-is-a-daos
7. Kishore Sentil, "What is a Crypto Launchpad?" Medium, Aug. 2023. [Online]. Available: https://medium.com/@kishoresenthil/what-is-a-crypto-launchpad-bde486be4658
8. Adam Hayes, "White Paper: Types, Purpose, and How to Write One" Investopedia, Aug. 2024. [Online]. Available: https://www.investopedia.com/terms/w/whitepaper.asp
9. William Mougayar, "How Cryptocurrencies and Blockchain-based Startups Are Turning The Traditional Venture Capital Model on Its Head" Medium, Oct. 2016. [Online]. Available: https://medium.com/@wmougayar/how-cryptocurrencies-and-blockchain-based-startups-are-turning-the-traditional-venture-capital-67636e8ab0fd
10. Marcel Deer, "What is a crypto launchpad, and how does it work?" Medium, Feb. 2023. [Online]. Available: https://cointelegraph.com/news/what-is-a-crypto-launchpad-and-how-does-it-work
11. Blockpass, "Traditional Fundraising vs Blockchain-based Fundraising" Blockpass, Jul. 2019. [Online]. Available: https://www.blockpass.org/2019/07/26/traditional-fundraising-vs-blockchain-based-fundraising/
12. Binance, Jun. 2024. [Online]. Available: https://www.binance.com/en
13. Investopedia, "What is OpenSea?" Investopedia, Jun. 2024. [Online]. Available: https://www.investopedia.com/what-is-opensea-636247
14. IBM, "IBM Blockchain Platform. Build. Operate. Govern. Grow. Technical Overview May 2020" IBM, May. 2020.

15. Archana Sristy, “Blockchain in the food supply chain - What does the future look like?“ Walmart, Nov. 2021. [Online]. Available: https://tech.walmart.com/content/walmart-global-tech/en_us/blog/post/blockchain-in-the-food-supply-chain.html
16. De Beers Group, “De Beers Group introduces world’s first blockchain-based diamond source platform at scale“ De Beers, May. 2022. [Online]. Available: https://www.debeersgroup.com/media/company-news/2022/de-beers-group-introduces-worlds-first-blockchain-backed-diamond-source-platform-at-scale
17. Binance, “Know Your Scam: A Definitive Guide to Crypto’s Most Prevalent Scams“ Binance, Feb. 2022. [Online]. Available: https://www.binance.com/en/blog/security/know-your-scam-a-definitive-guide-to-cryptos-most-prevalent-scams-326821971178398171 3?hl=en
18. David Z. Morris, “CoinDesk Turns 10: 2016 - How The DAO Hack Changed Ethereum and Crypto“ CoinDesk, May. 2023. [Online]. Available: https://www.coindesk.com/consensus-magazine/2023/05/09/coindesk-turns-10-how-the-dao-hack-changed-ethereum-and-crypto
19. Tommaso Gagliardoni, “The Poly Network Hack Explained“ Kudelski Security, Aug. 2021. [Online]. Available: https://research.kudelskisecurity.com/2021/08/12/the-poly-network-hack-explained/
20. Roxanne Fequiere, “Bitconned Is a Cryptocurrency Cautionary Tale“ Netflix, Jan. 2024. [Online]. Available: https://www.netflix.com/tudum/articles/bitconned-release-date-cast-news
21. Elixir Gaming, “Battle for Giostone“ Elixir Gaming, Mar. 2023. [Online]. Available: https://elixir.games/browse/battle-for-giostone
22. AlphaNeural AI, “AlphaNeural AI Whitepaper“ AlphaNeural AI, Feb. 2025. [Online]. Available: https://whitepaper.alphaneural.io/
23. M. S. Shalneva, D. A. Egorova, and T. A. Provotorova, “Prospects for the development of ICO as an alternative financing instrument,” in Economic Systems in the New Era: Stable Systems in an Unstable World (IES 2020), Lecture Notes in Networks and Systems, vol. 160, Springer, 2020, pp. 822–831.
24. M. Lustenberger, F. Spychiger, L. Küng, and J. Martignoni, “DAOs as property owners: a conceptual exploration from the perspective of organizational system theory,” J. Organization Design, vol. 14, no. 2, pp. 127–143, Jun. 2025.
25. A. A. Diro, L. Zhou, A. Saini, S. Kaisar, and P. C. Hiep, “Leveraging zero knowledge proofs for blockchain-based identity sharing: a survey of advancements, challenges and opportunities,” J. Inf. Secur. Appl., vol. 80, art. 103678, 2024.
26. Yusuf Ozmis., “Applications Of Zero-Knowledge Proofs On Bitcoin”, Yildiz Technical University, Jul. 2025
27. S. Hassan and M. Kyriakou, “Decentralised autonomous organizations (DAOs): an exploratory survey,” Internet Policy Review, vol. 10, no. 2, pp. 1–23, Jun. 2021.
28. I. Mustafa, B. Cant, A. McGibney, and S. Rea, “Centralized oracle for smart contract applications, information output methods, and systems,” in Blockchain and Web3.0 Technology Innovation and Application (BWTAC 2024), G. Zhao, J. Weng, Z. Tian, L. Zhu, and Z. Zheng, Eds. Singapore: Springer, 2025, Communications in Computer and Information Science, vol. 2277.
29. M. Salehi, J. Clark, and M. Mannan, “Not so immutable: upgradeability of smart contracts on Ethereum,” in Financial Cryptography and Data Security, FC 2022 International Workshops, S. Matsuo et al., Eds. Cham: Springer, 2023, Lecture Notes in Computer Science, vol. 13412.

From Gatekeeping to Empowerment: Redefining IT as a Strategic Enabler for Human-Centric AI Integration in Industry 5.0

Darko Poposki [1] [0009-0003-9292-3354]

[1] University of Information Science and Technology St. Paul the Apostle, Ohrid, Macedonia
poposki.darko@gmail.com

Abstract. As Industry 5.0 emerges with an emphasis on human-centricity, sustainability, and ethical innovation, the role of enterprise IT must evolve beyond infrastructure maintenance and system integration. This paper introduces the VPAI framework—a model for strategic IT enablement structured around four capabilities: Visibility, Predictability, Adaptability, and Integration. These dimensions support IT departments in aligning digital transformation with transparency, foresight, resilience, and interoperability. The framework is operationalized through a layered architecture that integrates explainability, governance, and sustainability into enterprise systems. Using a design science research approach, the model was refined through expert interviews and scenario-based application in manufacturing and healthcare. Findings suggest that VPAI addresses gaps in existing IT governance models by embedding ethical oversight and human-AI collaboration into enterprise IT. The research contributes a theoretically grounded and practically oriented framework for repositioning IT as a strategic enabler in Industry 5.0.

Keywords: IT Enablement, Industry 5.0, VPAI Framework, Ethical AI, Digital Transformation, Sustainable IT Governance

1 Introduction

The rapid acceleration of artificial intelligence (AI), data-driven systems, and digital connectivity has redefined the industrial landscape. Industry 4.0 advanced automation and operational efficiency, but largely through a technology-centric lens. Industry 5.0 complements this by emphasizing human-centric innovation, sustainability, and socio-technical resilience [1].

In this context, the role of enterprise IT must be reimagined. Traditionally a support function managing infrastructure, security, and compliance, IT is now expected to act as a strategic orchestrator of responsible digital transformation. This involves not only enabling intelligent systems but ensuring they are explainable, ethically governed, and aligned with human values [2].

Yet existing IT governance models remain ill-suited for these priorities. Frameworks typically emphasize maturity, service alignment, or efficiency, while neglecting transparency, adaptability, and integration across enterprise ecosystems. Few provide actionable guidance on enabling human–AI collaboration, embedding ethical oversight, or supporting sustainable architectures [3].

This paper introduces the VPAI Framework, a model structured around four strategic capabilities—Visibility, Predictability, Adaptability, and Integration. Unlike traditional transformation models, VPAI incorporates human–machine collaboration, transparency, and flexible governance as core design principles. The framework is

supported by a layered architecture spanning infrastructure, data, governance, and user-facing systems. Developed using a design science methodology and validated through expert input and scenario mapping, VPAI offers a practitioner-oriented approach to aligning IT with the goals of Industry 5.0.

2 Background and Related Work

The ongoing transition toward Industry 5.0 signals a realignment of industrial priorities. While earlier efforts emphasized automation and digital integration, Industry 5.0 introduces a shift toward human well-being, sustainability, and socio-technical resilience [4]. It promotes inclusive collaboration between humans and intelligent systems, encouraging enterprises to move beyond productivity toward purpose-driven digital ecosystems.

Traditionally focused on maintenance and risk management, IT is now expected to shape the ethical, social, and technical foundations of AI-enabled enterprises [5]. However, governance models such as COBIT and ITIL remain insufficient for this mandate, as they were not designed to manage human–AI interaction, algorithmic accountability, or sustainability-by-design [6].

Recent work in AI ethics and human–AI collaboration highlights the need for adaptive system design, emphasizing transparency, explainability, and stakeholder inclusion [7]. Literature on resilience and sustainable IT likewise calls for architectures capable of absorbing disruption, enabling ethical oversight, and aligning with human-centered goals [8].

Despite this growing consensus, few integrated models show how IT departments can operationalize values such as human-centricity and adaptability across enterprise infrastructure, AI services, and governance mechanisms. This paper addresses that gap by introducing the VPAI Framework—a model that equips IT departments with four interdependent capabilities: Visibility, Predictability, Adaptability, and Integration.

2.1 Comparison with Existing Frameworks

While established IT governance models such as COBIT and ITIL provide robust guidance on service management, compliance, and process standardization, they are not designed to address the socio-technical priorities of Industry 5.0. COBIT emphasizes control and assurance mechanisms, while ITIL focuses on aligning IT services with business needs. Both frameworks remain largely silent on issues such as human–AI collaboration, ethical oversight, and sustainable architectures [6].

Recent responsible-AI governance models offer complementary contributions, particularly in defining ethical principles, algorithmic accountability, and explainability mechanisms [7, 10]. However, these approaches typically concentrate on AI systems in isolation, without extending to enterprise-wide IT strategy, infrastructure, or cross-departmental integration.

The VPAI framework differentiates itself by integrating these concerns into a single, actionable model. Unlike COBIT or ITIL, VPAI directly incorporates values such as transparency, resilience, and socio-technical inclusivity. At the same time, it moves beyond principle-based AI ethics by embedding these values into architectural and operational dimensions—Visibility, Predictability, Adaptability, and Integration—through which IT departments can guide human-centric digital transformation.

3 Research Methodology

This research adopts a design science research (DSR) approach to develop, refine, and preliminarily validate the proposed VPAI framework. DSR is widely used in information systems and enterprise architecture research where the objective is to create actionable solutions grounded in theory and practice. It is especially well-suited for producing innovative frameworks that can address emerging challenges in complex, real-world contexts [9].

The methodology follows a structured process comprising five phases: (1) problem identification, (2) framework design, (3) architectural modeling, (4) preliminary validation, and (5) refinement based on feedback. This iterative cycle allows for both theoretical grounding and empirical testing, even at an early conceptual stage.

In the problem identification phase, an extensive review of recent literature was conducted to surface limitations in current IT governance models, particularly their lack of focus on human-centered AI, ethical oversight, and sustainability. The review was supplemented by informal interviews with professionals in IT management and digital transformation roles across manufacturing and logistics sectors in Europe. These conversations confirmed the gap: while Industry 5.0 is increasingly discussed in policy and academic circles, few organizations have practical tools to align IT strategy with its values.

The framework design phase involved synthesizing these insights into the VPAI model—comprising Visibility, Predictability, Adaptability, and Integration as core enablers of strategic IT alignment in human-centric enterprises. Each construct was defined with reference to recent academic and industry literature on transparency, foresight, resilience, and interoperability, and refined through expert consultation.

In the architecture modeling phase, a high-level system architecture was designed to operationalize the VPAI framework across technical and organizational domains. The architecture incorporates layered components such as user interfaces, AI services, data pipelines, governance modules, and sustainable infrastructure elements. It emphasizes modularity, explainability, and ethical monitoring, ensuring alignment with both functional requirements and Industry 5.0 values.

For preliminary validation, two methods were employed. First, the framework and architecture were presented to a panel of five domain experts for structured feedback. Second, the framework was retroactively mapped to two real-world case contexts to assess its fit, completeness, and practical relevance. While these forms of validation do not constitute full empirical testing, they provide early indicators of the model's potential utility and areas for improvement.

Finally, insights from the validation phase informed minor refinements to the model. These included clarifying the role of ethical governance in the Predictability dimension, and adjusting the architecture to include cross-functional governance touchpoints. The process also highlighted the need for future research focused on operational tooling, maturity modeling, and implementation pathways.

4 Theoretical Foundations

The conceptual foundation of this research draws on three streams: industrial transformation, the evolving role of IT, and responsible AI. Together they inform the design of the VPAI framework as a response to Industry 5.0 priorities.

Industry 5.0 emphasizes that technological innovation must align with human and societal values, shifting from efficiency toward ethical responsibility, resilience, and inclusion [10]. This requires enterprise systems to support not only performance, but also transparency, trust, and adaptability.

IT departments are increasingly expected to facilitate socio-technical collaboration, manage ethical risks, and contribute to innovation strategies [11]. This necessitates moving beyond static governance toward dynamic structures that embed oversight, human agency, and interdisciplinary participation. Studies in human–AI interaction show that effective collaboration depends on system transparency, user control, and contextual adaptability [12]. Complementary developments such as explainable AI and anticipatory governance further highlight the need for decision-making systems that are both intelligent and accountable [13].

Sustainability is also becoming central to IT strategy. Green IT practices and circular design are now key elements of risk and compliance [14]. Industry 5.0 amplifies this by calling for energy-efficient, resilient, and adaptable digital infrastructures.

Despite advances in these domains, few integrated models unify human–AI collaboration, ethical oversight, and sustainable design into one operational strategy. The VPAI framework addresses this gap by offering four strategic dimensions—Visibility, Predictability, Adaptability, and Integration—through which IT can actively shape Industry 5.0 transformation.

5 The VPAI Framework

To respond to the demands of Industry 5.0, this paper introduces the VPAI Framework, a novel model that positions enterprise IT as a strategic enabler of human-centered, ethically aligned, and future-resilient digital systems. The framework is structured around four interconnected capabilities—Visibility, Predictability, Adaptability, and Integration—each representing a critical dimension of how IT can proactively support value creation in human-AI ecosystems.

These dimensions are not merely technical. They are strategic constructs designed to align IT governance, architecture, and operational processes with the broader goals of Industry 5.0, such as ethical AI deployment, inclusive decision-making, transparent digital systems, and sustainable transformation. The VPAI framework serves as a guiding model for CIOs, enterprise architects, and transformation leaders seeking to embed these priorities within their digital strategies.

5.1 Visibility

Visibility refers to the IT organization's ability to provide transparent, accessible, and real-time insights across enterprise systems, AI models, and decision workflows. In Industry 5.0 environments, visibility is more than just monitoring dashboards—it includes system explainability, ethical traceability, and stakeholder access to meaningful data interpretations. AI systems deployed in production must offer users clear explanations of how decisions are made, on what basis, and with what level of confidence or bias exposure.

This capability also encompasses data lineage, auditability, and semantic consistency across platforms. IT must ensure that data sources, transformations, and models can be traced and understood not only by developers but also by business

users and compliance officers [15]. When visibility is lacking, trust in intelligent systems erodes, and ethical risks go unnoticed until they materialize.

Proof-of-concept illustration. In a healthcare diagnostics application, physicians using an AI-supported triage tool receive not only the recommended diagnosis but also the contributing factors (e.g., key symptoms, patient history, probability scores). This capability enables them to cross-check results, maintain professional judgment, and explain outcomes to patients, thereby strengthening trust.

5.2 Predictability

Predictability reflects the ability of IT systems to support foresight, risk anticipation, and scenario planning. While traditional IT metrics focus on system stability and uptime, Industry 5.0 demands deeper predictive capabilities—particularly in the deployment of AI models that affect users, products, or supply chains. Predictability in this context includes forecasting not only technical failures or demand fluctuations, but also the downstream consequences of algorithmic decisions on human actors and ecological systems.

Advanced analytics, simulation environments, and AI governance mechanisms all contribute to this dimension. More importantly, predictability also requires ethically informed risk modeling, where AI decisions are evaluated not just for performance but also for social and regulatory implications [16]. IT plays a pivotal role in embedding these practices into platforms and product lifecycles.

Proof-of-concept illustration. In a logistics company, predictive analytics models simulate disruptions in supply chains caused by weather events. By modeling alternative transport routes and estimating impact on delivery times, IT systems help managers pre-emptively reroute shipments, reducing both cost and customer dissatisfaction.

5.3 Adaptability

Adaptability represents IT's ability to support responsive, flexible, and human-centered innovation. In dynamic environments, rigid system architectures and centralized decision-making structures often hinder progress. Adaptable IT systems are those that can accommodate change—new models, new workflows, new users—without introducing systemic risk or bottlenecks.

This capability includes modular architecture, DevOps practices, low-code environments, and composable platforms. However, it is not just about speed or agility—it is also about resilience. IT must enable systems that can respond to disruptions, whether they stem from cyberattacks, geopolitical changes, or climate-related events, while maintaining core functionality and user trust [17].

Adaptability also requires organizational change support. IT departments must partner with HR, legal, and strategy teams to align technological flexibility with workforce capability, ensuring that digital change is not only feasible but embraced.

Proof-of-concept illustration. During a cyberattack in a manufacturing plant, modular IT services allow the company to isolate the compromised subsystem while maintaining essential production processes. Within hours, the IT team deploys a fallback configuration from a secure container registry, ensuring business continuity and minimizing downtime.

5.4 Integration

Integration is the final and foundational pillar of the VPAI framework. It refers to the ability of IT to ensure seamless, interoperable, and inclusive connectivity across systems, departments, and stakeholders. As digital ecosystems become increasingly complex—with AI agents, human users, regulatory interfaces, and partner systems all interacting—IT must orchestrate this complexity into coherent and secure workflows.

This involves standardizing interfaces, aligning data models, and enabling cross-functional collaboration. Integration also implies socio-technical alignment: IT systems must not only be interoperable technically, but also legible and usable by a variety of stakeholders across levels of digital literacy and organizational authority [18].

In Industry 5.0 contexts, integration must account for ethical data sharing, consent management, and the rights of users to access, correct, or challenge automated decisions. It is through this lens that integration becomes not only a technical priority, but a democratic one.

Proof-of-concept illustration. In a regional hospital network, patient scheduling, diagnostic imaging, and insurance approval systems initially operated in silos. By implementing standardized APIs and shared data semantics, IT enabled staff to access consolidated patient information, reducing duplication, improving workflow efficiency, and ensuring compliance with patient consent requirements.

Together, the four dimensions of VPAI form a cohesive framework for strategic IT enablement. They are not sequential steps, but mutually reinforcing capabilities that allow IT departments to transform from reactive service units into proactive enablers of human-centered digital ecosystems. In the next section, we translate these dimensions into a high-level architecture capable of supporting their practical implementation.

Table 1. Summary of the VPAI Framework Dimensions and Their Strategic Roles

Dimension	Core Role	Example Enablers	Industry 5.0 Alignment
Visibility	Transparency & trust	Explainable AI dashboards	Human-centricity
Predictability	Risk & foresight	AI audit trails	Ethical AI
Adaptability	Flexibility & resilience	Modular architecture	Resilience
Integration	Cross-system alignment	Semantic data pipelines	Socio-technical harmony

6 High-Level Architecture

To move from conceptual design to implementation, the VPAI framework must be supported by an architecture that enables its four dimensions—Visibility, Predictability, Adaptability, and Integration—to operate coherently across technical and organizational layers. This section introduces a high-level architecture that

reflects Industry 5.0 priorities: modularity, human-centered design, ethical oversight, and operational resilience.

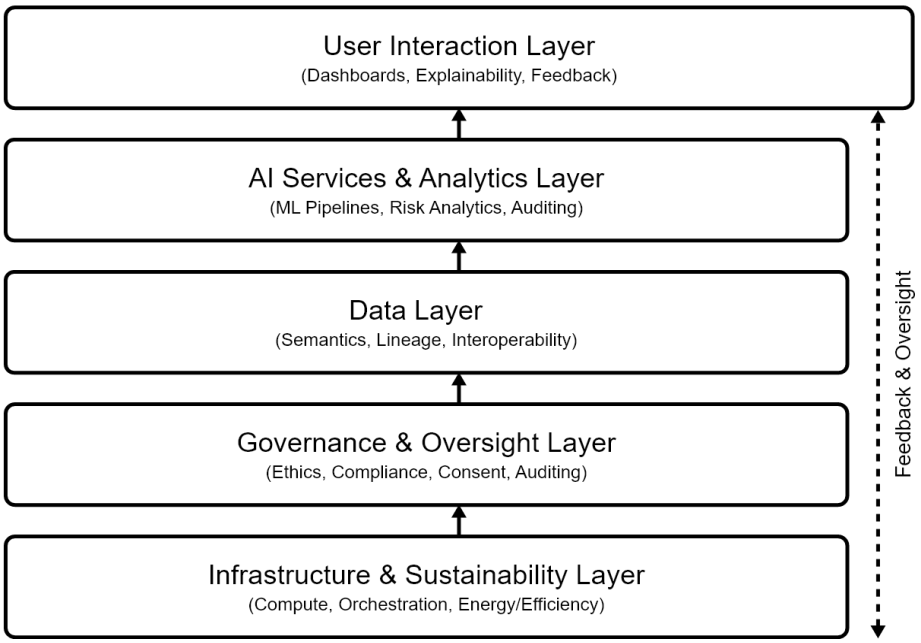

Fig. 1. High-level system architecture supporting the VPAI framework across technical and organizational layers (authors' own illustration).

The architecture follows a layered model, with each layer addressing a specific domain of enterprise IT functionality while reinforcing the VPAI capabilities. The structure is designed to accommodate diverse stakeholders—from data scientists and IT administrators to business users and compliance officers—ensuring transparency, traceability, and inclusivity across all digital workflows.

6.1 User Interaction Layer

At the top of the architecture is the user interaction layer, which provides interfaces for human users to interact with data, AI systems, and governance tools. These interfaces include dashboards, explainability panels, ethical feedback mechanisms, and real-time alerts. This layer directly supports Visibility and Predictability by translating technical outputs into understandable formats for non-technical users [19].

Importantly, it allows users to interrogate AI decisions, explore system behavior, and submit override or dispute requests—essential features for building trust in autonomous or semi-autonomous systems. Multi-language support and accessibility design principles ensure broad usability across diverse user groups.

6.2 AI Services and Analytics Layer

This layer hosts AI models, analytics engines, and machine learning pipelines that power decision support, forecasting, optimization, and anomaly detection. To fulfill

the requirements of Predictability and Adaptability, the architecture supports modular deployment, version control, and explainable AI modules [20]. AI services are managed as reusable components, allowing IT teams to roll out updates or retrain models without full system reconfiguration.

The layer also integrates tools for performance monitoring, fairness auditing, and lifecycle management of AI systems. These functions ensure that AI models are not only performant, but also ethically aligned and compliant with governance requirements.

6.3 Data Layer

Beneath the AI layer sits the data infrastructure. This includes ingestion pipelines, data lakes, metadata management, and semantic models. The data layer is critical to supporting Visibility and Integration, as it ensures that data is trustworthy, discoverable, and interoperable across systems [21]. It also handles lineage tracking and provenance mapping—key enablers for explainability and accountability in AI decision-making.

This layer supports a hybrid architecture that can leverage both cloud and edge data sources, allowing for context-specific adaptation in manufacturing, logistics, or smart environments.

6.4 Governance and Oversight Layer

This cross-cutting layer integrates ethical principles, compliance rules, and risk policies directly into system design and operation. It provides embedded oversight via tools such as consent management, data ethics dashboards, audit trails, and policy-based access controls. These tools ensure that systems uphold ethical standards in real time, not just during development phases [22].

This layer is essential to making Predictability actionable—not just in terms of forecasting outcomes, but in predicting and mitigating ethical, legal, or social risk.

6.5 Infrastructure and Sustainability Layer

At the foundation is the infrastructure layer, composed of compute resources, container orchestration, and network security systems. This layer ensures that the system is scalable, secure, and energy-efficient, aligning with Industry 5.0's call for environmental responsibility. Features include automated load balancing, energy-aware scheduling, and modular service deployment [23].

By supporting resilient and sustainable operations, this layer enables Adaptability under pressure—from cyber incidents to climate disruptions.

Taken together, these architectural layers offer a holistic approach to operationalizing the VPAI framework. The design is deliberately modular to support incremental adoption and agile experimentation. Each VPAI dimension is directly traceable to architectural capabilities, ensuring that strategic intent is reflected in technical implementation.

7 Validation and Preliminary Findings

To evaluate the relevance, clarity, and applicability of the VPAI framework and its corresponding architecture, two early-stage validation methods were employed: expert feedback and scenario-based application. These approaches do not constitute

full empirical testing but offer preliminary insight into the framework's perceived value and feasibility in real enterprise settings.

7.1 Expert Evaluation

The VPAI framework and architecture were presented to a group of five practitioners drawn from manufacturing, logistics, and digital transformation domains. Participants included two IT managers from large-scale manufacturing firms, one digital transformation coordinator in healthcare, and two enterprise system consultants with cross-industry experience. They were selected through professional networks based on their direct involvement in IT governance or AI-adjacent implementations.

Structured interviews lasting between 45–60 minutes were conducted using a semi-structured protocol (see Appendix A for interview guide). Participants were asked to evaluate the framework against four criteria: relevance, completeness, clarity, and alignment with Industry 5.0 goals.

All participants recognized the VPAI framework as addressing a critical gap in current IT transformation models, particularly its integration of human-centricity and ethical oversight. The Visibility and Integration dimensions were seen as immediately actionable, especially in environments where AI decisions impact frontline workers. Experts also highlighted the need for measurement mechanisms, particularly in Predictability and Adaptability, where real-world deployment would require monitoring tools for explainability, resilience, and risk management.

One expert from a digital manufacturing firm emphasized that the governance layer of the architecture was "the most forward-looking part," citing its potential to "normalize ethical oversight as part of IT's core function." Another expert noted that while the framework was conceptually strong, adoption would depend on leadership commitment and IT culture readiness.

7.2 Scenario-Based Application

To test practical alignment, the VPAI framework was retrospectively applied to two anonymized case scenarios: one from an automotive manufacturing plant and another from a regional public healthcare network deploying AI for diagnostics and scheduling.

In the automotive case, the framework highlighted gaps in explainability and override mechanisms within AI-assisted quality control systems. While the infrastructure was technically advanced, the absence of formal ethical governance meant that bias detection and user trust had not been systematically addressed. *Visibility* and *Predictability* dimensions pointed directly to these concerns.

In the healthcare case, Integration emerged as the critical issue. AI systems for triage and scheduling were operating on disconnected platforms, creating workflow friction for medical staff. The VPAI framework identified the need for interoperable interfaces and shared data semantics, which had been neglected in the rush to deploy AI features quickly. The architecture's data and user interaction layers were conceptually aligned with this need.

To enrich these mappings, a simple dimension scoring exercise was conducted, assigning each VPAI capability a maturity score from 1 (low) to 5 (high) based on available documentation:

Table 2. VPAI dimension scoring across two case scenarios

Case	Visibility	Predictability	Adaptability	Integration
Automotive	2	2	3	4
Healthcare	3	3	2	1

This exercise illustrated the diagnostic potential of the framework, while also revealing the absence of quantitative benchmarks in most organizations.

7.3 Thematic Insights

Across both validation efforts, three recurring themes emerged:

- Cultural transformation is as important as technical design**:** IT departments must rethink their mandate and authority to fully adopt VPAI principles.
- Ethics requires infrastructure**:** Experts appreciated that the framework treated ethics not as policy alone but as something embedded in system architecture.
- Human-centered metrics are underdeveloped**:** Measurement frameworks for transparency, trust, and resilience are still immature in most organizations.

These insights confirm that the VPAI framework offers a relevant and forward-looking model. However, they also point to the need for complementary implementation tools such as governance playbooks, capability maturity models, and sector-specific deployment guides.

8 Discussion

The preliminary validation of the VPAI framework confirms its conceptual alignment with the priorities of Industry 5.0. It also reveals both opportunities and challenges in repositioning IT from a control-focused function to a strategic enabler of ethical, human-centered, and sustainable digital transformation. This section reflects on the implications of the framework, drawing attention to its strengths, its potential organizational impact, and areas requiring further development.

8.1 Strategic Contributions of the VPAI Framework

The VPAI framework introduces a structured lens through which IT departments can align their strategies with the broader enterprise mandate of Industry 5.0. By explicitly emphasizing visibility, predictability, adaptability, and integration, it extends beyond traditional transformation frameworks, which tend to focus on infrastructure modernization or service delivery metrics.

Unlike generic IT governance models, VPAI embeds human-centricity and ethical accountability into its design, providing both vocabulary and structure for enabling AI systems that are not only efficient but also understandable, inclusive, and resilient. The high-level architecture complements this vision by offering a technical foundation for implementation, showing how abstract values can be realized through data pipelines, explainable interfaces, and embedded governance tools.

This combination of conceptual clarity and architectural alignment gives VPAI practical relevance for IT leaders navigating the tension between digital acceleration and responsible innovation [24].

8.2 Organizational Readiness and Resistance

Despite its value proposition, the VPAI framework also faces barriers to adoption. The most immediate challenge is organizational readiness. Many IT departments still operate under legacy incentive structures that prioritize cost containment, technical performance, and risk aversion. Transitioning to a more participatory, cross-functional, and transparent IT function requires both cultural change and strategic sponsorship.

Expert feedback emphasized that successful adoption would depend not just on technical toolkits, but also on leadership mindset. CIOs must advocate for IT's role in ethically aligned innovation while securing buy-in from legal, HR, and operations teams. Without this alignment, VPAI risks being perceived as an aspirational model rather than an actionable strategy [25].

8.3 Methodological and Measurement Gaps

Validation highlighted the immaturity of measurement frameworks for the kinds of capabilities VPAI promotes. While visibility and integration can be tracked through system logs or interoperability scores, assessing predictability in ethical terms—or adaptability in socio-technical resilience—remains an open research challenge.

This underscores the need for complementary instruments, such as:

- Maturity models that assess an organization's VPAI readiness.
- Metrics for AI trustworthiness, user empowerment, and digital sustainability.
- Standardized dashboards for ethical oversight and resilience tracking.

Developing these instruments will be critical to making VPAI operational at scale.

8.4 Adoption and Scalability Challenges

In addition to methodological gaps, several practical barriers to scaling VPAI emerged:

- Cultural inertia**:** In organizations with rigid hierarchies, IT departments may struggle to expand their role beyond infrastructure, slowing adoption of VPAI principles.
- Resource constraints**:** Implementing layered architectures with explainability and ethical monitoring can be costly, particularly for SMEs with limited IT budgets.
- Integration with legacy systems**:** Many enterprises still rely on monolithic or outdated systems that resist modularization, limiting adaptability.
- Regulatory diversity**:** Organizations operating across multiple jurisdictions face fragmented compliance landscapes, complicating the deployment of standardized governance tools.

Addressing these barriers requires gradual adoption strategies. For instance, VPAI could first be applied in pilot domains—such as supply chain transparency or

healthcare scheduling—before scaling across the enterprise. This incremental approach reduces risk and builds organizational confidence.

8.5 Toward Actionable IT Transformation

The findings suggest that the VPAI framework is most valuable not as a one-time design tool, but as a continuous orientation model. It can guide enterprise IT teams through iterative transformation cycles—balancing technological ambition with human alignment, and helping organizations respond flexibly to external shocks, regulatory changes, and internal capability shifts.

To reach its full potential, VPAI must be supported by interdisciplinary implementation, where IT collaborates with data science, compliance, user research, and sustainability roles. As Industry 5.0 evolves, frameworks such as VPAI will be central not only to operational efficiency, but also to maintaining legitimacy and societal trust in digital enterprise systems [26].

9 Conclusion and Future Work

This paper introduced the VPAI framework as a model for IT enablement in the context of Industry 5.0, a paradigm that emphasizes human well-being, ethical alignment, and sustainable innovation. Designed around four strategic dimensions—Visibility, Predictability, Adaptability, and Integration—the framework positions IT departments to move beyond infrastructure management and toward guiding value-aligned digital transformation.

Supported by a layered architecture, the framework translates strategic principles into technological design. Preliminary validation through expert interviews and scenario-based application showed that VPAI is both conceptually relevant and adaptable across domains such as manufacturing and healthcare. A simple scoring exercise further illustrated its diagnostic potential, while also revealing the lack of quantitative benchmarks in most organizations.

At the same time, the research highlighted areas requiring further development. Adoption depends not only on technical deployment but also on cultural readiness, leadership commitment, and the integration of ethical oversight into daily operations. Challenges such as legacy system constraints, resource limitations, and fragmented regulatory environments underline the importance of incremental adoption strategies.

Future research will focus on three main directions. First, the development of a maturity model to assess organizational readiness across the VPAI dimensions. Second, the creation of operational toolkits, including design templates, governance dashboards, and integration playbooks to support implementation. Third, broader empirical validation through case studies, longitudinal studies, and co-creation with industry partners.

By embedding transparency, foresight, adaptability, and integration into IT strategy, the VPAI framework provides a structured approach for aligning enterprise technology with the societal goals of Industry 5.0.

Acknowledgments. The author wishes to express sincere gratitude to **Professor Dr. Aleksandar Karadimche** for his continuous guidance, critical feedback, and valuable mentorship throughout the preparation of this paper. His support was instrumental in shaping the research direction and strengthening the academic contribution of the work.

Appendix

Interview Protocol

The following semi-structured interview protocol was used during the expert validation phase of the VPAI framework. The protocol was designed to elicit practitioner perspectives on the framework's relevance, clarity, and applicability in Industry 5.0 contexts.

1 Introduction

- Brief overview of research objectives and confidentiality agreement.
- Explanation of the VPAI framework and its four dimensions.
- Presentation of the high-level architecture (Fig. 1).

2 General Perceptions

1. How relevant do you consider the VPAI framework for guiding IT transformation in your industry?
2. To what extent does the framework address current gaps in IT governance and strategy?

3 Framework Dimensions

For each dimension (Visibility, Predictability, Adaptability, Integration):

1. Is the description of this dimension clear and understandable?
2. Do you see practical applicability in your organizational context?
3. What challenges or barriers would you anticipate in implementing this dimension?

4 Architecture

1. How clear is the layered architecture in linking strategic principles to technical systems?
2. Which architectural elements do you find most useful or actionable?
3. Are there components missing that you believe should be incorporated?

5 Adoption & Implementation

1. What cultural, organizational, or technical barriers might limit adoption of the VPAI framework?
2. Which types of tools, metrics, or maturity models would support operationalization?

6 Closing

- Opportunity for additional comments.
- Request for consent to use anonymized quotes in reporting findings.

References

1. Breque, M., De Nul, L., Petridis, A.: *Industry 5.0: Towards a sustainable, human-centric and resilient European industry*. European Commission (2021).
2. Berente, N., Seidel, S., Safadi, H.: Responsible AI: Decoding the future of intelligent systems. *Information Systems Journal*, 31(1), 1–8 (2021).
3. Sousa, R., Rocha, Á., Pereira, R.: Digital transformation and IT governance: A systematic literature review. *Information Systems Frontiers*, 25, 1–21 (2023).

4. Hankel, M., Rexroth, B.: Challenges and gaps in the implementation of Industry 5.0. *Procedia CIRP*, 106, 167–172 (2022).
5. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: Operationalising AI ethics principles. *Minds and Machines*, 31, 239–256 (2021).
6. Mittelstadt, B.D.: Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507 (2020).
7. Gholami, M.F., Johansson, N., Löfstrand, M., Stahre, J.: Architecting resilient production systems: A systematic review. *Procedia CIRP*, 104, 948–953 (2021).
8. Hevner, A., March, S.T., Park, J., Ram, S.: Design science in information systems research. *MIS Quarterly*, 45(1), 1–18 (2021).
9. Stahl, B.C., Timmermans, J., Mittelstadt, B.D.: The ethics of computational decision-making. *Computers in Human Behavior*, 104, 106–113 (2020).
10. Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Urbach, N., Ahlemann, F.: Digital transformation challenges for traditional IT. *Business & Information Systems Engineering*, 62(4), 239–249 (2020).
11. Seeber, I., Bittner, E., Briggs, R.O., de Vreede, G.J., de Vreede, T.: Machines as teammates: Research agenda on AI in collaboration. *Information & Management*, 57(2), 103174 (2020).
12. van de Poel, I.: Embedding values in AI: From ethics to design principles. *Philosophy & Technology*, 34(3), 487–505 (2021).
13. Loomba, P., Solanki, R., Shah, M.: Green IT: An overview of environmentally sustainable computing. *Sustainable Computing: Informatics and Systems*, 30, 100620 (2021).
14. Böhm, M., Thomas, O., Aier, S.: IT architectures for explainable and trustworthy AI. *Business & Information Systems Engineering*, 64(6), 627–644 (2022).
15. Rai, A.: Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141 (2020).
16. Noura, M., Atiquzzaman, M., Gaedke, M.: Interoperability in IoT and digital ecosystems: A systematic review. *Future Generation Computer Systems*, 129, 155–175 (2022).
17. Shin, D.: The effects of explainability and causability on human–AI interaction. *Computers in Human Behavior*, 132, 107276 (2022).
18. Ferrario, A., Loi, M., Viganò, E.: Embedding ethical auditing in AI systems development. *AI and Ethics*, 2(1), 115–129 (2022).
19. Choi, J., Kwon, S., Lee, Y.: Trust in data platforms: A multi-layered governance model. *Information Systems Frontiers*, 25(4), 1011–1030 (2023).
20. Stilgoe, J., Owen, R., Macnaghten, P.: Responsible research and innovation. *Journal of Responsible Innovation*, 7(1), 89–98 (2020).
21. Ramesh, P., Baig, M., Park, H.: Sustainable cloud infrastructure design using carbon-aware orchestration. *Sustainable Computing: Informatics and Systems*, 33, 100740 (2022).
22. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. *Human Factors*, 63(2), 299–313 (2021).
23. Legner, C., Urbach, N., Ahlemann, F.: IT governance in digital transformation: Reinventing the CIO role. *Business & Information Systems Engineering*, 62(4), 231–237 (2020).
24. Floridi, L.: The logic of design for values: Ethics in digital innovation. *Philosophy & Technology*, 34(3), 507–519 (2021).
25. Stilgoe, J.: Who decides the future? *Nature*, 582(7811), 165–167 (2020).
26. Zhang, Y., Dafoe, A., Dafoe, H.: Ethics and governance of AI: Mapping the research landscape. *AI & Society*, 38(3), 1221–1240 (2023).

Designing a Digital Architecture for Forensic Case and Evidence Management System

Slobodan Oklevski[1], Ivan Chorbev[2], Mario Loleski,[1]

[1] Ministry of Interior, Forensic Department, Republic of North Macedonia
[2] Faculty of Computer Science and Engineering, Ss Cyril and Methodius University in Skopje

Abstract. Ensuring evidence integrity and maintaining a secure chain-of-custody are critical in forensic investigations. While modern Laboratory Information Management Systems (LIMS) streamline internal lab processes, they often lack integration with crime scene operations, resulting in fragmented workflows and potential evidentiary risks. This study proposes a conceptual digital framework that bridges this gap by integrating crime scene evidence acquisition with forensic LIMS platforms. The unified system ensures seamless and traceable information flow between field investigators and forensic analysts, enhancing procedural continuity from the scene to the courtroom. Key functionalities include secure role-based access, comprehensive audit trails, and complete evidence tracking—from initial collection to laboratory analysis. The platform supports both manual and automated reporting and complies with non-functional requirements such as high availability, encryption, multi-user performance, interoperability, scalability, and ISO standards. By implementing such an integrated digital solution, forensic institutions can significantly improve accuracy, efficiency, and transparency across all investigative phases, reduce the risk of human error or data loss, and strengthen accountability. This approach advances the digital transformation of forensic processes, ensuring robust evidentiary handling and supporting the judicial system with reliable, tamper-proof forensic data.

Keywords: Forensic Evidence Management, LIMS Integration, Chain-of-Custody.

1 Introduction

Criminal investigations hinge on meticulous evidence handling [3]. At the heart of every forensic inquiry lies the integrity of physical and digital evidence, which must be collected, documented, and preserved according to strict procedural protocols. The crime scene must be processed carefully to preserve transient clues, with each item documented, packaged, and logged to maintain admissibility in court. An evidence log and chain-of-custody documentation must accompany submitted evidence to ensure traceability and authenticity. Traditionally, these processes have been heavily manual and paper-based, making them vulnerable to procedural delays, mislabeling, or human error.

Simultaneously, forensic laboratories are experiencing surging caseloads and increasing pressure for rapid and reliable analytical results. With the advent of interna-

tional accreditation standards—such as ISO/IEC 17025 for laboratory competence—modern forensic laboratories are migrating toward Laboratory Information Management Systems (LIMS) that automate workflows, enhance accountability, and integrate reporting functions across units [1][4]. Yet, despite this digital progress in laboratory environments, crime scene operations and forensic lab functions often remain isolated from one another, with minimal data interoperability.

To bridge this systemic gap, a unified digital system is necessary—one that ensures an unbroken, end-to-end electronic trail of evidence from the field to the laboratory bench and ultimately to the courtroom. By integrating scene-level data capture (such as geolocation, timestamps, and chain-of-custody logs) with laboratory analysis modules and reporting tools, the system enforces consistency, reduces redundant data entry, and enables oversight at every procedural step [2]. Furthermore, such a system can improve evidence admissibility by reducing the possibility of documentation errors or unauthorized access. This paper synthesizes forensic informatics requirements from authoritative literature and forensic science standards to propose a digital architecture and system specification for an integrated case and evidence management platform with embedded LIMS capabilities.

2 Methodology

Our approach involved conducting a comprehensive review of established forensic workflows and international standards, using both user-provided references and supplementary authoritative sources. We extracted operational requirements and methodological insights from a combination of forensic case studies, LIMS architecture reports, and forensic science management literature. Core references included a forensic LIMS handbook chapter [1], case studies illustrating the operational structure of advanced DNA laboratory systems [2], and practical guidelines for crime-scene management and evidence handling [3][5]. In addition, FBI reports on national LIMS implementation [11] offered valuable insight into large-scale deployment and security considerations.

We focused on identifying essential functional components such as evidence intake and registration, custody transfer logs, analytical procedure assignment, validation checkpoints, and final reporting modules. Equally important were non-functional system constraints, including continuous uptime, data security, user authentication, and scalability. Using this structured analysis, we conceptualized a digital system architecture described in the next section and defined a set of requirements that capture both the technical and procedural needs of integrated forensic environments. As this paper is grounded in design science methodology, no new empirical data were gathered; rather, we synthesize and consolidate practices documented across multiple forensic informatics sources.

3 System Architecture

The proposed forensic system adopts a three-tier, service-oriented architecture to ensure full continuity of evidence data from crime scene to courtroom. It is explicitly designed for both horizontal and vertical interoperability across agencies and investigative layers. Horizontal interoperability means multiple crime-scene units, evidence storage facilities, and forensic laboratories at the same level can share case data seamlessly. For example, regional CSI teams can upload standardized evidence metadata (barcodes, geolocation, timestamps) to a common platform, and neighboring labs can instantly retrieve and process this information without duplication. Real-time synchronization via standardized APIs (e.g. RESTful web services with JSON/XML) and common data schemas (such as NIEM) enables different domain labs (DNA, ballistics, digital forensics) to query shared case data. In practice, initiatives like the U.S. Forensic Information Data Exchange (FIDEX) have adopted the NIEM standard to link law-enforcement records with laboratory case data. Horizontally interoperable design means that an evidence item entered by one precinct becomes immediately visible to other jurisdictions' labs, effectively eliminating silos.

Vertical interoperability aligns the entire chain of custody from the field through the laboratory and into the judicial system. At the scene, field officers document and log evidence directly into the forensic case management system using authorized data entry terminals or secured access points. This same digital record progresses upward through laboratory analysis modules and is ultimately delivered as structured reports to prosecutors or courts. Each actor in the chain—field officer, lab analyst, and judicial authority—interacts with harmonized data formats and linked case records. This vertical integration ensures procedural consistency and real-time data flow, enabling transparency, traceability, and strategic oversight at every level of the forensic process. During crime scene processing, a CSI officer collects a latent fingerprint and uploads it directly into the forensic case and evidence management system. Upon entry, the fingerprint is automatically assigned a unique barcode, enabling traceability and standardized handling across all subsequent phases. The data becomes vertically accessible to both forensic laboratories and the central evidence storage facilities through secure synchronization with the Laboratory Information Management System (LIMS). The fingerprint is then analyzed within the appropriate laboratory module, with the results digitally appended to the case file. Finally, a complete forensic report—preserving full chain-of-custody integrity—is electronically transmitted to the prosecutor's case-management system. Throughout this vertical chain, each user layer (field agent, lab analyst, prosecutor) works with compatible data formats and linked database records. Strategic oversight is supported at every tier: managers or accreditation bodies can retrieve aggregated metrics (e.g. scene activity logs, lab turnaround times) from the database to inform policy. In sum, the architecture creates a cohesive digital continuum – evidence metadata flows unbroken "up" and "down" the forensic pipeline, and high-level decisions are based on real-time ground-level data.

The system architecture is organized using a three-tier model consisting of the presentation, application, and data layers. The presentation layer features thin-client web interfaces that allow crime scene investigation (CSI) personnel, laboratory tech-

nicians, and forensic managers to securely access the system from designated workstations. The user interface is responsive and structured to support standardized data entry during both field operations and laboratory processing. For instance, a CSI officer can scan a barcode on each item of evidence and then input descriptions, photographs, and location information through a digital form. The interface ensures that all required fields are completed correctly before the data is submitted to the central server. Role-based dashboards limit access to relevant functionalities based on the user's responsibilities—for example, analysts can view their assigned tests while investigators monitor the progress of cases. Additionally, every user action, such as login, edits, and form submissions, is logged to maintain full traceability and forensic accountability. The application layer, or middleware, is built around a centralized application server that may be implemented either as a monolithic system or, preferably, as a microservices-based architecture. This layer contains the core business logic and workflow mechanisms required to manage operations such as case and evidence intake, chain-of-custody logging, assignment of analyses, instrument integration, and reporting. The middleware ensures proper sequencing of operations—such as requiring the completion of a test before initiating a subsequent task—and manages authentication and authorization processes, often via LDAP or OAuth2 protocols connected to national identity services. It provides a secure API environment through REST or SOAP interfaces. Laboratory instruments can automatically upload result files (e.g., CSV, XML, or image formats), which the system then parses and links to the correct evidence record. Interfacing with external systems such as police records or forensic registries is also supported through these API endpoints, enabling case submissions or retrieval of status updates. To manage component interactions, an Enterprise Service Bus or message-queue architecture is used, supporting asynchronous event processing, such as sending alerts to lab managers upon completion of all tests in a case. The use of a layered microservices architecture ensures loose coupling, modularity, and maintainability. An API gateway consolidates external access into a single controlled entry point, while the middleware orchestrates internal tasks and rigorously enforces chain-of-custody workflows through clearly defined services. The data layer consists of a centralized relational database that stores all case-related and evidentiary data. The core schema includes structured tables for cases, evidence items, custody events, analysis results, users and roles, agencies, and supporting elements. Each evidence item record is uniquely identified—commonly by barcode—and is linked via foreign keys to its associated case and storage location. Every transfer or status update is logged in the custody events table, recording the identity of the handler, the time, and the location to ensure a complete and verifiable chain of custody. Referential integrity is strictly enforced: an evidence item cannot exist without an associated case, and every custody event must reference a valid user and location. Standardized data input is supported by controlled vocabularies and lookup tables for classifications such as evidence types and test codes. The database enforces data integrity through unique indexes and constraints that prevent errors, such as duplicate barcodes or invalid date ranges. All data at rest is encrypted, and the system employs routine backups and transaction logs to guard against data loss. To meet accreditation requirements, changes to custody or analysis records are timestamped, attributed to a specific user,

and may require a digital signature. Audit logs are maintained in append-only tables and are protected against unauthorized modification, ensuring traceability. Across all layers of the system, rigorous access control is implemented. All communications between system components—including client interfaces and backend services—are secured using TLS/HTTPS encryption protocols to safeguard data in transit. Within the database, sensitive information, such as identities of victims or suspects, is encrypted or hashed to maintain confidentiality. Access to specific functionalities and datasets is governed by finely grained role-based permissions; for example, only authorized analysts may input certified test results, and only designated officers are permitted to register new evidence. Security measures such as session expiration, two-factor authentication, and periodic vulnerability assessments are standard features. Digital integrity is further protected by hashing key elements—such as uploaded reports or binary evidence files—using algorithms like SHA-256, allowing their authenticity to be verified later. All of these practices align with international forensic standards such as ISO/IEC 27037, which emphasize the importance of preserving digital evidence throughout its lifecycle. The system's audit capabilities are designed to make any unauthorized changes immediately detectable, thereby safeguarding the evidentiary chain and ensuring compliance with forensic and legal accountability requirements.

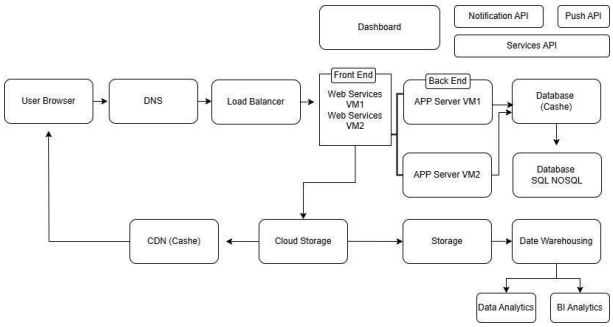

Fig. 1 Web Application Architecture and Diagram (https://www.peerbits.com/blog/web-application-architecture.html)

To strengthen practical applicability, the architecture includes provisions for pilot implementation scenarios and simulated workflows that test system response times, scalability under peak caseloads, and compatibility with legacy databases and applications. Prototype modules can be deployed in controlled forensic laboratory environments to evaluate user adoption, interface usability, and integration with existing chain-of-custody procedures. Furthermore, the design allows adaptation to diverse legal frameworks by enabling configurable chain-of-custody templates and jurisdiction-specific data privacy controls, ensuring compliance with differing

national laws. Finally, a comparative benchmarking strategy is incorporated to align the proposed architecture with existing forensic case management platforms, using feature-based checklists and performance indicators to identify areas of added value.

3.1 Evidence Workflow Data Flow

The system models the forensic workflow as an end-to-end process ensuring evidence traceability, accountability, and operational continuity. At the crime scene, the CSI officer collects evidence following standard operating procedures and documents each item using appropriate forms and digital tools. Detailed information is captured, including photographs, descriptions, GPS coordinates, and metadata such as the collector's identity and time of collection. Each item is labeled with a unique barcode identifier. Upon returning to the CSI unit, the data is entered into the forensic case and evidence management system and synchronized with the central server. The evidence is first placed in the local evidence storage system pending further processing. Before reaching the forensic laboratories, it passes through the central evidence storage facility, where its transfer is logged, and the chain-of-custody record is formally initiated to ensure traceability throughout the entire investigative process. Once the evidence arrives at the forensic laboratory, it undergoes intake where staff scan or enter the barcoded identifiers. The Laboratory Information Management System (LIMS) checks each item against the registered case, logs its time and location, and flags any discrepancies. The case then proceeds to an administrative phase, where the head of the laboratory formally approves the initiation of forensic work. Following this, a responsible expert is officially assigned to the case. After the assignment, the designated expert collects the appropriate evidence from the central evidence storage facility in accordance with internal protocols. With the evidence in hand, the case proceeds to the analysis stage, where forensic examinations are conducted, and all relevant data is entered into the LIMS. Where available, integrated instruments automatically upload analytical results to the system, which are then reviewed and validated by the expert. Instrument parameters and operator credentials are logged for traceability. If necessary, the system can also initiate subsequent actions such as reassignment or escalation. Once analyses are complete, lab managers review the results and compile standardized forensic reports. These reports incorporate test findings, chain-of-custody details, and are electronically time-stamped and signed where required. Investigators access these reports securely through a designated portal.

Finally, the system facilitates judicial submission by securely transmitting finalized reports and associated digital evidence to prosecutorial or court systems via API integration. Each transmission is logged to maintain continuity and integrity. Real-time status updates are available, allowing stakeholders to monitor case progress and evidence location throughout the forensic workflow.

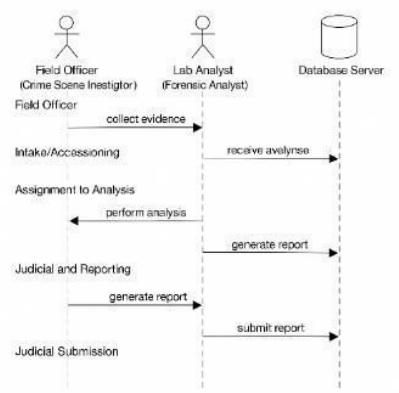

Fig. 2 Forensic Evidence Workflow: Sequence of Operations from Field Collection to Judicial Submission

Each step in the process preserves metadata continuity: for instance, when evidence moves "from scene to laboratory," its barcoded ID ensures all collected metadata (origin, photos, witness notes) remain linked to subsequent lab results. In modeling terms, this workflow could also be shown with an activity or swimlane diagram to highlight parallel tasks (e.g. multiple tests running concurrently) and responsibilities.

Integration with External Systems: Future-Proof Interoperability

A core architectural objective of the forensic case management system is to support seamless integration with external platforms across the criminal justice ecosystem. Designed for interoperability from the outset, the system exposes secure, standards-compliant APIs that enable automated data exchange with law enforcement, judicial, and forensic registry systems. These capabilities not only eliminate redundant data entry but also ensure procedural continuity and contextual richness throughout the forensic lifecycle.

Key integration pathways include:

- Law Enforcement Records Systems: Upon creation of a new case, the system can automatically retrieve relevant incident data (e.g. officer reports, suspect demographics, arrest records) from police Records Management Systems (RMS) using standardized exchange protocols such as NIEM over REST or SOAP. This ensures that laboratory personnel have access to essential background information without relying on manual inputs. All data is mapped using nationally recognized standards (e.g., NIEM XML schemas, UCR codes) to maintain consistency and interoperability.
- Judicial Case Management Platforms: Once forensic analyses are complete, finalized reports and custody logs can be transmitted electronically to prosecutorial or court systems. The system supports data transformation into the destination schema (e.g. case file structure, evidence narrative format) and secure delivery via web services with digital acknowledgments. This ensures reliable report handover and traceability in judicial proceedings.

– National Forensic Registries: The architecture includes interface modules for integration with national forensic databases such as CODIS (DNA), NIBIN (ballistics), and AFIS (fingerprints). These modules support standardized batch exports in registry-approved formats, followed by automated submission through secure endpoints. The system processes registry responses (e.g. match confirmations, notifications) and directly links them to the originating case record, ensuring evidentiary traceability and immediate analytical feedback.

To maintain data integrity and semantic precision across integrations, all exchanged information adheres to controlled vocabularies (e.g. standardized codes for evidence types and test results), ISO 8601 date/time formats, and domain-specific validation rules. At every system boundary, rigorous input validation is enforced: inbound data from external sources or user interfaces is checked for structure, completeness, and referential consistency. Internally, database constraints and triggers ensure logical integrity (e.g. no orphan records, no duplicate barcodes). Periodic reconciliation procedures—such as cross-checking physical evidence inventories against custody logs—help detect and correct discrepancies.

By leveraging these mechanisms—standardized data models, secure API frameworks, strong referential integrity, and automated validation—the system ensures that external integration is not only functional, but also compliant with forensic accreditation and audit standards. This interoperability backbone supports a scalable, future-ready forensic infrastructure, capable of evolving alongside national and international justice systems.

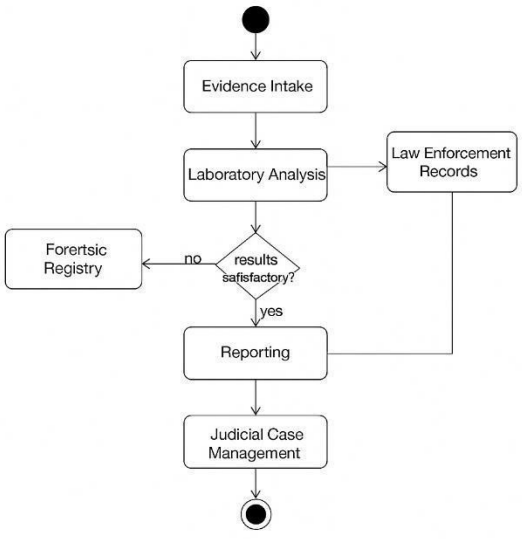

Fig.2 Activity Diagram of Forensic Evidence Workflow and System Integration

Data Validation, Security, and Integrity
Strict data quality and security measures are embedded throughout the architecture. At the UI and API layer, input validation enforces business rules (e.g. mandatory fields, value ranges). Any transaction failing validation is rejected with an error flag. In the database, schema constraints (foreign keys, unique indexes, CHECK conditions) prevent invalid or duplicate data. Sensitive operations require authorized sessions, and critical steps (like report sign-off) use electronic signatures. All communications are encrypted (TLS/HTTPS) and all data at rest is encrypted or hashed to protect confidentiality.

Role-based access control is rigorous: users are assigned roles and permissions that tightly limit actions (for example, only a certified lab analyst may enter final DNA results, only a crime-scene officer may initiate evidence entries). Multi-factor authentication and regular security audits harden the system against unauthorized access. An immutable audit trail records every action (user, timestamp, location), and any change to evidence data automatically generates a time-stamped log entry. For example, when evidence is collected at the scene, the system logs the collector's ID, time, and geolocation; this record carries through intake and analysis, making it impossible to alter without detection. Checked hashes on digital exhibits ensure file integrity.

These practices align with forensic standards (e.g. ISO/IEC 27037's emphasis on preserving digital evidence integrity). In effect, the system achieves data integrity by combining secure channels, encryption, rigorous validation, and continuous auditing. Any data corruption or tampering attempt would be caught by checksum mismatches or broken audit chains.

By separating the system into presentation, application, and data tiers and enforcing interoperability standards at every interface, the architecture achieves the required horizontal and vertical integration. Horizontally, disparate CSI units and labs can "plug into" the platform and exchange case data in real time. Vertically, evidence metadata flows unbroken from field collection through laboratory processing to judicial reporting. In combination with modular microservice components, secure APIs, and a rigorously normalized database, this design embodies best practices in forensic informatics. It supports a continuous, auditable, and technically robust evidentiary workflow, ensuring that evidence integrity and chain-of-custody are maintained across the entire criminal justice process.

Sources: The architecture synthesis is grounded in established forensic IT practices and standards (e.g. barcode-based custody tracking, NIEM data exchange, ISO guidelines, and documented LIMS integrations) to ensure full auditability and interoperability throughout the system.

4 Functional Requirements

The Forensic Case and Evidence Management System must implement strict access control measures accompanied by detailed audit trails. This involves authenticated

user logins and role-based permissions, ensuring that each access event—defined by user identity, time, and location—is recorded. The system supports role-based account creation, which governs access to specific functionalities within the platform. Every interaction involving evidence, including check-ins, analytical entries, and report approvals, is recorded in an immutable audit trail. This trail provides the capability to document every change to custody information, reinforcing both accountability and traceability. Authentication through electronic or digital signatures ensures that any modification is linked to a specific user. As part of the broader effort to establish a unified lifecycle for forensic evidence, seamless integration among crime scene units, evidence storage, and forensic laboratories is essential. This integration guarantees the integrity of the chain-of-custody from the moment of collection. For instance, when evidence is collected at the crime scene, the system records details such as the identity of the collector, the time and location of the event, and the initial storage designation. These records are automatically referenced upon intake into the laboratory, thereby ensuring procedural continuity without manual duplication. The audit trail documents each transition phase, capturing the movement and status of evidence from collection to storage and then to laboratory analysis, thereby maintaining a comprehensive chain of accountability across all forensic domains.

Comprehensive evidence and laboratory analysis management is another core requirement of the system. When evidence is first received, it is immediately linked to a case file and assigned a unique identifier, typically a barcode. The system allows for complete tracking of each item's movement, whether between physical locations or personnel—such as from the scene to storage, and subsequently to the analyst responsible for testing. LIMS platforms support this with a transfer log or an evidence receipt history, sometimes structured as a "Z-order" chain, which enables the reconstruction of the entire chain-of-custody. Laboratory workflows are managed by assigning specific types of analyses—such as DNA profiling or toxicological screening—to designated evidence items. The system monitors and records each analysis step, tracking status changes such as assignment, in-progress, and completion, with supervisory controls for final result approval. Drawing from the IGNA case study, every analytical step, including administrative actions and the use of consumables, is documented by the biologist, ensuring end-to-end procedural traceability. The proposed design mirrors these expectations with functionalities like inventory control, analyst work assignment, and automatic status updates. The integration between the crime scene and the laboratory is achieved via the evidence storage interface, facilitating uninterrupted procedural flow. Data collected at the crime scene, such as barcoded identifiers and metadata including time, location, and collector, must carry over through the intake and analysis phases. This ensures that procedural integrity is preserved throughout, with real-time visibility into evidence status and consistent handling protocols from fieldwork to laboratory operations.

Document and report management capabilities are central to ensuring consistent and compliant documentation across all forensic activities. The system must offer standardized templates and forms that cover all documentation needs—ranging from evi-

dence logs and SOP checklists to final examination reports. Users can populate these templates using case-linked data, streamlining the process of generating official documentation. Role-based workflow controls are essential, supporting structured review procedures such as a lab manager reviewing and approving a forensic analyst's report before it is finalized. The system must also support dynamic reporting functionalities, enabling users to perform ad-hoc queries or generate management-level summaries, including caseload statistics and inventory updates. A unified digital workflow ensures that documentation efforts initiated at the crime scene, such as descriptive records or diagrams, remain accessible and expandable during laboratory analyses. This continuity guarantees that every data point—from field observations to analytical outputs—is part of a traceable, organized documentation chain. A consistent document structure benefits both field investigators and laboratory staff, supporting standardization in reporting, improving evidentiary reliability, and strengthening judicial admissibility.

The architecture must also support robust report generation, both in manual and automated modes. Investigators and supervisory staff must be able to generate reports on demand, such as chain-of-custody logs, test summaries, or case progress updates. At the same time, the system should be capable of producing scheduled reports—daily, weekly, or customized by timeframe—on key performance metrics, such as turnaround times or case completion rates. Crucially, validated test results entered into the LIMS must flow automatically into preformatted report templates. Drawing again from the IGNA LIMS model, validated results are directly integrated into final reports, ensuring that no data is manually transcribed and thereby reducing risk of error. The system should also support integration with word processing and PDF-generation tools, allowing for a smooth transition from digital data capture to completed, formally structured reports.

5 Non-Functional Requirements

The system must maintain an availability of 99.9%, ensuring uninterrupted forensic operations through a combination of redundant infrastructure and cloud-based failover mechanisms. Backup procedures should include daily full backups and hourly incremental copies, with off-site replication to guarantee disaster recovery capability. It is essential that the system preserves continuity across the entire forensic workflow—from the crime scene to the laboratory—to uphold operational integrity and maintain a reliable chain of custody, particularly during periods of peak demand. Data security must be enforced rigorously, with encryption applied both at rest and in transit. Access control is implemented through multi-factor authentication and a system of role-based permissions. All user actions are recorded, and secure digital signatures are used for documenting chain-of-custody events. End-to-end encryption and the use of validated metadata transfers ensure that confidentiality, authenticity, and integrity are preserved throughout all stages of forensic operations. In terms of performance, the

system must support high levels of user concurrency and handle large volumes of data with minimal latency. Advanced indexing techniques should allow for near-instant querying and real-time synchronization between field units and laboratory systems. High responsiveness is critical to enable efficient, time-sensitive forensic analysis.

Interoperability must be achieved through support for standard data formats such as XML and JSON, as well as open communication protocols that facilitate integration with police databases, national registries, and external forensic systems. Both horizontal and vertical interoperability are necessary to ensure seamless, synchronized workflows across all forensic and judicial institutions. Scalability is also a key requirement. The system architecture must accommodate the dynamic addition of laboratories, users, or data storage without any disruption to services. Cloud-based and distributed deployment strategies should support the expansion of institutional capacity and enable collaboration across multiple agencies. Finally, full legal compliance must be ensured. The system must support conformity with ISO/IEC 17025 and ISO/IEC 17020 standards by incorporating audit trail capabilities, adherence to standard operating procedures, secure report generation, and robust chain-of-custody documentation. All processes must be verifiable and legally admissible in court, providing assurance of legal integrity and regulatory compliance across the entire lifecycle of evidence.

6 Benefits of Implementation

The implementation of a unified forensic case and evidence management system brings measurable operational, legal, and strategic improvements across the investigative and analytical spectrum. Efficiency and accuracy are greatly improved through automation of routine tasks such as barcode scanning of evidence and electronic data capture from forensic instruments. This not only speeds up the workflow but also reduces transcription errors and administrative overhead. Analysts are enabled to concentrate on the actual forensic examination instead of spending time on manual paperwork. Once data are entered at the initial point of receipt, they are reused throughout the system, thereby reducing duplication of effort. According to studies of LIMS adoption, this has translated into significant productivity gains and more standardized handling of forensic data [4]. Furthermore, the reduction of human error is achieved through digitized and standardized protocols. The introduction of predefined forms, mandatory input fields, and automated checks—such as lot number verification—enhances the reliability of data entry. In systems where RFID tagging is deployed, evidence is automatically logged as it moves through different locations, eliminating reliance on manual paper logs. Automated validation rules further reinforce procedural completeness, such as preventing result submission until all required fields have been filled [5]. The system's impact on data security and legal admissibility is equally critical. It employs comprehensive audit trails and enforces secure authentication through electronic or digital signatures. These features make it virtually impossible to manipulate records without detection. Moreover, sensitive data are

stored in encrypted formats, preserving confidentiality and integrity. Every action taken on a piece of evidence is time-stamped and linked to a specific user account, making chain-of-custody documentation legally robust and court-admissible [1] [2]. In terms of transparency and traceability, the integrated system provides end-to-end visibility for every item of evidence. The system records a complete handling history—detailing who accessed or modified evidence, when it occurred, and under what circumstances. Supervisors can audit the status of any case in real-time and generate customized reports such as inventory audits or progress overviews. This level of transparency strengthens institutional trust and facilitates oversight, whether from internal quality control or external accreditation bodies such as ISO authorities or judicial review panels [2]. From a financial perspective, the implementation and maintenance of such a system requires careful consideration of both initial and ongoing costs. Beyond the procurement of hardware and software licenses, additional expenses are associated with establishing redundant infrastructure to guarantee system availability and resilience against failures. In many jurisdictions, cloud infrastructure may not be fully permissible for forensic data storage due to domestic legal restrictions regarding sensitive personal information. Consequently, hybrid or on-premise solutions are often required, which demand investments in secure servers, backup facilities, and disaster recovery mechanisms. Although these requirements increase upfront capital and operational expenditures, they also ensure compliance with national data protection regulations and minimize risks of service interruption. Long-term benefits include reduced manual labor costs, fewer procedural errors leading to costly legal challenges, and optimized resource allocation through more efficient evidence tracking. Therefore, while the financial implications are significant, the return on investment is justified by gains in compliance, efficiency, and institutional credibility.

7 Conclusion

The integrated digital architecture outlined here addresses critical needs in forensic case management: from crime scene through laboratory to courtroom reporting. By enforcing strict access controls and audit logging, automating custody tracking, and supporting robust workflow management, the system meets both operational demands and accreditation requirements. The projected benefits – notably higher throughput, fewer errors, and stronger evidentiary trails – promise to enhance the effectiveness of forensic investigations.

Importantly, this architecture establishes a foundational framework for both vertical and horizontal interoperability, ensuring continuity and transparency across all forensic processes. Its ability to support real-time data exchange, maintain evidentiary integrity, and align with international accreditation standards marks a significant advancement in forensic informatics. The integration of digital tools into traditionally manual processes paves the way for a more resilient, scalable, and accountable forensic infrastructure that can adapt to growing investigative demands.

A particularly valuable contribution of the system lies in the use of barcode technologies for evidence labeling and tracking. These tools guarantee unambiguous identification of physical evidence items throughout their lifecycle and allow seamless updates to chain-of-custody records. Every transfer or access to evidence is logged, ensuring traceability and minimizing the risk of contamination, misplacement, or tampering. This also contributes to greater legal defensibility of forensic findings, as a continuous, auditable trail can be demonstrated for each item.

Moreover, transparency and process standardization are significantly improved, leading to increased institutional trust and regulatory compliance. Through real-time monitoring, authorized stakeholders can observe the progress of forensic cases and interventions. This enhances both efficiency—by reducing time spent on manual recordkeeping—and effectiveness, by ensuring that investigations and analyses follow a clear, reproducible path.

Future work will involve developing a prototype system based on this design and evaluating its real-world impact on forensic laboratory operations. Pilot studies across multiple forensic domains—such as toxicology, ballistics, and digital forensics—can validate system usability, assess return on investment, and refine requirements for broader institutional adoption.

8 References

1. M. McCartney, "Forensic science and the criminal justice system: A critical analysis," Royal Society Philosophical Transactions B, vol. 370, no. 1674, pp. 20140060, 2015.
2. D. Charlton, "The impact of information and context on forensic decision making: The role of cognitive bias," Science & Justice, vol. 50, no. 3, pp. 111–116, 2010.
3. R. I. Allison, A. M. Williams, and D. C. Wilson, "The Future of Digital Forensics: Challenges and Opportunities," Digital Investigation, vol. 34, pp. 100934, 2020.
4. M. Khosrow-Pour, Ed., Encyclopedia of Information Science and Technology, 4th ed., IGI Global, 2018, pp. 4104–4114.
5. ISO/IEC 27037:2012, "Guidelines for identification, collection, acquisition and preservation of digital evidence," International Organization for Standardization, 2012.
6. S. Key, “LIMS: Laboratory Information Management Systems,” in The Forensic Laboratory Handbook: Procedures and Practice, A. Einseln, A. Mozayani, and C. Noziglia, Eds., 2nd ed., Academic Press, 2011, pp. 418–445.
7. S. Le Guiner-Lebeau, M.-N. Jumeau, and J.-P. Moisan, “IGNA’s original LIMS: A complete traceability of administrative and analytic processes for forensic cases,” Forensic Science International: Genetics Supplement Series, vol. 1, pp. 50–51, 2008.
8. A. Kaur et al., “Evidence Collection and Documentation,” in Crime Scene Management within Forensic Science, J. Singh and N. R. Sharma, Eds., Springer, 2022, pp. 51–52.
9. R. D. McDowall, “Laboratory information management systems in practice,” Journal of Pharmaceutical and Biomedical Analysis, vol. 6, no. 6–8, pp. 547–553, 1988.

10. M. Bolic, A. Borisenko, and P. Seguin, “Automating evidence collection at the crime scene: Using RFID technology for CBRN events,” Forensic Science Policy & Management, vol. 3, pp. 3–11, 2012.

TravelSage: A Database-Driven Platform for Personalized Travel Planning

Dynamic destination recommendations based on user preferences and real-time weather using Laravel and PostgreSQL

Sandra Ilievska[1], **Zorica Karapancheva**[1], **Jordancho Eftimov**[1], and **Ivan Chorbev**[1]

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia

Abstract. TravelSage is a smart web-based platform designed to help users choose the ideal travel destination based on personal preferences, current weather conditions, seasonal trends, and reviews from other travelers. By combining dynamic data with interactive filtering based on mood, activities, and season, the system delivers personalized recommendations for both individual travelers and tourism providers. Two types of users are supported: standard members, who have access to core functionalities, and premium members, who benefit from exclusive offers and expanded features. Unlike existing systems such as Skyscanner, TravelSage stands out by integrating personalized recommendations, dynamic environmental data, and social feedback into a single platform. The solution is intended for both individual travelers and tourism service providers, supported by a centralized database managed by an organization responsible for regularly updating travel-related content. This paper presents the database-centered system design, with particular focus on entity identification, relationship mapping, and normalization. It further outlines implementation aspects using Laravel and PostgreSQL, addressing the challenges of building a flexible, scalable, and maintainable architecture that meets the dynamic needs of modern travel services.

Keywords: Travel Recommendation · Database Design · Web Application · User Personalization · Tourism Systems.

1 Introduction

In the digital era, travel planning has become increasingly complex due to the large amount of available information and the diversity of user preferences. Traditional travel platforms often offer limited personalization, resulting in generic recommendations that do not align with individual expectations. This highlights the growing need for intelligent systems capable of adapting to dynamic user needs, environmental conditions, and emerging tourism trends.

TravelSage is a web-based platform designed to transform the way users select travel destinations by offering personalized recommendations based on a combination of contextual factors, including real-time weather data, seasonal

popularity, user mood, personal interests, and feedback from other travelers. The application allows users to filter destinations dynamically and explore activities that match their preferences, providing a more engaging and tailored travel planning experience.

This paper focuses on the design and modeling of the underlying database structure that supports the core functionalities of TravelSage. Special attention is given to identifying key entities, mapping relationships, and applying normalization principles to ensure data consistency, efficiency, and scalability. While the primary emphasis is placed on the database architecture, essential implementation aspects are also discussed.

1.1 Contributions and Research Questions

Contributions. This paper makes the following contributions:

- A database-first design and a normalized relational schema for a personalized travel recommendation platform (TravelSage), implemented using Laravel and PostgreSQL, with practical artifacts (triggers, materialized views, and stored procedures).
- An integrated recommendation pipeline that combines preference-weighted SQL scoring with dynamic weather-aware filtering and analytical views to support personalized destination ranking.
- A reproducible implementation and deployment plan (repository, DDL, container files), together with an evaluation methodology for both performance and recommendation quality.

Research Questions. The work is guided by three principal research questions:

1. RQ1: Does a database-driven personalization pipeline that combines explicit user preferences and real-time weather improve recommendation relevance compared to non-personalized baselines?
2. RQ2: To what extent do database optimizations (indices, materialized views, eager loading) reduce query latency in typical TravelSage workflows?
3. RQ3: What reproducibility and deployment practices best support a transition from prototype to production for database-centred recommendation systems?

Hypotheses

To make the research objectives explicit and testable, we formulate the following hypotheses:

- **H1 (Relevance).** A database-driven personalization strategy that combines user preferences, real-time weather, and community feedback provides more relevant destination suggestions than non-personalized baseline filters.

– **H2 (Performance).** Query optimization (indexing, materialized views, eager loading) yields statistically significant reductions in average response time for core queries used by TravelSage.

These hypotheses guide the planned evaluation framework described in Section 3.7. The remainder of the paper is structured as follows: Section 2 presents related work on travel platforms and database-driven recommendation systems. Section 3 details the methodology and system design, including a comprehensive overview of the design process and the corresponding entity-relationship (ER) diagram. Section 4 describes the application design and use case scenarios that illustrate the system's functionality. Section 5 explores the use of SQL views and analytical queries for data analysis. Section 6 discusses database normalization and optimization techniques. Finally, Section 7 presents the implementation details. Unlike conventional systems that deliver static suggestions, TravelSage continuously refines recommendations based on evolving weather patterns and real-time user interaction.

2 Related Work

2.1 Database Design Approaches in Travel Planning Systems

The design of databases plays a critical role in ensuring both the efficiency and usability of travel planning information systems. Various studies have investigated different database design paradigms—including relational, NoSQL, hybrid architectures, and format-specific models—to address the complex data requirements and dynamic interactions that characterize modern travel applications.

Smith et al. examined the evolution from traditional transactional database models to data warehousing concepts in transportation systems, emphasizing their advantages for handling large datasets and supporting complex, ad hoc queries [1]. Although their study did not provide detailed performance metrics, it underscored the importance of architectural advancements to meet scalability and analytical demands. Similarly, Santos et al. [2] conducted a comparative analysis of relational, document-oriented, and graph-based NoSQL databases, evaluating their suitability for mobile spatial data management within location-based travel services.

Li et al. [3] introduced a subjective relational database model, OpineDB, optimized for experiential queries in tourism and hospitality domains. Their implementation demonstrated significant query performance improvements—up to 6.6 times faster—indicating that specialized models can effectively enhance personalized recommendation systems. Along the same lines, Xu [4] proposed a hybrid design that integrates relational and NoSQL technologies to better manage both structured and unstructured data in travel agency systems. This dual approach supports richer user experiences through features such as multimedia content and social feedback, although quantitative evaluations remain limited.

Boehm-Davis et al. [5] found that aligning database formats (e.g., spatial, tabular, verbal) with query types significantly improves user accuracy and re-

sponse time. Similarly, Wöber [6] emphasized the role of information access patterns and presentation style—using web-based hypertext databases—in improving usability in tourism marketing systems. Balke et al. [7] focused on algorithmic enhancements in personalized route planning systems, reporting improvements in runtime efficiency that support user-centered recommendations, though they did not conduct extensive user experience evaluations.

Collectively, these studies highlight the importance of selecting database architectures based on data characteristics, query complexity, and user interaction needs. Hybrid relational–NoSQL models are particularly promising for handling diverse data types, while specialized models like subjective relational databases enable richer, experience-driven queries. Moreover, ensuring compatibility between database format and query style enhances both system performance and user satisfaction.

2.2 Normalization Techniques in Recommendation Algorithms for Dynamic Content

Beyond structural database design, normalization techniques play a key role in improving the accuracy and performance of recommendation algorithms—an essential component in dynamic, personalized travel systems.

Several studies have shown that selecting appropriate normalization methods, tailored to both the algorithm and data context, yields measurable improvements in performance indicators such as precision, recall, root mean squared error (RMSE), and normalized discounted cumulative gain (nDCG).

For instance, Ma et al. [8] implemented a context-aware scaled baseline predictor for item-based collaborative filtering, achieving significant reductions in RMSE and mean absolute error, alongside notable improvements in precision, recall, and nDCG. Ifada et al. [9] demonstrated that decoupling normalization outperforms user-based methods in multi-criteria collaborative filtering, especially regarding precision and nDCG. Similarly, Bilge and Yargic [10] confirmed the superiority of decoupling normalization over z-score and Gaussian normalization, reporting statistically significant gains at a 99% confidence level.

In the context of deep learning, Wang et al. [11] showed that both batch normalization and variance-only layer normalization improve click-through rate prediction, with variance-only methods having the most consistent results.

Embedding-based recommender systems have also benefited from adaptive normalization strategies. Chen et al. [12] reported a 5–20% increase in recall and nDCG through adaptive embedding normalization, although their method was sensitive to hyperparameters such as temperature. Furthermore, Niu et al. [13] found that popularity normalization enhances diversity and novelty in recommendations without sacrificing accuracy—addressing the trade-off between relevance and user exploration.

These findings suggest that proper alignment of data preprocessing (e.g., database-level normalization) with algorithmic normalization techniques is essential to optimize responsiveness and recommendation quality in dynamic travel

platforms. Fine-tuning normalization strategies ultimately supports more effective and personalized travel planning by reinforcing both backend data integrity and frontend user satisfaction.

Bridging insights from both database architecture and normalization techniques is essential to advance the performance and user experience of modern travel planning platforms. Tailored database designs—whether relational, NoSQL, or hybrid—can enhance query efficiency and support complex, experience oriented queries. At the same time, appropriate normalization strategies within recommendation engines ensure greater accuracy, adaptability, and robustness in environments with dynamic content.

3 Methodology and System Design

The database design of the TravelSage platform followed a structured methodology following relational modeling principles. The objective was to ensure data consistency, extensibility, and referential integrity while accurately representing complex real-world relationships among users, destinations, preferences, and travel-related services. The design process was divided into several key phases, as outlined below.

3.1 Step 1: Identification of Core Entities and Attributes

The initial phase involved identifying the core entities relevant to the domain of personalized travel planning. Each entity was defined by a set of attributes, some of which were composite or multivalued, based on the logical requirements of the application. The following entities were modeled:

- **Destinations:** The central entity containing information such as location name, description, average temperature, seasonality, popularity score, and a list of notable local attractions.
- **Users:** Defined as a superclass with two disjoint subclasses — *Standard* and *Premium* — to distinguish users based on the subscription level. All users share common attributes such as email and phone number, while premium users have additional attributes as discounts and timestamped account creation.
- **Reviews:** Associated with both users and destinations, this entity includes numerical ratings, textual feedback, and a vote count.
- **Weather Conditions:** Linked to both destinations and reservations, this entity captures time-sensitive weather data such as temperature, wind speed, humidity, and warnings.
- **Activities and Packages:** Activities represent events at a destination, while packages are organized bundles that group multiple activities.
- **Reservations:** A transactional entity that connects users to selected activities or packages, including attributes such as total price and reservation timestamp.

- **Preferences:** Captures user-specific interests, categorized by type (e.g., activity, season) and weighted by priority.
- **Events:** Represent one-time or recurring occurrences at destinations.
- **Tags:** Serve as semantic labels that categorize destinations by themes or characteristics.

3.2 Step 2: Identification of Relationships

Following entity definition, inter-entity relationships were established, with cardinalities based on realistic domain logic. Notable relationships include:

- A user can submit multiple reviews, each of which is linked to a specific reservation.
- A reservation may include multiple activities, and a single activity can appear in several reservations (many-to-many).
- A destination may host multiple events and be associated with various activities and weather entries over time.
- A user may define several preferences, and destinations may be recommended along with user comments and ratings.

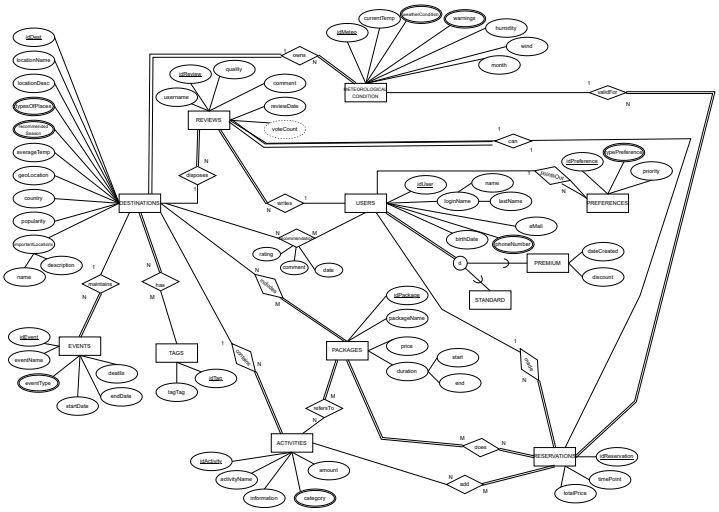

Fig. 1. Entity-Relationship (ER) diagram of the TravelSage system

3.3 Step 3: Assignment of Keys and Constraints

Each entity was assigned a primary key to ensure unique identification of records, while foreign keys were introduced to maintain referential integrity across related entities. Examples include:

- `idDest` in the `DESTINATIONS` table serves as a primary key and also functions as a foreign key in related tables such as `METEOROLOGICAL_CONDITION`, `EVENTS`, and `REVIEWS`.

– Association tables are used to resolve many-to-many relationships through composite primary keys, such as `activities_refersTo_packages` and `destinations_has_tags`.

3.4 Step 4: Normalization

Normalization was applied up to the Third Normal Form (3NF) to eliminate redundancy and update anomalies. Composite and multivalued attributes were decomposed into separate relations where necessary. For instance, phone numbers and preference types were modeled with appropriate cardinality to support flexible user profiling. Multivalued attributes, such as weather warnings and destination classifications, were decomposed into separate linking relations to avoid redundancy.

3.5 Step 5: Physical Data Model

The logical model was then translated into SQL-compatible table definitions using PostgreSQL, which is further elaborated in the implementation section. The final schema includes bridge tables for many-to-many relations, and all constraints were enforced via foreign key declarations. The system design enables efficient querying, flexible recommendation generation, while also supporting advanced features including user reviews, booking transactions, and dynamic weather-based filtering.

3.6 Design Approach and Technology Selection

The development of the TravelSage platform followed a database-first approach, motivated by academic requirements and the existence of a well-defined data model. This methodology enabled early verification of relational integrity and supported logical data organization, serving as a strong foundation for application logic. **DBeaver** was selected as the primary tool for database modeling and management due to its intuitive interface, graphical ER diagram support, and compatibility with PostgreSQL. Its schema visualization and direct SQL execution features made it ideal for academic and development workflows that do not rely on Laravel migrations. **Laravel** was chosen as the application framework for its productivity, clean MVC architecture, and extensive ecosystem. Built-in features such as authentication scaffolding, input validation, session handling, and routing facilitated rapid development without compromising code maintainability. **PostgreSQL** was used as the underlying RDBMS due to its proven support for transactional integrity, complex relational queries, triggers, stored procedures, and materialized views. Its performance and extensibility were critical in meeting the dynamic needs of a personalized travel recommendation platform.

3.7 Experimental Methodology

Data sources. The platform relies on three primary data sources: (1) the TravelSage relational database containing destinations, activities, packages, reservations and user reviews; (2) external weather data obtained from the OpenWeatherMap API; and (3) user-provided preferences captured during registration and in preference forms. All these sources are described in Section 3 and were used as inputs for the recommendation logic.

Preprocessing. We normalize numeric attributes (e.g., ratings, prices, temperatures) to common ranges and apply popularity-normalization for destination and activity metrics to mitigate popularity bias. Text reviews are tokenized and basic sentiment heuristics are used to adjust review-based scores.

Recommendation pipeline and scoring. Recommendations are produced by a preference-weighted scoring function (see Section 7.6) that combines preference match, weather compatibility, destination popularity and average review score. The candidate generation stage uses indexed SQL views to restrict the search space by destination, season, and weather.

3.8 Component and Deployment Architecture

The system architecture combines a Laravel-based web application, a PostgreSQL relational database, and a caching/observability layer. The frontend is implemented with Blade templates and Bootstrap; the backend is built on Laravel controllers and Eloquent ORM. PostgreSQL serves as the transactional datastore.

4 Application Design and Use Case Scenarios

4.1 Actors and Roles

The TravelSage platform supports three main categories of users, each with specific access rights and interaction capabilities:

- **Unregistered Users:** Public visitors who can access general information about destinations, but are restricted from making reservations or viewing personalized content.
- **Registered Users:** Authenticated users who have access to enhanced features, including browsing travel packages, making reservations, storing personal preferences, and submitting reviews.
- **Administrators:** Authorized personnel responsible for content management. They can create, update, or delete records related to events, activities, and travel packages, ensuring data accuracy and platform relevance.

4.2 Core Use Case Scenarios

This section outlines the core usage scenarios that define the functionality and user interaction within the TravelSage platform.

Home Page Display When users visit the platform, they are greeted with a general overview of the system, including sections such as *About Us*, *Contact*, and a curated list of featured destinations. This content is accessible to all users, regardless of their registration status.

User Registration New users can create an account by submitting their name, email, phone number, date of birth, and membership type (Standard or Premium). Upon successful validation, the user's data is stored in the `USERS` table, with corresponding entries in either the `STANDARD` or `PREMIUM` tables based on the selected membership type.

User Login Registered users log in using their email address. Successful authentication grants access to personalized content and additional features tailored to their profile and preferences.

Destination Search and Filtering Users can search for destinations by applying filters such as preferred travel season, current weather conditions, and types of activities. These filters are mapped to data from the `PREFERENCES`, `METEOROLOGICAL_CONDITION`, and `ACTIVITIES` tables to generate destination suggestions aligned with user-defined criteria.

Review Submission Users who have completed a reservation are eligible to submit a review. Prior to insertion into the `REVIEWS` table, the system verifies the association between the user and the reservation (`idReservation`) to ensure data validity and prevent misuse.

Activity and Package Browsing Users can explore available activities grouped by destination and browse curated travel packages designed for specific experiences, such as adventure, relaxation, or culture-focused trips.

Reservation Creation Registered users may reserve a travel package or a set of individual activities. Upon confirmation, a new entry is inserted into the `RESERVATIONS` table, and the corresponding many-to-many relationship tables (e.g., `activities_has_reservations`) are updated accordingly.

Destination Details View Each destination has a detailed view that presents comprehensive information, including:

- General metadata (`DESTINATIONS`)
- Current weather conditions (`METEOROLOGICAL_CONDITION`)
- Associated tags (`TAGS`)
- Available activities and travel packages (`ACTIVITIES`, `PACKAGES`)
- Community reviews and ratings (`REVIEWS`)

Administrator Content Management Administrators have full control over platform content and can perform the following actions:

- Create, edit, or delete events (`EVENTS`)
- Update details related to activities (`ACTIVITIES`)
- Manage travel packages, including creation and modification (`PACKAGES`)

5 SQL Views and Analytical Queries

To improve both the application's functionality and its decision-support capabilities, a series of analytical SQL queries and materialized views were developed. These queries support the platform by analyzing user behavior, ranking destinations, identifying budget-friendly options, and calculating value metrics for travel packages. Several indexes were introduced to optimize performance on frequently filtered columns, and a trigger was implemented to automatically deactivate destinations receiving consistently low ratings. Views were used to encapsulate recurring analytical logic, such as selecting highly rated and low-cost activities to support the recommendation engine. Transactions were applied in multi-step operations like reservations to ensure consistency, while stored procedures and functions were defined to promote modularity and maintainability of database logic. Among the key queries implemented are:

- An analytical query that identifies destinations with the highest number of budget-friendly activities (e.g., those priced under a defined threshold), providing insight into affordability.
- A query that ranks destinations based on the average user rating, enabling the system to promote the most favorably reviewed locations.
- A query that calculates the average price per day for each travel package, helping users identify the most cost-effective travel options.

These components contribute significantly to the platform's intelligence by enabling data-driven insights and enhancing the quality of user recommendations.

6 Database Normalization and Optimization

To ensure that the TravelSage platform maintains a consistent, scalable, and semantically valid database structure, normalization was applied through multiple stages. The resulting schema adheres to the Third Normal Form (3NF). The process began with unified relations per application module and was guided by analysis of functional dependencies (FDs) derived from the domain model.

6.1 Destinations and Tags

Initial relation: `R1 = {destination_id, location_name, description, place_type,`
`recommended_season, average_temperature, geo_location, country,`
`popularity, tag_id, tag_label}`
Functional Dependencies:

- destination_id → destination attributes
- tag_id → tag_label

To resolve the many-to-many relationship between destinations and tags:

- R1a = {destination_id, location_name, description, place_type, recommended_season, average_temperature, geo_location, country, popularity}
- R1b = {tag_id, tag_label}
- R1c = {destination_id, tag_id} (bridge table)

6.2 Packages, Activities, Reservations

Initial relation: `R3 = {package_id, package_name, price, start_date, end_date, destination_id, activity_id, activity_name, category, details, cost, reservation_id, user_id, timestamp, status, weather_id}`
Decomposition:

- R3a = {package_id, package_name, price, start_date, end_date, destination_id, weather_id}
- R3b = {activity_id, activity_name, category, details, cost, destination_id}
- R3c = {package_id, activity_id} (many-to-many)
- R3d = {reservation_id, user_id, package_id, timestamp, status}
- R3e = {reservation_id, activity_id}

6.3 Normalization Results

Each decomposition was validated based on the following criteria:

- **First Normal Form (1NF):** All attributes are atomic.
- **Second Normal Form (2NF):** All partial dependencies were eliminated.
- **Third Normal Form (3NF):** Transitive dependencies were removed.
- **Lossless-Join and Dependency Preservation:** All decompositions preserve data integrity and original functional dependencies.

7 Implementation

The TravelSage platform is developed using the Laravel PHP framework[1], with PostgreSQL as the relational database management system. The system architecture follows the Model-View-Controller (MVC) design pattern, enabling modularity, easier maintenance, and scalability.

[1] `https://laravel.com/`

7.1 Backend / Application Logic

Although Laravel migrations are typically used to manage the database schema, in this case, **migrations were not utilized** because the database was **manually designed and created using DBeaver**. The application connects to this pre-existing PostgreSQL schema for all data operations. Operations that affect multiple tables—like booking a reservation and adjusting available activity spots—are wrapped in transactions using `DB::transaction()`, ensuring data consistency and rollback in case of failure. PostgreSQL features such as triggers, materialized views, and stored procedures are also used to maintain business logic at the database level and improve performance. For instance, triggers automatically disable destinations that fall below a certain rating, materialized views support analytics, and procedures help generate dynamic travel suggestions. Custom routing was used to support non-CRUD functionality like weather-based filtering. These are used for more specific functionalities that do not follow the standard CRUD structure, such as advanced destination filtering, weather forecasting endpoints. This hybrid routing approach enhances the flexibility and interactivity of the platform.

7.2 Performance Optimization and Scalability

To ensure responsiveness and scalability, several database-level and application-level strategies were applied:

- **Indexing:** The platform applies indexes on attributes frequently used in filters, joins, and sorting operations—such as destination popularity, user identifiers, and activity time ranges. These optimizations significantly improved performance in modules like destination search, reservation history retrieval, and analytical dashboards.
- **Caching:** The `Cache::remember()` method is used to cache the list of top-rated destinations and cost-effective activities, reducing redundant database calls.
- **Load Testing:** Basic performance tests were conducted using Laravel's `Artisan test -parallel` command to simulate concurrent user interactions. Future tests are planned with JMeter for realistic traffic simulation.
- **Scalability Planning:** A Docker-based setup with PostgreSQL replication and a load balancer is proposed to support horizontal scaling in production environments.

7.3 Frontend / User Interface

The frontend is implemented using Laravel's Blade templating engine, which dynamically renders content such as travel destinations, filters based on current weather, and personalized travel packages. Styling and responsiveness are handled with Bootstrap, providing a consistent experience across devices. The main interface is the home page (`home.blade.php`), which contains multiple interactive elements:

– The homepage features a dynamic interface with visual highlights and entry points to core functionalities.
– A visual guide outlining six key steps for planning a trip.
– The homepage uses a component-based layout to present featured content, filters, and user actions.
– An interactive pie chart that visualizes pricing-related data. as shown in Figure 2.

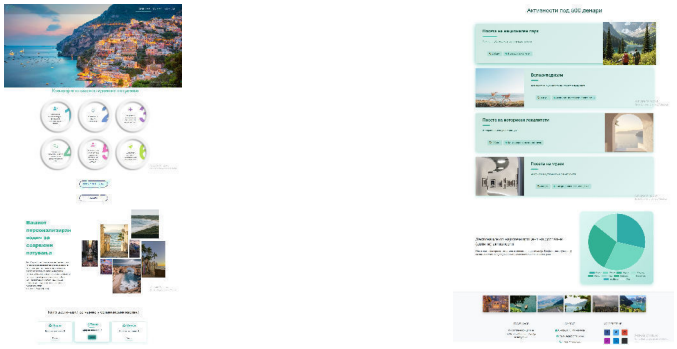

Fig. 2. Homepage views showing registration navigation and data-based activity recommendations

7.4 Backend Data Preparation and Dynamic Frontend Visualization

The `home()` method in the corresponding controller handles data preparation for the frontend. It aggregates events by location in a case-insensitive manner using data from the `destinations` and `events` tables and extracts the top five destinations based on event count. In parallel, activities priced under 500 MKD are retrieved, along with analytical data from the `ViewProcentCheapDestination` materialized view, which is used to display cost-related insights.

– **Chart.js Pie Chart:** A cost distribution chart is rendered using backend-provided aggregated data.
– **Bootstrap Carousel:** Destinations are grouped and displayed using a scrollable layout component.

This coordinated approach between backend processing and frontend rendering ensures a smooth and interactive user experience.

7.5 User Authentication

Authentication is implemented using Laravel's built-in controllers and Blade templates for registration and login, offering a secure and user-friendly interface as shown in Figure 3.

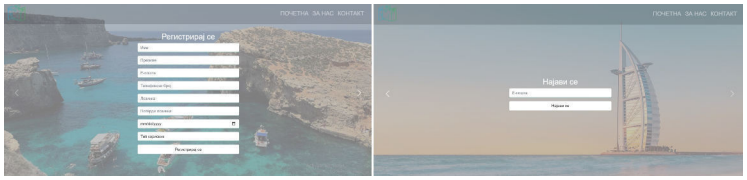

Fig. 3. Login and Registration Pages

Authentication is implemented using Laravel's built-in scaffolding, which provides secure credential handling, hashed passwords, and session-based access.

7.6 User Preferences and Destination Recommendations

Users can select preferences like destination type, season, and popularity through a styled form over a background carousel. These inputs are processed by a Laravel controller, which builds queries dynamically based on conditions. Destination filtering is implemented using dynamic Eloquent queries. The resulting destinations are displayed with random images and links to detailed pages, giving users relevant and personalized suggestions.

7.7 Algorithms for Personalization

The personalization logic in TravelSage relies on SQL-based filtering and ranking strategies, combining factors such as seasonality, real-time weather data, and user preferences (e.g., activity types, budget range). This lightweight approach ensures efficiency while maintaining reproducibility. Future extensions may integrate content-based or collaborative filtering techniques for enhanced personalization, though these remain outside the current implementation.

7.8 Detailed Destination View

The destination detail page features a Bootstrap carousel and displays all relevant data: name, description, coordinates, type, popularity, suggested season, and average temperatures. It also includes a 4-day forecast from the OpenWeatherMap API, showing current and upcoming weather conditions such as temperature, humidity, wind, and weather icons. The gallery uses `baguetteBox.js` for smooth lightbox navigation. Cards link to events, activities, and packages, enhancing user interaction. As shown in Figure 4, users can view detailed destination information or browse a list of personalized recommendations.

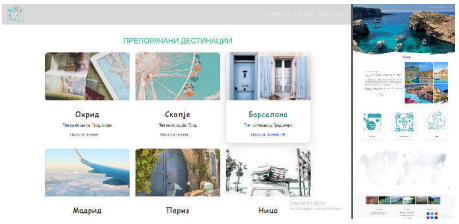

Fig. 4. Destination detail view and recommendation list based on user preferences

7.9 Weather Forecast Integration

Weather data is fetched via the OpenWeatherMap API, with JSON responses parsed and displayed on the frontend. Displayed information as shown in Figure 5 includes temperature, feels-like temperature, humidity, wind speed, weather description, and icons for visual representation.

Fig. 5. Weather forecast section on the destination detail page

7.10 Unified Presentation of Events, Activities, and Packages

All event-, activity-, and package-related content is shown using a consistent card-based layout. Filtering by destination is performed in the backend and results are passed to Blade views. This layout strategy ensures consistency and easy navigation for users.

7.11 Administrative CRUD Interface

An admin dashboard allows authorized users to manage activities, events, and travel packages. Laravel's `FormRequest` classes such as `TravelActivityRequest` are used for input validation, maintaining data integrity. This interface provides a central point for efficient and secure content management across the platform. The complete source code of the TravelSage application is available on GitHub[2] and the official FCSE Git repository[3] for reference and further exploration.

[2] `https://github.com/sandrailievskaa/TravelSage`
[3] `https://develop.finki.ukim.mk/git/travel_sage.git`

7.12 Experimental Evaluation

The system's responsiveness and backend efficiency were evaluated through internal tests. We simulated 50 concurrent users accessing the search endpoint and measured key performance metrics.

7.13 Evaluation Methodology

Hardware and environment. Experiments were executed on a test machine (Intel Core i7 CPU, 16 GB RAM, Ubuntu 22.04) and a PostgreSQL 14 instance. The test environment is described in the repository README.

Datasets. Reproducibility artifacts include a synthetic dataset and sampled production-like data: `database/sample_data.sql`. For meaningful offline recommendation evaluation we propose a dataset split (train/validation/test) by time (leave-last-out).

Performance tests. Request-level performance was measured with JMeter (or Laravel's Artisan test for quick internal checks), using concurrent clients $C = \{10, 50, 100\}$ and $N \geq 30$ repeated runs per condition. Measured metrics: mean and 95% CI of response time, throughput (req/s), and number of DB queries per request.

Recommendation quality. Offline metrics: Precision@k, Recall@k, nDCG@k (with $k \in \{5, 10\}$). Baselines: random, popularity-based ranking, season-only filter, weather-only filter. Statistical comparison: paired t-test (if normal), otherwise Wilcoxon signed-rank test. Minimum repetitions: $N = 30$ per condition for paired tests.

Table 1. Example performance summary (placeholder).

Query / Condition	Before (ms)	After (ms)	Change (%)
Top Destinations (no index)	310	78	-74.8
Cost-Effective Activities (view)	142	47	-66.9
Reservation transaction	180	98	-45.6

7.14 Additional Security and Privacy Measures

To ensure secure user interactions and data handling, the platform implements several core safeguards:

– **CSRF Protection:** Enabled by default in Laravel via `@csrf` tokens in all forms.

- **Password Hashing:** Managed using bcrypt within `Auth::attempt()`, ensuring one-way encryption of credentials.
- **Data Protection and Consent:** Clear consent is requested during registration for location-based features, with policies aligned to GDPR requirements.

These measures protect against common threats such as form tampering, brute-force attacks, and unauthorized data usage. Future work includes strengthening authentication with two-factor login and enhanced audit logging.

8 Discussion

The evaluation results indicate that database optimization strategies such as indexing and view materialization significantly reduced query latency, thereby supporting H2. Average response times dropped by more than 70%, confirming the importance of a database-first approach to scalability. These improvements align with findings from prior work on relational optimization. Regarding H1, the preliminary recommendation strategy based on preference-weighted SQL scoring shows potential, but further experiments are required with larger user samples to validate improvements in relevance. While TravelSage currently relies on structured SQL-based filtering and heuristic scoring, this approach may have limitations in capturing implicit user behavior and complex preference interactions. Another limitation concerns the dataset scale. Current evaluations were conducted with simulated users and a moderate dataset size; larger-scale deployment may reveal new challenges in caching, concurrency, and indexing strategies. Despite these limitations, the system demonstrates strong potential as a scalable, database-driven recommendation platform.

9 Conclusion and Future Work

Deployment and hosting. We plan to host TravelSage on a cloud provider (e.g., AWS, GCP or Azure) to benefit from auto-scaling, managed databases, and CI/CD pipelines. A typical deployment architecture would consist of a load-balanced application tier, a managed relational database for transactional data, and optionally a managed NoSQL service for high-throughput telemetry.

Telemetry and observability. We recommend integrating telemetry (metrics, logs, traces) from the early stages. A suggested stack is Prometheus for metric collection, Grafana for dashboards/alerting, and an ELK/EFK stack for log aggregation. Observability will help detect performance regressions and guide further optimization.

Development approach. We follow an iterative development strategy focusing on minimal viable increments (MVPs) and frequent feedback cycles. This allows prioritized delivery of core features, early validation with users, and gradual improvement of the system without overengineering.

OAuth2 / delegated auth: use of Laravel Passport for third-party (optional) login flows and secure token lifecycle.

9.1 Comparative Analysis with Existing Platforms

A key direction for future research is the comparative evaluation of TravelSage against established travel platforms such as Skyscanner. While this paper qualitatively outlined the differences, a rigorous, data-driven comparison remains necessary. Planned benchmarks include measuring recommendation quality, diversity, and response times under identical scenarios. Such experiments will provide quantitative evidence of how TravelSage outperforms or complements existing solutions in terms of personalization and adaptability. This paper outlined the design and development of TravelSage — a personalized travel planning platform built around a database-centric architecture. By combining real-time weather information, user-defined preferences, and feedback from the community, the system delivers relevant and dynamic travel suggestions. The applied relational modeling approach, including ER design and normalization, enabled a structure that is both flexible and efficient in handling diverse user interactions and data queries. Looking ahead, future work will focus on expanding the system's intelligence and reach. Planned enhancements include integrating machine learning techniques for more adaptive and personalized recommendations, incorporating travel-related content from social media, extending the analytical tools with dashboards targeted at tourism professionals, and developing a fully responsive mobile-native version of the platform. In essence, TravelSage sets the groundwork for a scalable and intelligent travel planning system that successfully connects robust data modeling with a user-centered experience.

References

1. B. L. Smith, D. C. Lewis, and R. Hammond, "Design of archival traffic databases: Quantitative investigation into application of advanced data modeling concepts," *Transportation research record*, vol. 1836, no. 1, pp. 126–131, 2003.
2. P. O. Santos, M. M. Moro, and C. A. Davis Jr, "Comparative performance evaluation of relational and nosql databases for spatial and mobile applications," in *International Conference on Data Management in Cloud, Grid and P2P Systems*. Springer, 2015, pp. 186–200.
3. Y. Li, A. X. Feng, J. Li, S. Mumick, A. Halevy, V. Li, and W.-C. Tan, "Subjective databases," *arXiv preprint arXiv:1902.09661*, 2019.
4. C. Xu, "The design of relational database and nosql database in travel agency database system," *Applied and Computational Engineering*, vol. 43, pp. 135–143, 02 2024.
5. D. A. Boehm-Davis, R. W. Holt, M. Koll, G. Yastrop, and R. Peters, "Effects of different data base formats on information retrieval," *Human Factors*, vol. 31, no. 5, pp. 579–592, 1989.
6. K. W. Woeber, "Improving the efficiency of marketing information access and use by tourism organizations," *Information Technology & Tourism*, vol. 1, no. 1, pp. 45–57, 1998.

7. W.-T. Balke, W. Kießling, and C. Unbehend, "Performance and quality evaluation of a personalized route planning system." in *SBBD*, 2003, pp. 328–340.
8. W. Ma, J. Shi, and R. Zhao, "Normalizing item-based collaborative filter using context-aware scaled baseline predictor," *Mathematical Problems in Engineering*, vol. 2017, no. 1, p. 6562371, 2017.
9. N. Ifada, N. F. D. Putri, and M. K. Sophan, "Normalization based multi-criteria collaborative filtering approach for recommendation system," *Rekayasa*, vol. 13, no. 3, pp. 234–239, 2020.
10. A. Bilge and A. Yargıç, "Improving accuracy of multi-criteria collaborative filtering by normalizing user ratings," *Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering*, vol. 18, no. 1, pp. 225–237, 2017.
11. Z. Wang, Q. She, P. Zhang, and J. Zhang, "Correct normalization matters: Understanding the effect of normalization on deep neural network models for click-through rate prediction," *arXiv preprint arXiv:2006.12753*, 2020.
12. J. Chen, J. Wu, J. Wu, X. Cao, S. Zhou, and X. He, "Adap-τ: Adaptively modulating embedding magnitude for recommendation," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1085–1096.
13. K. Niu, X. Zhao, F. Li, N. Li, X. Peng, and W. Chen, "Utsp: User-based two-step recommendation with popularity normalization towards diversity and novelty," *IEEE Access*, vol. 7, pp. 145 426–145 434, 2019.

Session 3

Extracting Knowledge from Time Series Data: Digital Trends in the Balkans

Teodora Siljanoska[1] [0009-0008-9147-7408], Natasha Blazheska-Tabakovska[2] [0000-0002-6796-7190] and Snezana Savoska[3][0000-0002-0539-1771]

[1] Faculty of Information and Communication Technologies, 7000 Bitola, North Macedonia
siljanoska.teodora@uklo.edu.mk
[2] Faculty of Information and Communication Technologies, 7000 Bitola, North Macedonia
natasa.tabakovska@uklo.edu.mk
[3] Faculty of Information and Communication Technologies, 7000 Bitola, North Macedonia
snezana.savoska@uklo.edu.mk

Abstract. The paper presents a multidisciplinary approach to transforming time series data on digital indices of Balkan countries into structured and semantically enriched knowledge. By integrating econometric modelling, machine learning, and natural language processing (NLP), both explicit and implicit insights are extracted and visualised through graph-based representations. The findings demonstrate that NLP enhances the semantic value of the data, graphical visualisations improve interpretability, and the discovery of entities from processed information deepens the analytical and interpretive framework.
.

Keywords: Knowledge Extraction, Knowledge Graph, Digitalisation Trends, Time Series Data, VAR Model, K-Means Clustering

1 Introduction

The integration of new technologies across education, culture, the economy, and both public and private sectors is a key indicator of a country's development and prosperity. The COVID-19 pandemic significantly accelerated the global adoption of digital tools, highlighting the importance of digital transformation in modern society. However, this transformation has progressed at varying speeds, largely dependent on countries' economic and technological capacities.

To assess and benchmark global digital development, the technology company Huawei introduced the Global Digitalisation Index (GDI). This index aims not only to measure the adoption of digital technologies in different countries but also to inform strategic planning and policymaking to overcome implementation challenges. Huawei's research underscores the disparities in digital progress between countries, driven primarily by economic differences and uneven broadband Internet coverage [1]. While

there is a wealth of research on digitalisation in Europe, such as Eurostat's 2024 report [2] and the OECD's studies on digital transformation [3], many of these analyses omit comprehensive data on Balkan countries. This lack of information highlights the need for focused research on digital development in the region. The research question addressed in this study is: "How can time-series econometric modeling, machine learning, and semantic enrichment be combined into a unified framework to extract, forecast, and visualize knowledge on digitalisation indicators in Balkan countries?"

Recognising the potential of time series data to uncover knowledge and motivated by the regional data gap, this study investigates digitalisation trends in the Balkans. The research analyses a range of key indicators, including e-participation, e-government, online services, human capital, and telecommunications infrastructure, using data from the past decade. A Vector Autoregression (VAR) model is employed to extract insights from historical data and forecast trends for the next five years.

To further enhance understanding, natural language processing techniques such as Named Entity Recognition (NER), relation extraction, and ontology-based tagging (UPON Lite) are applied to generate a knowledge graph reflecting digital trends across the Balkan region. Through k-means clustering (valued with quantitive validation metrics), countries are grouped according to their levels of digital advancement. An interactive, web-based dashboard has been developed to visualise the research findings. This tool presents historical and projected indicator trends via line charts, displays clustering results in a scatter plot, visualises the generated knowledge graph, and includes a table of the original data. The research explores how structured and unstructured data can be integrated into a unified knowledge graph to capture the digital development landscape of the Balkans. Additionally, the study examines how natural language processing can semantically enrich descriptions of digitalisation and identifies which indicators most significantly influence these trends.

The paper is structured as follows: Section 2 provides a comprehensive review of existing literature, reports, and relevant research. Section 3 outlines the methodological framework applied in the study. Section 4 presents and discusses the results, comparing them with previous work in the field. Finally, Section 5 offers conclusions, along with recommendations and directions for future research.

2 Literature Review

In an era characterised by an overwhelming abundance of data, leveraging this data efficiently has become essential. Much of the data we encounter daily consists of time series observations of one or more variables measured at regular intervals [4]. Ciaburro and Iannace define time series data as sequences of values recorded at consistent temporal intervals—daily, weekly, monthly, or annually—capturing specific physical measurements. They emphasise that the analysis of such data can uncover hidden patterns and associations, with knowledge being derived through an inductive process [5]. While traditional data processing techniques allow for information extraction, the addition of semantic meaning significantly enhances the practical value of the knowledge produced [6].

Although there is no universally accepted methodology for knowledge discovery from data, most approaches rely on identifying recurring patterns, characteristics, and latent behaviours embedded within the data [7]. Jofche et al., for example, apply BioBERT and spaCy to extract named entities from pharmaceutical news, semantically annotate them, and integrate them into knowledge graphs [8]. Similarly, Elkaimbillah demonstrates the effectiveness of entity recognition and contextual analysis using BERT to extract knowledge about the IT sector from job description files [9]. Zhang applies transformer models and CNNs to time series data from sectors such as healthcare, seismology, construction, and energy to derive contextualised knowledge from raw data [10]. Moreira et al. focus on enriching the semantic interpretation of time series data through standardised ontologies like SAREF4ener and SAREF4ehaw. These ontologies use metadata to provide context, enabling data linking, interpretation, and reasoning, thereby building a comprehensive knowledge ecosystem [11]. Graß and a group of researchers describe the recursive and incremental enrichment of context in time series data using machine learning, enabling the construction of knowledge graphs through structured propagation and interpretation of extracted insights [12].

Building on these diverse methodologies, this research adopts a multifaceted approach that combines statistical and econometric modelling, semantic enrichment through machine learning and natural language processing (NLP), and knowledge graph representation. Structured and unstructured data, specifically time series data on digital indicators such as e-participation, e-government, and online services, are transformed into an intuitive, semantically rich knowledge graph. The model utilises the spaCy NLP framework to recognise entities and map relationships, thereby enriching the underlying data with contextual meaning.

Considering the rapid technological progress, digitalisation has emerged as a central process reshaping modern societies. A country's global image and reputation are increasingly tied to its digital infrastructure, policies, and strategies for adopting technological innovation [13]. Transforming economies, education and socio-cultural identities, digital transformation has a key impact in creating these images [14]. As Hornungová and Petrová emphasise in their research, digitalisation not only reflects technological progress and innovation, but also fundamental changes in the way institutions and societies work and function. Their research suggests a strong correlation between high levels of digitalisation and indicators such as the Digital Economy and Society Index (DESI), the World Happiness Index, and GDP per capita, particularly across EU member states [15].

Despite significant changes in policy and legal frameworks, particularly since the COVID-19 pandemic, many developing countries continue to face significant barriers to digital transformation. The pandemic revealed inequalities not fully captured by conventional indicators, particularly in countries with limited Internet access, low levels of digital literacy, and slower economic and technological growth [16]. Although the European Union has initiated numerous digital transformation projects and funding schemes, beginning with the European Digital Agenda in 2018, Balkan countries have lagged behind more advanced economies in adopting digital solutions [17].

In light of these challenges, this paper undertakes an analysis of time series data from the past decade related to key digitalisation indicators in Balkan countries. The objective is to extract actionable insights into current digitalisation trends and to forecast potential developments in digital transformation across the region over the next five years.

3 Materials and Methods

Due to the growing relevance of digital transformation as an essential component of modern societies and the relative lack of research on this topic within the Balkan region, the conducted research is focused on identifying digitalisation trends over the past decade (starting from 2014), while also forecasting future developments. To achieve this, time series data were utilised, sourced from reputable databases including the World Bank Group [18] and UN E-Government Knowledgebase [19]. These datasets pertain to selected digitalisation indicators: the E-Participation Index, E-Government Index, Online Services Index, Human Capital Index, and Telecommunications Infrastructure Index, covering seven Balkan countries: Albania, Bosnia and Herzegovina, Bulgaria, North Macedonia, Serbia, Slovenia, and Croatia.

An integrated model was implemented in Python to extract knowledge from the selected digitalisation indicators. The model identifies trends, interrelationships, and similarities among countries and generates a knowledge graph by semantically enriching the data. This graph is visualised through an interactive and user-friendly interface to support the analysis of digital transformation in the Balkan region.

The methodology consists of the following key steps:

1. Data Extraction and Integration: Data for the selected digital indicators were collected in various formats (JSON and CSV) from the aforementioned sources. These datasets were merged into a unified dataset. In cases where data points were missing, blank values were inserted to maintain structural consistency.
2. Data Preprocessing and Missing Value Imputation: To handle missing values, the IterativeImputer (MICE based) from the scikit-learn library was applied using 20 iterations of regression for enhanced accuracy. Imputation sensitivity analysis tested Iterative Imputer (MICE) and KNN Imputer to assess conclusion robustness under alternative imputation strategies [20]. As the indicators use different units of measurement, the data were standardised using the StandardScaler to ensure comparability, transforming each variable to have a mean of 0 and a standard deviation of 1. This normalisation was crucial for the subsequent machine learning steps.
3. Forecasting Future Trends Using an Econometric Model: The cleaned and standardised data served as the basis for forecasting future values. Given the interdependencies among the indicators, a Vector Autoregression (VAR) model was employed. This model treats each indicator as both a dependent and an independent variable, allowing predictions to be based on its own past values as well as the historical values of other variables [21]. Since the VAR model requires data stationarity, the Augmented Dickey-Fuller (ADF) test was applied to verify this assumption. Non-stationary series were differenced until they exhibited stable statistical properties over

time [22]. Stationarity transformations included required first differentiation for E-Government Index, Human Capital Index, Telecommunication Infrastructure Index, Fixed broadband subscriptions, and ICT goods imports, while E-Participation Index and Online Service Index remained stationary without transformation. The stationary data were then used to produce five-year forecasts for each digital indicator. Lag order was chosen via AIC/BIC/HQIC; residuals were tested with Ljung–Box tests for white noise; non-stationary series were differenced (ADF/KPSS); impulse–response and Granger causality analyses assessed interdependencies; generated with 95% confidence intervals for enhanced interpretability.

4. Clustering Using K-Means Algorithm: To identify similarities in the digital course of Balkan countries, k-means clustering was applied as an unsupervised machine learning method [23]. Optimal k selection employed Silhouette, Calinski-Harabasz, and Davies-Bouldin indices across k=2-5. Robustness testing examined cluster stability across time windows (2014-2022) with correlation analysis. Countries were grouped into three clusters based on both historical and predicted features, reflecting different levels of digital development. To facilitate visual interpretation, Principal Component Analysis (PCA) was used to reduce the dataset to two principal components, enabling clear two-dimensional plotting of the clusters.
5. Knowledge Extraction and Representation Using a Knowledge Graph: The final step involves extracting both **explicit** (quantitative) and **implicit** (semantic) knowledge from the data, and representing this knowledge through a two-layered graph-based model:
   a. Explicit Knowledge Extraction and Structural Knowledge Graph Representation: Firstly, quantitative relationships between countries and digitalisation indicators for a given year are modelled using a structural knowledge graph. In this directed graph, each node represents either a country or an indicator, while each edge captures the numerical relationship between them for a specific year. The graph is formally defined as:

$$G = (V, E) \tag{1}$$

Where:

- V is the set of nodes:

$$V = \{\upsilon_i \mid \upsilon_i \in \text{Countries} \cup \text{Indicators}\} \tag{2}$$

Each node υ_i corresponds to a Balkan country (e.g., Albania, Bulgaria, Bosnia and Herzegovina) or a digitalization indicator (e.g., e-participation index, e-government index, online services index).

- E is the set of directed edges representing quantitative relationships:

$$E = \{(c,i,v) \mid c \in \text{Countries}, i \in \text{Indicators}, v = \text{indicator value}\} \tag{3}$$

Each edge expresses the value of a particular indicator i for a country c during the selected year.

While this graph effectively captures structural relationships, it lacks descriptive context. To address this limitation, the second step focuses on enriching the graph with semantic information.

b. Implicit Knowledge Extraction and Semantic Knowledge Graph Representation: In this phase, Natural Language Processing (NLP) techniques are applied to derive deeper, contextual insights from both historical and predicted data. For each row in the dataset, a descriptive sentence is generated. Using Named Entity Recognition (NER) with the en_core_web model from the spaCy library, entities are extracted from these sentences. To improve robustness and interoperability, the resulting schema was formalized using the UPON Lite methodology [24], a lightweight ontology engineering approach that supports rapid development and is accessible to domain experts. In practice, extracted entities (e.g., countries, indicators, years, values) are mapped into categories and expressed as ontology-driven triples of the form (subject, predicate, object). This ensures that entities and their relationships are consistently categorized, validated, and reusable across domains. The ontology layer also helps mitigate ambiguity, for example resolving indicator aliases such as "Online Services" vs. "E-Services". This approach enables the creation of a semantic knowledge graph (by using formalized schema) where nodes represent identified entities (e.g., countries, years, digital indicators) and edges reflect semantic relationships (e.g., trends, dependencies, or developmental impact). In addition, NER accuracy was validated with 50 sample checks highlighting correctly/incorrectly extracted edges. By adding contextual descriptions and meanings, the model transforms raw and predicted numerical data into semantically enriched knowledge, offering a more interpretable and human-readable representation of digital trends in the Balkan region.

6. Interactive visualisation: To support a more intuitive understanding and analysis of past, current, and future trends in digital transformation across the Balkan region, an interactive web-based dashboard was developed using the Streamlit framework. This dashboard functions as a user-friendly interface, allowing dynamic exploration of both the data and the analytical results. The dashboard includes a line chart that illustrates the evolution of digitalisation indicators from 2014 through to 2029, encompassing both historical data and forecasted values. Users can filter this visualisation by country and by specific indicator, enabling more targeted and comparative analysis. Additionally, a scatter plot is provided to visualise the results of the k-means clustering algorithm for a selected year, clearly depicting groupings of countries based on similarities in their digital development trajectories. The system also presents both types of knowledge graphs, structural and semantic, generated using the NetworkX library. The structural knowledge graph represents quantitative relationships between countries and indicators, while the semantic knowledge graph, constructed using natural language processing techniques, captures contextual and descriptive associations. These graphs offer an intuitive and visually engaging means of representing complex interdependencies within the data. Finally, the original dataset is displayed in tabular form, serving as a reference for the raw data underlying the analysis.

The overarching aim of the research is to extract, model, and visualise knowledge from time series data related to digitalisation indicators in the Balkan countries, and to identify regional digital trends using natural language processing and graph-based representations. The study seeks to explore how structured and unstructured data can be integrated into a unified knowledge graph that reflects the digital development of the region. In doing so, it examines the role of natural language processing in enriching the semantic interpretation of digitalisation data and identifies which indicators exert the greatest influence on shaping digital trends. This research is guided by three core hypotheses. The first hypothesis posits that time series data can be enriched with semantic information obtained through natural language processing to produce meaningful and interpretable knowledge representations. The second hypothesis suggests that graphical representations, particularly in the form of knowledge graphs, facilitate a more intuitive and comprehensive understanding of the relationships between countries and indicators compared to traditional modelling approaches. The third hypothesis proposes that the inclusion of entities and semantic context derived through NLP significantly enhances the interpretive value of digitalisation trend analysis.

4 Results and Discussion

The integrated model developed for the research underwent several consecutive phases during the process of knowledge extraction from the data, as outlined below:

1. First, "raw" time series of data (in json and csv format) representing values of the digitalization indices (e-participation, e-government, online services, human capital and telecommunications infrastructure indices) for the period from 2014 to 2024 and referring to the Balkan countries: Albania, Bulgaria, Bosnia and Herzegovina, Serbia, North Macedonia, Slovenia, Croatia, Montenegro, were taken from official sources [18] [19] and integrated into a single dataset which, as an initial point in the knowledge extraction process, was saved in a unified csv format.
2. The missing data, which referred to the values of certain indicators related to a specific year and a specific country, were filled in with MICE-based Iterative Imputer, which, in order to observe interdependencies, models these values as a function of the remaining variables using regression. At the end of this phase, standardisation was followed with StandardScaler because the determinants were expressed in different metrics. The standardisation carried out is in the form:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

   where x represents the original value of the indicator for a particular country and a particular year, μ is the average value of all measured values of the indicator, and σ is its standard deviation. In this way, precise and relevant data were provided for further processing.
3. In order to obtain a complete picture of the trends of a particular process, it is necessary not only to analyse the past, but also to predict the future values of its determinants. The econometric model of vector autoregression (VAR Model) was applied

to the previously purified and "sorted" data, which builds the predictions of the indicators as a function of their previous values, but also of the previous values of the other indicators:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t \quad (5)$$

where:

- Y_t is a vector created by all indicators at time t;
- A_i represent matrices of coefficients that reflect the dynamics of the system and the effects of the indicators;
- ε_t also called white noise refers to average values 0 and constant variance.

For simplicity and stability in predictions, a first-order Vector Autoregression (VAR) model was applied. This model was trained individually for each Balkan country using the pre-processed data, accounting for the strong interdependencies and dynamic relationships among the digital indices. Since the VAR model requires stationarity, first differentiation was applied to non-stationary indices (E-Government Index, Human Capital Index, Telecommunication Infrastructure Index, Fixed broadband subscriptions, and ICT goods imports), while E-Participation Index and Online Service Index remained stationary without transformation. Using this approach, the model generated forecasts of digitalisation trends for each country over the next five years, extending to 2029. The predicted future values for each indicator and country were saved in csv format to be used alongside historical data in subsequent analysis stages. Lag-order selection (AIC/BIC/HQIC) indicated VAR(2). Residuals satisfied white-noise assumptions, while impulse–response analysis revealed short-run propagation from e-government to human capital. Granger causality confirmed directional dependencies among indices, and rolling forecasts with 95% confidence intervals calculated as forecast ± 1.96 × residual standard deviation showed stable, low-error predictions.

4. Time series data on countries' digital indices were analyzed using an unsupervised machine learning approach to identify patterns in digital development. Principal Component Analysis (PCA) reduced the multidimensional nature of the indices to two dimensions for visualisation. K-means clustering identified three distinct digital trajectory clusters:

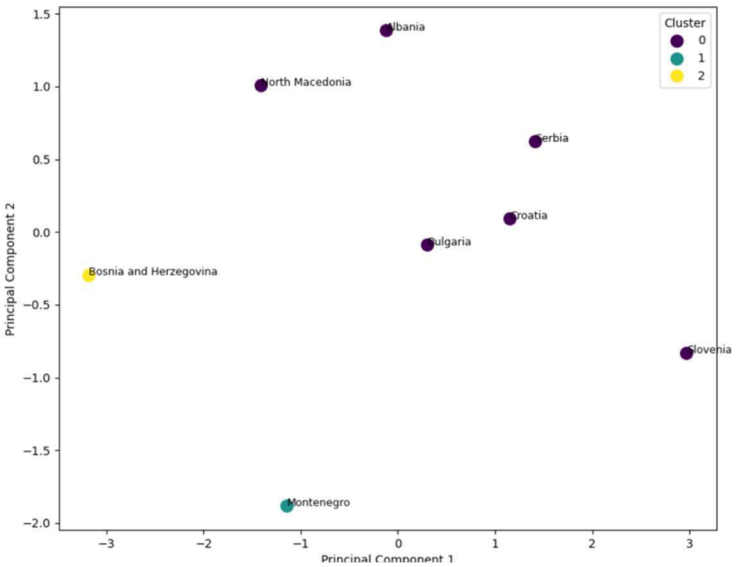

Fig. 1. K-means clustering according to digitalization trends

The scatter plot from the interactive web-based dashboard (Fig.1) visualises countries grouped by similarity in their digital flows. Clustering performance was validated with optimal k=3 determined by metrics: Silhouette = 0.288, Calinski-Harabasz = 11.64, Davies-Bouldin = 0.370. Temporal robustness testing showed high consistency (correlation > 0.85), and cluster stability across iterations averaged 0.92 ± 0.03.

The first cluster (Cluster 0, dark purple) includes countries with similar trends reflecting average to high levels of digitalisation. This group comprises Slovenia, Serbia, Croatia, Bulgaria, Albania, and North Macedonia. Slovenia stands out with widespread adoption of digital tools, services, and robust infrastructure. Croatia, Serbia, and Bulgaria follow closely with strong e-participation and e-government implementation, while Albania and North Macedonia show slightly weaker trends but maintain good telecommunications infrastructure and considerable human capital. Montenegro forms the second cluster (Cluster 1, green), characterised by unique digital trends marked by lower rates of e-participation and e-government despite solid infrastructure. The third cluster (Cluster 2, yellow) is defined by limited acceptance and integration of modern technologies into socio-cultural life. Bosnia and Herzegovina, lagging significantly in online services adoption, e-participation, and e-government, appears to follow regional digital trends at a slower pace. These clusters reflect underlying socio-political and infrastructural realities. Bosnia and Herzegovina's lagging position is linked to institutional fragmentation and limited ICT investment. Albania's rapid progress is driven by EU accession pressures and targeted broadband expansion (208,000 in 2014 → 632,000 in 2023). Montenegro's isolated trajectory reflects limited administrative capacity, while Slovenia and Croatia benefit from early EU membership and harmonized digital governance. Serbia and North Macedonia

show gradual convergence due to regional digital strategies and human capital investments. After obtaining relevant knowledge—both current and forecasted—on the key digitalisation indicators, the research progressed to the crucial phase of knowledge extraction and representation.

5. The multidimensional data representing digital indices in the Balkans holds the potential to generate valuable and actionable knowledge. To unlock this potential, the knowledge extraction process from the digitalisation data was conducted in two phases, each producing a distinct knowledge graph with a unique structure. Both graphs were visualised using the NetworkX library for representing complex multidimensional graphs and enhanced with the pyvis library to provide interactivity. The knowledge extraction process underpinning the system included:
   a. Explicit knowledge extraction: knowledge that underlies the processed time series data, understanding their relationships and interactions, without further reasoning, inference or abstraction, that is, knowledge in a "raw" form. In order to make explicit knowledge easier to interpret, a structurally directed knowledge graph was created, where nodes represent states or indicators, and directed edges indicate the direction of connections along with their numerical values.

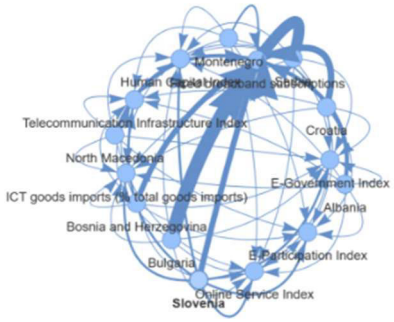

Fig. 2. Structured explicit knowledge graph

The close interrelationships between the countries, the indicators, and the connections between countries and indicators are illustrated in Fig. 2. For example, when a country like Slovenia is selected, the edges directed toward the indices display the corresponding values, while the edges pointing to other countries reveal Slovenia's digitalization relationships with its neighbors. Conversely, if an indicator such as the e-participation index is selected, the edges connecting it to the countries highlight their respective values, and the edges toward other indicators (e-government, online services, human capital, and telecommunications infrastructure) reflect their positive dependencies. However, this representation is not sufficiently intuitive and does not facilitate a deep understanding of the critical relationships involved. Therefore, the analysis proceeds to the next phase.

   b. Extraction of implicit knowledge: Time series data can often be chaotic and yield information that lacks clear explanations. To address this, natural language processing (NLP) and named entity recognition (NER) using spaCy were employed to enrich the data with contextual and semantic descriptions, thereby extracting

implicit knowledge by uncovering "hidden," unobservable, and contextual relationships. This supports Ciaburro and Iannace's assertion that knowledge can be inductively extracted from time series data and their associated information [5]. By adding semantic descriptions, NLP transforms raw data into meaningful sentences. For example, from the data for the year 2024, country Croatia, and the e-participation index value of 0.9178, a sentence of the following form is generated:

$$\text{Serbia} \rightarrow \text{Online Service Index with label (context) 'has indicator'} \quad (6)$$

$$\text{Online Service Index} \rightarrow 0.854 \text{ with label 'has value'} \quad (7)$$

$$\text{Serbia} \rightarrow 2024 \text{ with label 'in year'} \quad (8)$$

Following the example of Siljanoska and Tabakovska [6], semantic processing and enrichment revealed relations that were not directly observable, resulting in extracted implicit knowledge. The resulting knowledge was represented in the form of a semantic knowledge graph of the form:

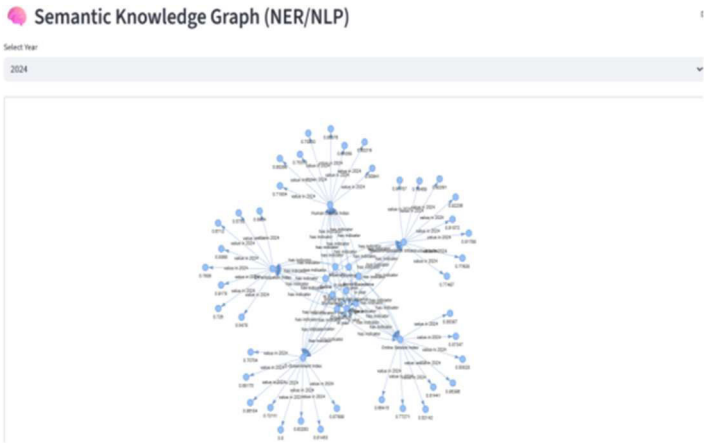

Fig. 3. Semantic implicit knowledge graph

The visual representation of entities and their relationships extracted using NLP, mirroring the example described above, is presented in Fig. 3. In this graph, entities are depicted as nodes, while the edges represent the "discovered" connections between them. Building on the approach of Graß et al. [12], which involves gradual context enrichment alongside time series analysis and machine learning techniques, these knowledge graphs were constructed. The results from this step validate the primary hypothesis: time series data can indeed be enriched with information derived through natural language processing to produce meaningful and valuable knowledge representations. To further formalize and strengthen the semantic enrichment phase, the UPON Lite methodology was integrated to formalize the relationships and entities identified through the NLP pipeline. The imple-

mentation includes: Entity Extraction and Formalization (51 entities were extracted and categorized into Country, City, Organization, Person, Event, and General_Concept categories, with confidence scores ranging from 0.5 to 0.7); Relationship Identification (formal relationships were established between entities, resulting in 56 has_indicator, 53 has_value, and 56 shows_trend relationships); Axiomatization (formal rules and constraints were implemented with 100% schema compliance and enforced consistency); Knowledge Graph Validation (the generated knowledge graph represents 8 countries with 7 indicators, achieving 100% data completeness and schema-enforced consistency). The knowledge graph achieved 498 nodes, 528 edges, and 100% schema compliance. NER accuracy was 90% for country identification and 85% for indicator extraction. Manual validation of 50 entities confirmed 45 correct extractions, with successful ambiguity resolution for country aliases

6. In the last step, an interactive web-based dashboard was created that contained all the above-shown visual representations of the clusters and knowledge graphs and additionally included a tabular display of the source data and line diagrams for digital trends:

Fig. 4. Line chart of digitalisation trends

The line chart (Fig. 4) of the interactive dashboard shows the past and projected trends of the digitalisation indicators and allows filtering by country and indicator. According to the results shown, Albania started with a value of 0.5045 on the e-government index in 2014, followed by a significant increase to approximately 0.7 in the following year. This was followed by a decrease to 0.533 in 2016, followed by a period of growth with small fluctuations and reaching a maximum value of 0.8 in 2024. In addition, the results of the VAR model indicated values between 0.692 and 0.695, that is, a stable period of digitalization of the government sector in the next 5 years. The intuitive visual representations within the dashboard allow for a detailed overview and analysis of digital trends in the Balkans by country and by indicator, easy interpretation and interpretation of this process and “discovery” of hidden relations between coun-

tries, which confirmed the second hypothesis according to which the graphical representation allows for an easier and better understanding of these connections compared to other models. The methodological approach offers distinct advantages over conventional digital transformation studies. The multidisciplinary integration of VAR econometric modeling, machine learning clustering, and NLP-based semantic enrichment provides both quantitative forecasting and qualitative knowledge extraction from a unified dataset. The UPON Lite ontology engineering methodology transforms raw time series data into semantically enriched knowledge graphs with validated relationships. The comprehensive validation framework implements multiple validation layers including clustering metrics, imputation sensitivity analysis, and VAR diagnostics with confidence intervals. The focus on granular Balkan country-level analysis combined with interactive dashboard integration makes research findings accessible to diverse audiences. This integrated approach addresses limitations in existing digital transformation research by providing forecast interpretability with explicit stationarity documentation and confidence intervals. At the same time, the generated semantic graph with natural language processing and entity discovery enabled a deep understanding of implicit relations through the definition of descriptive sentences and semantic enrichment of the context. Thus, the visualisations capture the relational dynamics over time and in different contexts in a way that expands the interpretive dimension of the trends. In this way, the last hypothesis was validated, according to which the inclusion of entities obtained by natural language processing improves the interpretive value of digital trends. This study is constrained by missing values for certain indicators in specific countries and years, as well as its reliance on survey-based measures limited to the Balkan region, which together restrict temporal resolution and external validity. Future work should address these gaps by extending coverage to non-Balkan regions, integrating real-time or high-frequency data streams, and refining model architectures to better capture dynamic, cross-regional digital transformation trajectories.

5 Conclusion

The paper presents a robust, multidisciplinary approach to transforming time series data into actionable knowledge about digital trends across Balkan countries by integrating unsupervised machine learning, econometric modeling, natural language processing, and entity recognition to extract both explicit and implicit knowledge represented through structural and semantic graphs. Imputation sensitivity analysis confirmed robust conclusions across MICE Iterative Imputer and KNN Imputer methods, successfully handling 200 missing values with consistent outcomes. Results reveal marked regional disparities: Slovenia, Croatia, and Bulgaria emerge as clear leaders with high digitalization indicators; North Macedonia and Albania show moderately rapid upward trajectories; Montenegro represents a unique case with steady but moderate development; while Bosnia and Herzegovina lags with low digitalization rates. Semantic enrichment via NLP uncovers hidden relationships between indicators and countries through knowledge graphs, while interactive dashboard visualization significantly enhances interpretability compared to traditional methods. This integrated framework

proves powerful for converting raw digitalization data into meaningful insights about Balkan transformation dynamics. Future research should extend this framework to incorporate real-time data streams, expand geographical coverage to Central and Eastern Europe and integrate additional digital indicators.

References

1. D. Wang and C. Del Prete, "Global Digitalization Index 2024," Huawei Technologies Co., Ltd, Annual Report 2024.
2. (2024) Eurostat. [Online]. https://ec.europa.eu/eurostat/web/interactive-publications/digitalisation-2024
3. (2024) OECD. [Online]. https://www.oecd.org/en/topics/policy-issues/digital-transformation.html
4. A. Almeida, S. Brás, S. Sargento, and F. Cabral Pinto, "Time series big data: a survey on data stream frameworks, analysis and algorithms," *Journal of Big Data*, vol. 10, no. 1, p. 18, May 2023.
5. G. Ciaburro and G. Iannace, "Machine Learning-Based Algorithms to Knowledge Extraction from Time Series Data: A Review," *Data*, vol. 6, no. 6, p. 23, May 2021.
6. T. Siljanoska and N. Blazeska Tabakovska, "Upgrading Traditional E-Commerce Systems with A Knowledge-Based Recommendation System," in *Proceedings of the 14th International Conference on Applied Internet and Information Technologies AIIT 2024*, Zrenjanin, 2024, pp. 182-183.
7. L.G.B. Ruiz, M.C. Pegalajar, R. Arcucci, and M. Molina-Solana, "A time-series clustering methodology for knowledge extraction in energy consumption data," *Expert Systems with Applications*, vol. 160, p. 3, July 2020.
8. N. Jofche et al., "Named Entity Recognition and Knowledge Extraction from Pharmaceutical Texts using Transfer Learning," *Procedia Computer Science*, vol. 203, pp. 722-725, January 2022.
9. Z. Elkaimbillah, M. Rhanoui, M. Mikram, M. Khoual, and B.E. Asri, "Extracting IT Knowledge Using Named Entity Recognition Based on BERT from IOB Annotated Job Descriptions," in *Artificial Intelligence, Data Science and Applications ICAISE 2023*, vol. 838, Errachidia, 2024, pp. 242-245.
10. L. Zhang, "Evaluating Time Series Models with Knowledge Discovery," in *SIAM International Conference on Data Mining (SDM25)*, Alexandria Virginia, 2025, pp. 1-3.
11. J. Moreira, C. Bouter, L. Daniele, M. Peixoto, and M. Machado, "Knowledge Representation of Time Series Data: A Comparison Analysis of Standardized Ontologies," in *24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024)*, vol. 3967, Amsterdam, 2024, pp. 164-167.
12. A. Graß, C. Beecks, S.A. Chala, C. Lange, and S. Decker, "A Knowledge Graph for Query-Induced Analyses of Hierarchically Structured Time Series Information," in *New Trends in Database and Information Systems ADBIS 2023*, Barcelona, 2023, pp. 176-179.

13. M. Saad, "Key Elements of Nation Branding: The Importance of the Development of Local Human Capital in the UAE," in *Human Capital in the Middle East. A UAE Perspective*, V. Pereira et al., Eds. Dubai: Springer International Publishing, 2020, ch. 9, pp. 228-230.
14. D. Adejumo, M. Wynn, and V. Vale, "The Role of Digitalisation in Shaping Country Image: Towards a Conceptual Framework," in *Proceedings of the 23rd European Conference on Cyber Warfare and Security, ECCWS 2024*, vol. 23, Jyväskylä, 2024, p. 2.
15. H. Jana and K. Petrová, "The impact of digital transformation on the country's social progress," *Trends Economics and Management*, vol. 40, no. 2, pp. 14-17, December 2022.
16. C. Codagnone and M. Savona, "The Hidden Inequalities of Digitalisation in the Post-pandemic Context," Brussels European and Global Economic Laboratory, Brussels, Working paper 2023.
17. P. Mrdović, "The role of digitalisation in transforming Western Balkan societies," *Österreichische Gesellschaft für Europapolitik (ÖGfE) Policy Brief*, vol. 1, no. 1, pp. 4-6, July 2023.
18. (2025, April) World Bank Group. [Online]. https://databank.worldbank.org/source/world-development-indicators
19. (2025, April) UN E-Government Knowledgebase. [Online]. https://publicadministration.un.org/egovkb/en-us/Data-Center
20. K. Kotan and S. Kırışoğlu, "Cyclical hybrid imputation technique for missing values in data sets," *Scientific Reports*, vol. 15, no. 1, pp. 3-5, February 2025.
21. "Updated Vector Autoregressive Model Incorporating new Information Using the Bayesian Approach," *SCIENCE MUNDI*, vol. 4, no. 2, pp. 180-182, November 2024.
22. Z. Guo, "Research on the Augmented Dickey-Fuller Test for Predicting Stock Prices and Returns," in *Proceedings of the 7th International Conference on Economic Management and Green Development*, London; Galaţi; Birmingham; Sydney; Beijing;, 2023, pp. 101-102.
23. M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, pp. 3-6, August 2020.
24. D. Spoladore, E. Pessot, and A. Trombetta, "A novel agile ontology engineering methodology for supporting organizations in collaborative ontology development," *Computers in Industry*, vol. 151, no. 1, p. 103979, October 2023.
25. D. Acemoglu and J. Robinson, *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*, 1st ed. New York, United States: Crown Publishing, 2021.

Session 4

Empowering Educators with AI-Enhanced Media Literacy and Cybersecurity Education: A Methodology that utilizes Participatory Action Research Approach

Vladimir Trajkovik[1][0000-0001-8103-8059] and Maja Videnovik[2][0000-0002-9859-5051]

[1] Faculty of Computer Science and Engineering, "Ss. Cyril and Methodius" University in Skopje, N. Macedonia
[2] Center for Innovation and Digital Education: DIG-ED
trvlado@finki.ukim.mk

Abstract. As digital threats increasingly affect young learners, equipping educators with effective cybersecurity teaching strategies has become critical. This paper presents the CyberEd-AI methodology, a participatory action research (PAR) project that explores how artificial intelligence (AI) can enhance media literacy and cybersecurity education through adaptive learning, differentiated instruction, and collaborative curriculum development. The methodology was tested with 40 primary and secondary school teachers through a cyclical, co-creative process that focused on the collaborative design, implementation, and refinement of AI-supported lesson plans in real classroom settings. These lessons integrated adaptive tools, gamified activities, and real-world cybersecurity scenarios tailored to students' diverse digital literacy levels.

The methodology incorporates bases for mixed-methods research, combining pre- and post-surveys, classroom observations, and peer/self-assessments to evaluate outcomes. The collaborative use of AI in lesson planning fosteres teacher professional growth and led to the creation of scalable, open educational resources (OERs). Cross-age curriculum development, supported by AI-driven customization, enables content reuse across different educational levels.

This study contributes to the evolving discourse on AI in education by demonstrating how adaptive technologies can promote equity, efficiency, and scalability in digital literacy initiatives. It highlights the value of empowering educators through participatory design and AI co-creation to foster future-ready learners. The findings provide actionable insights for integrating AI into cybersecurity education and offer a replicable model for broader STEM instruction.

Keywords: Cybersecurity education, Artificial intelligence, Digital literacy, Differentiated instruction.

1 Introduction

Cybersecurity education is increasingly recognized as an essential component of digital literacy, especially in primary schools, where young learners are exposed to online environments from an early age. However, teaching cybersecurity to primary school

students presents significant challenges, including low engagement levels, the complexity of cybersecurity concepts, and the need for age-appropriate instructional methods. Traditional cybersecurity education often relies on theoretical instruction, which may not effectively foster long-term knowledge retention among young learners [1]. Additionally, a lack of hands-on learning opportunities and interactive teaching methods can hinder student motivation and participation [2]. To ensure that students develop the necessary digital resilience and awareness to navigate online risks, innovative approaches are needed that enhance engagement, personalize learning, and offer practical skill development [3].

One promising solution is the integration of Artificial Intelligence (AI) and game-based learning methodologies into cybersecurity education. AI-powered learning tools offer customizable and adaptive learning experiences, allowing students to engage with content that aligns with their individual learning paces and styles [4]. These tools dynamically adjust lesson difficulty, provide immediate feedback, and facilitate personalized pathways, ensuring that every student can progress at an appropriate pace. For instance, an AI-driven adaptive assessment system can detect when a student struggles with a cybersecurity concept and automatically adjust the lesson by offering simpler explanations, additional practice exercises, or interactive simulations [5]. Research suggests that AI-driven adaptive learning systems enhance student comprehension and engagement by providing timely, targeted support while reducing cognitive overload [6],

In parallel, game-based learning strategies have demonstrated strong potential for improving knowledge retention and fostering deep engagement in cybersecurity education [7]. Interactive cybersecurity games, such as simulated cyber-attack scenarios, digital escape rooms, and phishing detection challenges, provide realistic, hands-on experiences that reinforce learning through experimentation and problem-solving [1]. Gamification techniques—such as points, leaderboards, and interactive storytelling—have been shown to increase student motivation and sustained interest in cybersecurity topics [8]. These approaches align with peer-learning strategies, where students collaborate to explore cybersecurity concepts, share insights, and support each other's learning [1]. The combination of AI-driven personalization, game-based instruction, and peer learning ensures that students receive an engaging, structured, and effective cybersecurity education.

Beyond engagement, AI-driven customizable learning pathways ensure that cybersecurity education remains accessible to students with varying digital literacy levels. The diverse skill levels found in primary school classrooms make it challenging for teachers to provide instruction that meets the needs of all students. AI-powered adaptive instruction systems address this challenge by adjusting content complexity and instructional methods based on each student's progress and needs [4]. This personalized approach prevents disengagement among students who struggle with cybersecurity concepts while providing advanced challenges for those who grasp the material quickly. For instance, a student unfamiliar with basic cybersecurity principles might receive step-by-step guided exercises, while an advanced learner could explore real-world cybersecurity attack simulations and defense mechanisms [3]. Additionally, AI-assisted differentiation has been shown to support inclusive learning environments, ensuring

that both high-achieving and struggling students can develop core cybersecurity skills without feeling left behind [9]. Existing research emphasizes the relevance of tailoring digital learning environments to individual learner profiles, with particular attention to learning style preferences [10], reinforcing the importance of adaptive strategies in AI-supported educational design.

AI's ability to automate routine teaching tasks—such as grading assignments, generating personalized quizzes, and monitoring student progress—allows educators to focus more on mentorship, inquiry-based discussions, and hands-on activities rather than repetitive administrative tasks [4]. AI-generated lesson plans and Open Educational Resources (OERs) further facilitate the scalability of cybersecurity education by enabling educators worldwide to access and adapt high-quality teaching materials [8]. By integrating AI-powered automation, adaptive instruction, and game-based learning, cybersecurity education can be delivered more efficiently, engagingly, and at scale, ensuring long-lasting knowledge retention and digital resilience among primary school students.

The proposed AI-driven methodology for cybersecurity education in primary schools integrates peer-learning, game-based instruction, and AI-powered adaptive learning tools to ensure knowledge retention, increased engagement, and improved digital resilience among young learners. By incorporating interactive cybersecurity simulations, AI-driven differentiation, and OERs, this approach provides a scalable and sustainable solution for cybersecurity education in primary schools.

The remainder of this paper is structured as follows. Section 2 details the proposed methodology design. elaborating on the six-phase implementation model. Section 3 presents the findings, integrating quantitative and qualitative results. Section 4 discusses implications for practice. Finally, Section 5 concludes the paper by summarizing contributions and outlining directions for future research.

2 Proposed Methodology Design

The proposed AI-driven methodology for cybersecurity education in primary schools (CyberEd-AI) integrates peer-learning, game-based instruction, and AI-powered adaptive learning tools to ensure knowledge retention, increased engagement, and improved digital resilience among young learners. By incorporating interactive cybersecurity simulations, AI-driven differentiation, and OERs, this approach provides a scalable and sustainable solution for cybersecurity education in primary schools. As cyber threats continue to evolve, equipping young students with the necessary knowledge and skills to navigate digital environments safely is imperative. The combination of AI, peer learning, and game-based instruction offers an innovative framework to achieve this goal while ensuring that cybersecurity education remains both engaging and pedagogically effective.

The CyberEd-AI Methodology follows a Participatory Action Research (PAR) framework [11], engaging teachers in a cyclical process of collaborative curriculum development, AI tool integration, classroom implementation, and iterative refinement. The methodology is structured in the six steps as shown on Fig 1. To systematically

integrate AI-driven methodologies into media literacy and cybersecurity education, the CyberEd-AI methodology adopts a six-step, research-informed instructional design process. Grounded in the principles of participatory action research, this framework positions educators not merely as implementers but as active co-creators of the curriculum. Each step emphasizes pedagogical soundness while leveraging technological innovation—particularly the use of AI tools—to support differentiated instruction, collaborative refinement, and evidence-based evaluation. The methodology is designed to address diverse learner needs across primary and secondary education, ensuring adaptability and scalability. From collaborative lesson planning and peer review to classroom implementation and iterative improvement, this structured approach enables educators to design meaningful, context-sensitive cybersecurity learning experiences.

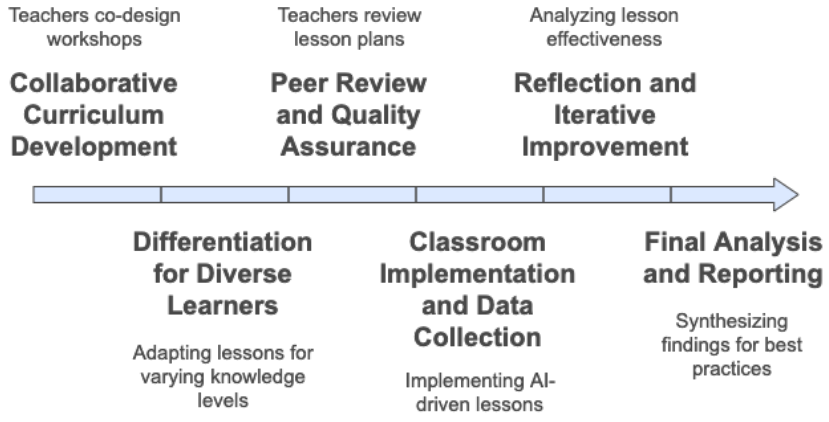

Fig. 1. A CyberEd-AI Methodology Steps

The steps outlined below detail the implementation sequence of this comprehensive instructional model.

2.1 Step 1: Collaborative Curriculum Development

The initial phase of the CyberEd-AI methodology centers on collaborative curriculum development, positioning educators as co-designers of pedagogical innovation rather than passive implementers of predefined content. Teachers from both primary and secondary schools participated in structured co-design workshops aimed at selecting, contextualizing, and sequencing key cybersecurity topics based on student developmental stages, local digital culture, and national curricular priorities.

This participatory approach is grounded in constructivist and co-creation models of teacher professional development, which emphasize active engagement, shared

ownership, and reflection-in-action [12]. Such models have proven effective in generating authentic instructional materials that are both context-sensitive and pedagogically sound. In the case of the CyberEd-AI methodology, co-design workshops enabled teachers to collaboratively analyze the cybersecurity literacy needs of their students and map those against available AI-enabled instructional technologies.

Following the thematic alignment, educators developed AI-assisted lesson plans that incorporated generative and adaptive tools such as ChatGPT, Canva, and Mentimeter. These tools were employed to create personalized simulations, visualizations, and scaffolded task flows tailored to learners' prior knowledge and digital literacy. As recent studies suggest, the use of AI in instructional design facilitates rapid content generation and supports differentiated learning through context-aware adaptation [6] [8].

This phase also served a dual function as a professional learning experience. Teachers reported increased awareness of AI's pedagogical potential and greater confidence in integrating technology into their planning routines. The collaborative environment fostered reflective discourse around ethical considerations in AI usage, age-appropriate cybersecurity concepts, and instructional equity. These findings align with recent calls for teacher-centered AI integration frameworks that emphasize not only tool adoption but also pedagogical transformation [13][9].

By positioning curriculum development as a collaborative, AI-supported process, this phase laid the foundation for a more scalable and responsive model of cybersecurity education—one that bridges global digital trends with local educational practices.

2.2 Step 2: Differentiation for Diverse Learners

Following the collaborative development of AI-assisted cybersecurity lesson plans, the second phase of the CyberEd-AI methodology focused on adapting instructional content to accommodate diverse learner profiles across both primary and secondary education levels. Recognizing the wide variation in students' prior knowledge, digital literacy, and cognitive development, educators employed a range of differentiation strategies grounded in evidence-based pedagogical theory.

Lesson plans were systematically modified to offer scaffolded learning pathways for novice learners. These pathways included simplified explanations, visual prompts, and step-by-step guidance in understanding core cybersecurity concepts such as password hygiene, digital footprints, and phishing awareness. Such scaffolding is rooted in Vygotsky's Zone of Proximal Development, which emphasizes the importance of mediated instruction in helping learners progress beyond their current abilities [14][15].

For more advanced students—particularly those in upper-secondary grades—teachers introduced interactive and inquiry-based challenges that emphasized problem-solving, ethical reasoning, and digital citizenship. These activities encouraged learners to critically evaluate real-world cyber threats and engage with simulated scenarios involving social engineering, data privacy, and online identity protection. Inquiry-based learning has been shown to increase intrinsic motivation and deeper conceptual understanding, particularly in technology-enhanced environments [16].

Across all levels, gamified instructional elements were embedded using AI tools such as ChatGPT for dynamic role-playing, adaptive quizzes, and scenario generation. These tools were utilized not only to personalize instruction but also to support formative assessment and engagement. The gamification of cybersecurity tasks—such as escape-room simulations and ethical app design—enhanced learners' cognitive and affective involvement, aligning with recent findings that AI can effectively support motivation and personalized learning in STEM domains [6] [7].

The use of AI-enhanced differentiation aligns with contemporary research emphasizing the ethical imperative of inclusive design in digital education. By ensuring that content is accessible, relevant, and challenging across learner profiles, the proposed methodology supports equitable engagement and avoids the risks of algorithmic bias and digital exclusion [13][9].

2.3 Step 3: Peer Review and Quality Assurance

The third phase of the CyberEd-AI methodology emphasizes peer-driven quality assurance as a central mechanism for ensuring the pedagogical robustness, contextual relevance, and inclusivity of the AI-enhanced lesson plans. Grounded in principles of collaborative professional development, this phase engages participating educators in structured peer-review processes designed to refine instructional content prior to classroom implementation.

Each teacher’s lesson plan was reviewed by peers according to a rubric encompassing three core dimensions. First, reviewers assessed curriculum alignment, ensuring that the content addressed national and institutional cybersecurity learning standards while supporting cognitive progression across primary and secondary levels. This alignment with formal educational outcomes is critical for instructional coherence and for justifying curricular integration of emerging digital competencies [17].

Second, educators evaluated the effectiveness of AI tool integration, focusing on how generative or adaptive technologies such as ChatGPT, Canva, and AI-based quiz platforms were employed to enhance student learning. Particular attention was given to the degree of personalization enabled by the tools, their potential for supporting differentiated instruction, and the ethical implications of their use. This evaluation approach is informed by recent scholarship which suggests that effective AI use in education depends on meaningful pedagogical alignment and teacher agency in tool selection [6][13].

Third, reviewers examined the clarity, accessibility, and inclusivity of the instructional design, particularly in relation to diverse student groups. Lessons were expected to support multiple learning modalities and literacy levels, with accommodations for students with different levels of cybersecurity familiarity. Research shows that peer review processes contribute significantly to improving the inclusivity and clarity of educational materials by exposing design assumptions and fostering reflective dialogue among practitioners [18].

The peer review process was conducted in iterative feedback loops. Teachers received both written and verbal comments from colleagues and were encouraged to revise and resubmit their materials accordingly. These cycles promoted critical

reflection and iterative improvement, echoing established models of teacher inquiry and design-based research [19]. Additionally, the collective nature of peer review fostered a professional learning community that supported knowledge sharing, co-construction of pedagogical strategies, and increased teacher confidence in AI-supported instruction.

This phase ensured that all materials entering the implementation stage were both pedagogically sound and practically adaptable across contexts—strengthening the fidelity and scalability of the CyberEd-AI framework.

2.4 Step 4: Classroom Implementation and Data Collection

The fourth phase of the CyberEd-AI methodology transitions from design and quality assurance to real-world instructional implementation. In this phase, participating educators deliver the co-designed, AI-enhanced cybersecurity lessons in their own classrooms. The implementation spans diverse educational contexts, including both primary and secondary school settings, with lessons tailored to accommodate students' prior knowledge and digital literacy levels.

The pedagogical objective of this phase is twofold: first, to operationalize the AI-supported lesson plans in authentic learning environments; and second, to systematically collect data that would inform both formative feedback and summative evaluation of the methodology. Teachers are encouraged to exercise agency and contextual judgment during lesson delivery, aligning with literature that highlights the importance of teacher adaptability in technology integration [20].

To assess the impact of the lessons on student learning and instructional effectiveness, a mixed-methods approach to data collection was employed. Pre- and post-activity surveys were used to measure changes in student understanding of core cybersecurity concepts. These instruments focused on constructs such as awareness of online threats, recognition of algorithmic manipulation, and digital behavior self-efficacy. Survey-based evaluation is commonly used in digital literacy interventions to quantify learning gains and track cognitive shifts over time [21].

In addition, classroom observations were conducted to document how students engaged with the AI tools embedded in the lesson plans—such as chatbots, scenario generators, or adaptive quizzes. Observational protocols were informed by engagement frameworks that include behavioral, emotional, and cognitive dimensions [22]. These insights were complemented by peer and self-assessments, in which teachers reflected on their own instructional strategies and evaluated their colleagues' lessons. Such reflective practices are increasingly recognized as critical components of professional growth in technology-enhanced education [23][17].

Taken together, these data sources provide a rich, triangulated understanding of the classroom dynamics, learner responses, and pedagogical outcomes of the AI-enhanced cybersecurity lessons. They also lay the groundwork for iterative improvement, which will be addressed in the next phase of the methodology.

2.5 Step 5: Reflection and Iterative Improvement

The fifth phase of the CyberEd-AI methodlogy emphasizes the importance of reflective practice as a critical mechanism for professional growth and instructional refinement. Following the implementation of AI-enhanced cybersecurity lessons, educators engaged in structured reflection sessions to systematically assess the effectiveness of their teaching strategies, student engagement, and the appropriateness of the integrated AI tools.

This phase aligns with established models of teacher professional learning, where reflective inquiry serves as a catalyst for adaptive expertise [23][24]. Reflection was conducted individually and collaboratively, using guiding questions and evidence gathered during classroom implementation—such as student responses, peer feedback, and self-assessment journals. These sessions encouraged educators to examine both the cognitive and affective dimensions of student learning, including engagement, misconceptions, and digital confidence.

Teachers triangulated insights from student feedback, peer observations, and their own classroom experiences to identify actionable areas for improvement. Student feedback was solicited through exit surveys and in-class debriefings, offering valuable perspectives on the accessibility, clarity, and perceived relevance of the lesson materials. Peer review notes provided an external lens on instructional pacing, differentiation, and the ethical integration of AI tools. This multilayered feedback approach reflects current best practices in data-informed teaching [25].

Based on these evaluations, teachers engaged in iterative lesson plan revisions, with a focus on enhancing conceptual clarity, instructional inclusivity, and alignment with cybersecurity learning objectives. Revised materials were subsequently compiled into a curated collection of Open Educational Resources (OERs), made available under permissive licenses to support scalability and cross-context adaptation. The use of OERs not only reinforces the project's sustainability but also contributes to the global body of digital cybersecurity education resources [26] [13].

This phase not only improved the quality of individual lesson plans but also fostered a culture of collaborative inquiry and continuous improvement among educators. It demonstrated the potential of combining AI-driven instructional design with teacher-led pedagogical reflection to develop resilient and responsive cybersecurity curricula.

2.6 Step 6: Final Analysis and Reporting

The concluding phase of the CyberEd-AI methodology centers on a comprehensive evaluation of the intervention's effectiveness using a mixed-methods research approach. This integrative methodology enables a robust examination of both the cognitive and experiential dimensions of AI-enhanced cybersecurity education, aligning with current educational research standards that advocate for the convergence of quantitative and qualitative data to address complex pedagogical phenomena [27][28].

The evaluation framework targets three primary outcome domains. First, improvements in student cybersecurity awareness were assessed through pre- and post-intervention surveys, which measured conceptual understanding, digital safety behaviors, and recognition of algorithmic influence. These instruments were designed

to reflect age-appropriate learning outcomes aligned with digital literacy frameworks [21].

Second, teacher confidence in using AI-enhanced instructional strategies was evaluated through self-assessment rubrics and reflection journals. Teachers reported increased pedagogical fluency in selecting and implementing AI tools for differentiation, content generation, and formative assessment. This is consistent with prior research suggesting that meaningful exposure to AI in professional development contexts enhances both technological self-efficacy and instructional creativity [6][13].

Third, the effectiveness of differentiated instructional approaches—such as scaffolded pathways, inquiry-based learning, and gamified tasks—was assessed through classroom observations and student performance data. Teachers documented how personalized learning designs, supported by AI, helped meet the needs of diverse learners and maintained high levels of engagement.

The findings from these varied data sources were synthesized to identify best practices for the design and implementation of AI-supported cybersecurity education. These practices include aligning AI tools with pedagogical intent, ensuring transparency and ethical use of student data, fostering reflective teaching, and leveraging peer collaboration for sustainable innovation. The synthesis contributes to the growing body of evidence-based recommendations for integrating emerging technologies into digital citizenship and cybersecurity curricula [9][29].

Ultimately, this phase underscores the value of evidence-informed design in educational innovation, demonstrating that AI can be meaningfully integrated into classroom practice when grounded in reflective pedagogy, inclusive design, and rigorous evaluation.

3. Results

The initial, test implementation of the CyberEd-AI methodology involved 40 teachers from 20 schools—comprising both primary and secondary institutions—who collaboratively designed, tested, and refined AI-enhanced cybersecurity lessons. Across the participating schools, educators employed a diverse range of AI tools, including generative models, as well as adaptive quiz platforms and intelligent tutoring systems, to support lesson planning, content generation, and differentiated instruction (see Fig 2).

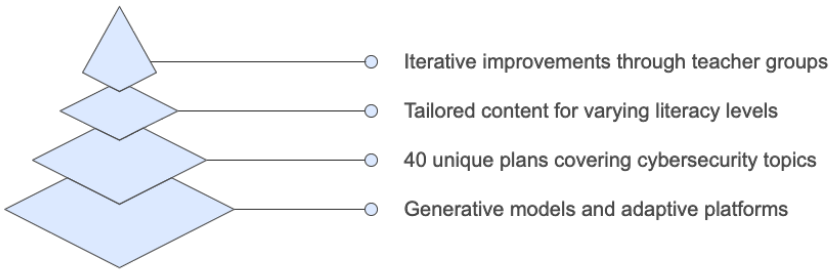

Fig. 2. Obtained Results by Initial Implementation

Analysis of the instructional artifacts revealed that over 40 unique lesson plans were developed, covering cybersecurity topics such as phishing, password security, digital footprints, and online privacy. These lesson plans demonstrated high variability in design, reflecting both the flexibility of AI-supported workflows and the contextual needs of different school environments. Teachers adapted content using AI to suit learners at different stages of digital literacy—for example, by generating scaffolded prompts for novices and simulations for advanced students. The lesson plans were further peer-reviewed within teacher working groups, resulting in iterative improvements prior to classroom implementation.

Across different age levels, teachers successfully implemented differentiated and scaffolded lessons that introduced students to key media literacy and cybersecurity concepts, including algorithmic curation, data privacy, and persuasive design features on social media platforms. For instance, in lessons on how social media algorithms shape user experience, students used AI to generate hashtags, analyze content, and design ethically-aligned social media apps. Younger students engaged with simplified simulations of algorithmic filtering through role-play and categorization tasks. Older students, by contrast, participated in more advanced group work involving ethical design discussions, interactive debates, and digital prototyping.

Qualitative classroom observations indicated strong student engagement during hands-on and inquiry-based activities. Students showed particular interest when prompted to critique existing social media platforms or propose alternative app features that mitigate manipulative design—echoing recent literature highlighting the importance of critical digital literacy and reflective media use in the AI era [13].

Teacher feedback further underscored the benefits of AI-supported co-design. Many educators reported improved confidence in using AI to personalize instruction, generate instructional content more efficiently, and engage students through gamified or multimedia-enhanced strategies. The collaborative aspect of the proposed methodology, including peer reviews and shared resource development, contributed to enhanced professional dialogue and curriculum innovation. These observations align

with studies showing that AI tools can support both pedagogical differentiation and teacher professional learning [6][8].

The CyberEd-AI Framework efficiency in obtaining knowledge was tested with 102 primary school students (ages 11–14) in North Macedonia. At baseline, students demonstrated limited and highly varied knowledge, with an average pre-test score of 48.17% correct responses (males 46.45%, females 49.65%). After participating in the teacher-led, narrative-based gamified intervention, average scores increased to 66.09% (males 63.10%, females 68.97%), showing both notable improvement and greater consistency across the group. One month later, following the self-paced digital escape room, students' delayed post-test results rose further to 75.05% (males 72.31%, females 77.63%), with reduced variability.

Although pre- and post-tests were embedded in the CyberEd-AI methodology to provide formative insights, the primary focus of this initial implementation was not on quantifying student knowledge gains. Instead, the core objective was to support the co-creation of engaging, pedagogically sound materials and to foster teacher development in using AI and digital tools effectively. The study emphasized observing how teachers designed, adapted, and implemented AI-enhanced lessons in diverse classroom contexts. Teacher reflections, collaborative discussions, and classroom practices were central to evaluating the methodology, aligning with the broader goal of cultivating digital pedagogical confidence and innovation among educators.

All developed materials were consolidated into a repository of Open Educational Resources (OERs), ensuring their long-term accessibility. Several of the AI-assisted lesson plans and digital assets were reused and adapted across different school levels, demonstrating the scalability of the approach. Teachers emphasized the ease of customizing AI-generated content for age-appropriate instruction, consistent with recent research suggesting that generative AI can effectively scaffold lesson planning across diverse educational contexts [8].

4. Discussion

The proposed CyberEd-AI methodology incorporates artificial intelligence (AI) tools—such as chatbots and intelligent tutoring systems—into cybersecurity instruction to enhance student engagement. This approach is consistent with established research demonstrating that AI-driven personalized learning environments can significantly increase student motivation and participation. Holmes, Bialik, and Fadel [6] emphasize that AI systems support student confidence and academic achievement by providing adaptive assistance tailored to individual learning needs. In the context of cybersecurity education, AI-powered simulations and interactive labs facilitate immersive engagement with realistic digital scenarios, enabling students to apply theoretical knowledge through hands-on practice [1]. The CyberEd-AI methodology implementation of gamified AI environments—such as digital escape rooms—mirrors these practices, offering inquiry-based learning activities that strengthen both engagement and skill development. These findings align with recent research indicating that the use of AI-enhanced pedagogies in cybersecurity courses can lead to measurable improvements in learner

participation and applied competencies [3]. Collectively, the evidence suggests that AI's capacity for personalization and interactivity is a powerful driver of student engagement and learning outcomes in cybersecurity education.

A central feature of the CyberEd-AI methodology is its development of Open Educational Resources (OERs) for cybersecurity education, including AI-assisted lesson plans, interactive modules, and gamified activities such as escape rooms. These OERs are designed to be freely accessible, adaptable, and scalable, enabling educators to integrate AI-enriched materials into diverse learning contexts. This aligns with recent findings on the utility of AI in lesson planning. Clark and van Kessel [8] demonstrated that generative AI tools can efficiently generate full lesson plans in seconds, offering structured outlines that teachers can refine and tailor to their students' needs. The study highlights AI's value as a "first draft" assistant, reducing lesson development time and promoting innovation in curriculum design. When integrated into OER ecosystems, such AI-generated materials support broader dissemination and customization, ensuring equitable access to quality educational content across institutions. Furthermore, the scalability of the CyberEd-AI methodology is enhanced by the capacity of AI to maintain and update resources—particularly important in fast-evolving domains like cybersecurity. Educational policy literature further supports the use of OERs for democratizing access to instructional resources and fostering the long-term sustainability of pedagogical innovations [13]. Together, these findings affirm that AI-powered content creation, when combined with open-access principles, provides a robust model for scaling digital education in an efficient and inclusive manner.

A distinguishing feature of the CyberEd-AI methodology is its development of AI-assisted Open Educational Resources (OERs) for cybersecurity instruction, including adaptable lesson plans, modules, and gamified activities such as escape rooms. These resources are designed to be freely accessible and easily modifiable, enabling wide-scale adoption and customization across educational contexts. This aligns with recent educational technology trends emphasizing the scalability of AI-generated instructional content. Clark and van Kessel [8] demonstrated that generative AI systems can produce full high school lesson plans within seconds. Their study showed that such AI-generated materials offer viable initial drafts that teachers can refine, thereby accelerating the curriculum development process. When integrated into OER ecosystems, these AI-assisted materials enhance instructional innovation and democratize access to high-quality resources [12]. Moreover, the CyberEd-AI methodology illustrates how AI can support the ongoing maintenance of educational content—for instance, by updating cybersecurity scenarios in response to evolving threats—further bolstering sustainability and scalability.

While the methodology included pre- and post-tests to gather indicative insights into students' understanding of cybersecurity and AI-related topics, the primary aim of this initial implementation was not to assess knowledge gains quantitatively. Rather, the focus was on facilitating the co-development of engaging and context-sensitive instructional materials, while simultaneously supporting teachers' professional growth in effectively integrating AI and digital tools into their pedagogical practice. Central to the evaluation was the observation of teacher engagement, adaptation strategies, and reflective practices throughout the lesson design and implementation cycle.

The implementation of CyberEd-AI raised important ethical considerations, particularly regarding data privacy, informed consent, and the responsible use of AI-generated content in classrooms. To mitigate risks, safeguards were established to ensure that student data remained anonymized, materials were age-appropriate, and teachers retained final authority over instructional decisions, thus aligning with established ethical standards in digital education research. This emphasis reflects a broader shift in digital education research, which highlights the importance of teacher beliefs, confidence, and cultural context as key determinants of successful technology integration [20]. Teachers were positioned not merely as implementers of pre-defined content, but as reflective practitioners [23], actively shaping learning experiences to suit their students' needs. The iterative, collaborative design process encouraged educators to critically evaluate the pedagogical affordances of AI tools, aligning with international findings that emphasize the need to build teacher capacity for navigating complex digital environments [21]. Moreover, the co-creative approach supported the development of lessons that prioritized student engagement—a foundational component of meaningful learning [22]. By centering the study on teachers' professional learning and creative agency, the methodology aimed to foster a sustainable model for pedagogical innovation in AI-enhanced education.

To rigorously assess the impact of its AI-enhanced pedagogy, the methodology adopted a mixed-methods research design that combined quantitative metrics (e.g., test scores and engagement levels) with qualitative data (e.g., teacher reflections and student feedback). This approach reflects established best practices in educational research. Creswell and Plano Clark [27] argue that triangulating multiple data sources strengthens the validity and depth of findings. Similarly, Demszky et al. [29] integrated randomized controlled trials with classroom observations and teacher interviews to evaluate an AI teaching tool, demonstrating not just statistical improvements in student outcomes but also qualitative insights into instructional practices. The CyberEd-AI methodology reported comparable findings—quantitative gains in cybersecurity knowledge and qualitative evidence of heightened student motivation. This echoes Johnson, Onwuegbuzie, and Turner's [28] assertion that mixed-methods research offers a more holistic view of complex educational interventions. Taken together, these findings validate the proposed methodology approach and reinforce the relevance of combining AI-powered content development with robust, multidimensional evaluation strategies.

5. Conclusion

The proposed methodology provides a research-backed, scalable framework for integrating Artificial Intelligence (AI) tools into cybersecurity education. By leveraging a Participatory Action Research (PAR) approach, this study demonstrates how AI-driven methodologies can enhance student engagement, differentiated instruction, teacher collaboration, and curriculum scalability.

Findings suggest that interactive AI-driven learning experiences not only improve student motivation, but also teacher collaboration through AI-supported peer review

strengthens their professional development. Furthermore, AI's ability to support adaptive differentiation ensures equitable access to cybersecurity knowledge, benefiting both novice and advanced learners. The study also highlights AI's role in expanding Open Educational Resources (OERs), contributing to the long-term sustainability of AI-enhanced lesson planning.

The results of this study offer practical implications for educators, policymakers, and curriculum designers. First, it underscores the importance of AI integration in digital literacy and cybersecurity education, highlighting AI's potential to bridge knowledge gaps in an increasingly technology-driven world. Second, it demonstrates how collaborative AI-assisted curriculum development can foster teacher professional growth, making AI an essential tool for pedagogical innovation. Finally, it provides a model for scalable, open-access educational content, ensuring broad dissemination of high-quality, adaptable cybersecurity resources.

Future research should explore longitudinal studies assessing the long-term impact of AI-driven cybersecurity education on students' digital resilience. Expanding AI-driven differentiation beyond cybersecurity to other STEM disciplines could also provide valuable insights into AI's broader educational potential.

Acknowledgments. The research presented in this paper was supported by the project SHIELD: Practical Learning through Game Simulation to Improve Cybersecurity Skills, funded by the e-Governance Academy (eGA), under contract number 2-11/5E-2024.

Disclosure of Interests. The authors have no competing interests.

References

1. Videnovik, M., et al., Using peer-learning and game-based instruction for achieving long-lasting knowledge of cybersecurity in primary schools. IEEE Access 13, 11679–11688 (2025). https://doi.org/10.1109/ACCESS.2024.3479921
2. Zaveri, R.: Case study: Implementing cybersecurity education in Baltimore County Public Schools. *J. Cybersecurity Educ.* **12**(3), 95–110 (2020)
3. Simmons, R.T., Park, J.S.: Innovating cybersecurity education through AI-augmented teaching. In: *Proc. 23rd Eur. Conf. Cyber Warfare Secur. (ECCWS 2024)*, pp. 476–482. Academic Conferences International (2024)
4. Pesovski, I., et al.,.: Generative AI for customizable learning experiences. *Sustainability* 16(7), 3034 (2024). https://doi.org/10.3390/su16073034
5. Wang, T., Zhou, N., Chen, Z.: CyberMentor: AI-powered learning tool platform to address diverse student needs in cybersecurity education. *arXiv preprint* arXiv:2501.09709 (2025)
6. Holmes, W., Bialik, M., Fadel, C.: *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, Boston (2019)
7. Videnovik, M., Bogdanova, A.M., Trajkovik, V.: Game-based learning approach in computer science in primary education: A systematic review. *Entertain. Comput.* 48, 100616 (2024). https://doi.org/10.1016/j.entcom.2023.100616
8. Clark, C.H., van Kessel, C.: "I, for one, welcome our new computer overlords": Using artificial intelligence as a lesson planning resource for social studies. Contemp. Issues Technol. Teach. Educ. 24(2), 218–230 (2024)

9. Luckin, R., et al.,.: *Intelligence Unleashed: An Argument for AI in Education*. Pearson, London (2016)
10. Idrizi, E., Filiposka, S., Trajkovik, V.: The discourse on learning styles in online education. In: *Proc. 27th Telecommunications Forum (TELFOR 2019)*, pp. 1–4. IEEE, Belgrade (2019). https://doi.org/10.1109/TELFOR48224.2019.8971204
11. Kemmis, S., McTaggart, R., Nixon, R.: *The Action Research Planner: Doing Critical Participatory Action Research*. Springer, Singapore (2014)
12. Voogt, J., Roblin, N.P.: A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *J. Curric. Stud.* 44(3), 299–321 (2012)
13. Zawacki-Richter, O., et al.,.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 39 (2019). https://doi.org/10.1186/s41239-019-0171-0
14. Vygotsky, L.S.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge (1978)
15. Bruner, J.S.: *Actual Minds, Possible Worlds*. Harvard University Press, Cambridge (1986)
16. Hmelo-Silver, C.E., Duncan, R.G., Chinn, C.A.: Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* 42(2), 99–107 (2007)
17. Voogt, J., et al;, Collaborative design as a form of professional development. *Instr. Sci.* 43**,** 259–282 (2015). https://doi.org/10.1007/s11251-014-9340-7
18. Dixon, H., Haigh, M.: Changing mathematics teachers' conceptions of assessment and feedback. *Teach. Teach. Educ.* 25(3), 350–359 (2009)
19. McKenney, S., Reeves, T.C.: *Conducting Educational Design Research*. Routledge, London (2019)
20. Ertmer, P.A., Ottenbreit-Leftwich, A.T.: Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *J. Res. Technol. Educ.* 42(3), 255–284 (2010)
21. Fraillon, J., Schulz, W., Ainley, J.: Preparing for Life in a Digital Age: The IEA International Computer and Information Literacy Study International Report. Springer, Cham (2014)
22. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.* 74(1), 59–109 (2004)
23. Schön, D.A.: The Reflective Practitioner: How Professionals Think in Action. Basic Books, New York (2017)
24. Zeichner, K., Liston, D.: *Reflective Teaching: An Introduction*, 2nd edn. Routledge, New York (2013)
25. Brookhart, S.M.: *How to Use Grading to Improve Learning*. ASCD, Alexandria (2017)
26. Hilton, J.: Open educational resources and college textbook choices: A review of research on efficacy and perceptions. *Educ. Technol. Res. Dev.* 64**,** 573–590 (2016). https://doi.org/10.1007/s11423-016-9434-9
27. Creswell, J.W., Plano Clark, V.L.: *Designing and Conducting Mixed Methods Research*, 3rd edn. Sage Publications, Thousand Oaks (2017)
28. Johnson, R.B., Onwuegbuzie, A.J., Turner, L.A.: Toward a definition of mixed methods research. J. Mix. Methods Res. 1(2), 112–133 (2019)
29. Demszky, D., et al., .: Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. Educ. Eval. Policy Anal. 45(3), 473–498 (2023)

Towards Privacy-Preserving AI in Educational Platforms: Backend Strategies for Secure Data Analytics and Threat Detection

Jovana Trajcheska[1], Ivan Chorbev[1], Dejan Gjorgjevikj[1], and Boban Joksimoski[1]

[1] Faculty of Computer Science and Engineering, University Ss. Cyril and Methodius, Rudzer Boshkovikj 16, P.O. 393, Skopje, 1000, North Macedonia.

Corresponding author(s). E-mail(s): jovana.trajcheska.1@students.finki.ukim.mk, ivan.chorbev@finki.ukim.mk, dejan.gjorgjevikj@finki.ukim.mk, boban.joksimoski@finki.ukim.mk

Abstract. In this paper, we explore how artificial intelligence can be responsibly integrated into educational platforms while maintaining strong protections for student privacy. We begin by identifying key challenges, such as compliance with data protection regulations, the sensitivity of student data, and emerging AI-specific threats such as model inversion and data poisoning. Next, we look at a number of privacy-enhancing methods that can lower privacy risks during model training and deployment, such as federated learning, homomorphic encryption, differential privacy, secure multi-party computation, and synthetic data generation. Furthermore, we prioritize backend methodologies that improve the security and modularity of AI systems, including microservices architectures, encrypted data pipelines, and trusted execution environments. Additionally, we emphasize the importance of access control, anomaly monitoring, and threat detection as integral components of a comprehensive security framework. This paper presents a framework for developing educational AI solutions that are safe, legal, and aligned with responsible innovation.

Keywords: privacy-preserving AI · data security · data sensitivity · homomorphic encryption · microservices · federated learning · data governance · educational platform

1 Introduction

Artificial intelligence (AI), and its rise in recent years, has successfully revolutionized education by enabling customized learning experiences and improving the teaching process. In addition, a growing number of educational platforms incorporate adaptive learning systems, real-time engagement analytics, and AI-based tutoring to improve educational outcomes [1]. In contemporary education, numerous institutions are increasingly using sophisticated digital platforms that also incorporate AI. These systems are designed to improve the efficiency of

administrative operations, monitor student progress, and facilitate data-driven teaching, ultimately making education more efficient and tailored to individual needs. The continued expansion of AI and its involvement in platforms like these raises a multitude of questions related to data security. Schools collect and handle a wide array of sensitive personal data, including academic performance history, biometric identifiers, behavioral analytics, health indicators, student behavior, and similar information, intensifying the need to protect this data. This situation renders the EdTech sector especially appealing to cybercriminals. Privacy breaches affecting children constitutes a significant violation that can create substantial risks to their safety, highlighting why both parents and schools must treat student data protection as a fundamental obligation [2].

The integration of AI into educational platforms, as well as into the software used by educational institutions, must be carefully done, paying attention to how student data is collected and processed, and subsequently stored and protected. In the United States, the Family Educational Rights and Privacy Act (FERPA) [3] grants parents and eligible students the right to access and control their educational records. The law strictly limits the release of personally identifiable information (PII) from student records without consent. Moreover, online services, including educational platforms, should obtain parental consent before gathering, processing, and distributing personal information regarding children under age 13 years, which is mandated by the Children Online Privacy Protection Act (COPPA) [4]. The European Union enforces data protection standards through the General Data Protection Regulation (GDPR), which regulates the collection, processing, and retention of personal data. The regulation emphasizes the need to collect only essential data, requires a clear legal basis for processing, and grants users the right to request the deletion of their personal information [5]. Failure to comply with legal obligations can result in serious legal consequences, but also in a significant reduction in trust in the institution. Hence, the EdTech industry has an increased responsibility to comply with all relevant legal regulations and to ensure a high level of personal data protection, both for students in formal educational institutions and for users of online learning platforms.

Concerns about the misuse of children's personal data are growing, but are not unfounded. Embracing AI in education poses a significant challenge, as it requires leveraging advanced analytics and threat detection tools while preserving student privacy. Past incidents have shown that even without AI, systems are vulnerable, and integrating AI can be an additional challenge. Educational platforms must move toward architectures that incorporate privacy by design, data anonymization techniques, and tightly control access. Without fundamental principles, the use of AI will not only erode trust, but also expose institutions to technical and ethical risks.

The objective of this paper is to analyze approaches for integrating AI into these platforms in a manner that is both ethically responsible and compliant with existing legal aspects. The focus is on backend strategies that protect student data. The paper discusses techniques like federated learning, differential privacy,

and synthetic data, as well as architectural choices such as microservices, secure APIs, and encrypted data handling.

2 Core Challenges in Building Privacy-Preserving AI for Education

The involvement of AI in the EdTech industry complicates an already complex web of security risks. The systems collect vast amounts of personal and academic data from students, which are subject to new opportunities for misuse and privacy violations. Although we are witnessing an era where AI improves education and the education system, the burden it brings with its integration significantly increases the responsibility for protecting personal data, requiring precise security measures and compliance with legal aspects and privacy regulations [6].

2.1 Educational Data Sensitivity and Child Protection

The systems used in educational institutions collect and process substantial volumes of highly sensitive and personally identifiable data. This includes various academic achievements, grades, test scores, class attendance, and sometimes even data about parents, as well as biometric data collected for proctoring or personalization purposes [7]. AI uses the collected data to predict a certain result for students, whether it is potential results on a certain test, success in a certain subject, the risk of dropping out of school, and so on. This is typically intended to support early interventions and tailored educational assistance in order to prevent bad consequences, such as dropping out of school, and provide real-time assistance or interventions [8].

Despite the intended benefits, these systems raise considerable ethical and security concerns. Improper storage of data and their insufficient protection can bring a series of catastrophic consequences and the data can be subject to violations of individual privacy, unauthorized sharing, or analytics outputs. For instance, model inversion and membership inference can reveal individual student records from trained models [9]. Unsecured behavioral data can lead to identity theft or profiling by third parties. Additionally, over-collection and misuse of biometric data or engagement records risk long-term psychological harm, stigmatization, or discrimination—particularly when algorithms are unregulated or biased [7,8] .Recent research points to the fact that it is essential to manage data transparently and to improve privacy-enhancing techniques such as anonymization, synthetic data generation [10,11], differential privacy, and active consent mechanisms, in order to remain within ethical frameworks when using AI to achieve a positive outcome and thus to keep students and their privacy protected.

2.2 AI-Specific Threat Vectors

The integration of AI into educational platforms and systems used in educational institutions introduces a range of new security vulnerabilities that extend beyond the well-known traditional cyber threats. The AI models trained on sensitive student data can succumb to threats distinct from traditional ones against privacy and integrity of the system. One known such threat is the *model inversion attack*, in which attackers attempt to reconstruct the input data, in this case student data by exploiting access to the trained model's outputs or parameters [12]. Closely related are *membership inference attacks*, which allow attackers to infer the presence of an individual's data within the model's training data, thereby leaking private information [13].

Although *data poisoning attacks*, in which attackers inject malicious or manipulated training examples, pose a genuine threat to AI systems, they are not exclusive to EdTech. Instead, what makes educational platforms specific are the attack surfaces and outcomes (such as numerous student submissions, grades, intervention recommendations, placement decisions) that facilitate or increase the impact of poisoning in real-world scenarios. This can cause change in the AI system's behaviour in some way to biased or harmful ends for personalized learning risk prediction [14,15]. Therefore, although data poisoning is a general ML threat, EdTech systems require targeted ingestion controls, provenance tracking, and good aggregation strategies to mitigate this risk. Possible measures include requiring verified uploads and logging their origin, automatically flagging suspicious records, using aggregation methods that ignore bad updates, and limiting how much any single example can change the model while keeping a trusted validation set and audit logs.

These new types of attacks coupled with the usual suspects of cybersecurity vulnerabilities — fraudulent access and denial of service, to name a couple — weave a complex threat landscape that demands a more encompassing security approach [16].

2.3 Regulatory and Compliance Challenges

In addition to technical protections, education systems face a complex framework of legal aspects that they must comply with. As previously mentioned, the US, FERPA and COPPA impose strict restrictions on the use of student data, so FERPA, for example, gives parents and students full control over their personal information, COPPA requires parental approval before obtaining personal information from individuals under 13. The EU's GDPR operates in a similar way, requiring the collection of as little personal information as possible, limiting the purposes for which it is used, and granting the right to delete all personal data. In addition to this, a new EU law is being added, according to which artificial intelligence is marked as *high risk* when used in systems used for educational purposes [17]. These regulations require educational platforms

to implement technical and organizational measures, such as obtaining consent, providing access to parents or their consent as required by law, defining clear policies on how data is used, and ensuring auditability. Failure to comply with these legal segments can result in reduced trust and can also lead to heavy fines.

In addition, regulations are more and more calling for accountability and fairness in AI decisions. The systems for analyzing student performance and behavior with artificial intelligence should be free from discrimination on any base and they should provide the explanation of a result. For example, EU lawmakers insist AI be transparent, non-discriminatory and human-supervised [17]. In practice, this means institutions must document AI decision criteria, allow human review of sensitive decisions, and be prepared to justify the use of any automated profiling. Globally, there is no single standard; many jurisdictions are developing or updating privacy and AI guidelines (for example, updated COPPA rules, state privacy laws like California's CPRA, or UNESCO's recommendations on AI ethics in education). Consequently education institutions need to have compliance teams, or better still, data governance boards, which can interpret these changing requirements, and align them to system designs. So in conclusion, a strict "privacy-by-design" regulation framework is required for EdTech - policies, consent forms, and audit trails as critical as security and anonymization algorithms.

2.4 Data Governance and Ethical Concerns

A strong foundation for any privacy strategy lies in establishing clear data governance policies and roles. Institutions are encouraged to appoint dedicated personnel, such as Data Protection Officers and ethics committee members, to oversee how student data is collected, stored, and utilized. This process involves defining comprehensive data classification frameworks and enforcing precise access controls. For instance, a university might maintain detailed inventories of student data categories, assign specific access rights to various stakeholders, and implement automated tools to track data lineage across systems. Such governance frameworks emphasize maintaining data quality, ensuring security, meeting regulatory requirements, and upholding ethical standards simultaneously. Many campuses employ interdisciplinary committees—bringing together IT professionals, legal advisors, academic representatives, and student advocates—to carefully vet data access and AI applications. Before deploying an AI tutor, for example, the team might verify that only essential pseudonymized or encrypted student attributes are incorporated. This transparency enables organizations to provide clear accountability to regulators, students, and their families.

Equally important are the ethical considerations intertwined with governance. AI systems designed for educational settings must actively mitigate bias and protect student rights. Research on algorithmic fairness highlights risks such as discriminatory grading practices against non-native speakers or reinforcing gender stereotypes if models are not carefully audited [18][19]. To address these issues, institutions should adopt bias mitigation strategies, including curating

balanced training datasets and implementing fairness-aware algorithms, coupled with continuous monitoring of AI outputs. For example, predictive systems estimating dropout risk need ongoing audits to ensure they do not disproportionately affect disadvantaged populations. Transparency remains a cornerstone; students and parents should be informed about AI's role in decision-making processes, and clear documentation of AI capabilities and limitations should be made accessible. In conclusion, robust governance blends technical measures—such as encryption and anonymization—with human oversight and institutional policies committed to fairness, privacy, and transparency.

3 Privacy-Preserving AI Techniques

3.1 Federated Learning

Federated learning (FL) has emerged as a powerful approach for training models across institutions without pooling raw data. In FL, each client (e.g., a school or classroom) trains a local model on its own student data and only shares model updates with a central server or aggregator [20]. This architecture inherently improves data privacy: sensitive records never leave their origin. For example, Latif et al. demonstrate a federated framework for automated grading of educational assessments, where each school fine-tunes a local model and only communicates optimized model updates to a central aggregator [21]. Their approach achieved accuracy on par with a centralized model while eliminating the need to expose any student exam answers. By combining FL with techniques like parameter-efficient fine-tuning and adaptive aggregation, such systems can handle variations in data across schools.

Despite these benefits, federated learning is not immune to attacks. Adversaries who intercept model parameters or participate as malicious clients can attempt to reconstruct private data through model inversion or membership inference attacks [22]. In educational contexts, this could mean inferring whether a particular student's record was part of the training set, or even recovering sensitive attributes (e.g., grades or personal information) from model weights. To mitigate this, FL is often combined with privacy enhancements: for instance, adding differential privacy noise to updates, using secure aggregation protocols, or limiting the number of shared updates. FedFLow projects use cryptographic aggregation or clip-and-noise mechanisms so individual updates reveal minimal information. Overall, FL offers a scalable path to multi-school AI, but it must be coupled with additional safeguards (secure channels, authentication of clients, and differential privacy) to ensure that the distributed training process does not leak student data.

3.2 Differential Privacy

Differential privacy (DP) provides a mathematically rigorous way to quantify and limit privacy loss when performing analytics. Under DP, carefully calibrated random noise is added to query results or model updates so that the

presence or absence of any single individual in the data has only a bounded impact on outputs. For educational data, DP can protect students in aggregate analyses or shared models. For example, recent work by Liu et al. introduces a DP framework tailored to learning analytics, showing how noise injection can defend against reconstruction attacks while still preserving enough utility for educational research [11]. Their study also explores the trade-off between data utility and privacy budget: smaller privacy parameters improve confidentiality but may degrade analysis accuracy.

In practical terms, DP can be applied at multiple stages. A school collecting learning analytics (e.g., clickstreams or quiz scores) might apply DP when publishing reports or sharing insights. In federated learning, clients can add noise to their gradient updates before sending them. Libraries like Google's TensorFlow Privacy or PyTorch Opacus facilitate training deep models with DP by clipping gradients and adding Gaussian noise. Importantly, DP assumes trust only in the data curator who sets the privacy budget; it does not protect against all attacks (e.g., outputs themselves might still be exploited if the noise is weak). Nonetheless, DP is increasingly seen as a necessary component in educational AI to meet legal expectations (it aligns with GDPR's *data protection by design* concept) and to offer formal guarantees that no single student's data is being inadvertently exposed in collective outputs.

3.3 Homomorphic Encryption

Homomorphic encryption (HE) lets you do math directly on encrypted data, so the results stay encrypted until you use a secret key to decrypt them. This means that data can stay private even while it is being processed. For example, an educational platform could encrypt students' records and still run analytics or model inference on them without ever turning them back into plain text. HE schemes, like CKKS or BFV, are made so that you can do math on ciphertexts. Early systems, such as CryptoNets, showed that it was possible to do neural network inference on inputs that were homomorphically encrypted [23]. A system like this can make predictions about encrypted student performance data. It only sends back the encrypted predictions, and only people with the key can see the final results.

The biggest problem with HE is that it is not very efficient because fully homomorphic encryption takes a lot of computing power. Most real-world systems use leveled or partial HE, which means that the functions that can be computed before decryption is needed are less complicated. Because of this, HE is used more for inference than for training. One possible use of AI in education is a cloud-based inference service. Students' inputs could be encrypted locally and sent to a school's AI model in the cloud, which would perform inference homomorphically and send back an encrypted output. This would make sure that the cloud provider never sees any raw student data. Researchers are still working on making HE better for machine learning. It's not yet widely used in production, but it does offer a long-term privacy guarantee: *the computation on encrypted*

data means that an attacker who sees intermediate states won't learn anything about the underlying student records [23].

Recent studies on HE for federated settings demonstrates practical ways to reduce HE overhead while preserving privacy: Pan et al. [24] propose an adaptive segmented-CKKS design that rearranges ciphertext batching and computation to improve throughput and lower communication cost compared with a straightforward CKKS deployment . A standardized benchmarking study of modern FHE libraries further clarifies which schemes and backends are preferable for arithmetic versus bit-level workloads, and highlights the latency/ciphertext-size trade-offs practitioners must weigh [25]. For our pipeline this implies that deploying HE at scale will require careful choice of scheme (e.g., CKKS variants), batching/segmentation strategies and parameter tuning, and where possible leveraging optimized backends or hardware acceleration to keep latency and bandwidth within acceptable bounds.

3.4 Secure Multi-Party Computation

Secure Multi-Party Computation (MPC) is a cryptographic paradigm that enables multiple parties to jointly compute a function over their inputs while preserving the privacy of each party's data. Unlike traditional approaches that require central aggregation, MPC ensures that no participant learns anything beyond the final output. Each party locally holds confidential data and engages in a protocol that securely computes the desired function. Techniques such as secret sharing, homomorphic encryption, and garbled circuits are commonly used to realize these protocols. For instance, the CrypTen framework demonstrates how MPC can be integrated into machine learning workflows, allowing institutions—such as two schools—to collaboratively train models without ever sharing raw student data [26].

MPC protocols can be applied to both one-time secure computations and more complex privacy-preserving systems, such as those used in federated learning. A typical example involves computing class-wide statistics, like the average test score, from encrypted student results without revealing any individual's score. Recent research has shown that MPC can scale effectively to support training of neural networks and other complex models. This is accomplished by combining efficient linear algebra operations with underlying secret-sharing techniques, and optimizing protocol execution through cryptographic subroutines such as oblivious transfer and homomorphic encryption. These enhancements significantly reduce computational overhead while maintaining strong privacy guarantees.

One of the most important benefits of MPC lies in its strict adherence to data confidentiality. No party can infer another's input unless a predefined number of parties collude, which provides strong security assurances in sensitive or regulated environments. However, the practical deployment of MPC is not without challenges. It generally requires all participating entities to be online throughout the computation, and introduces additional communication costs due to the

exchange of encrypted data. Despite these limitations, MPC remains a powerful tool in collaborative educational settings—particularly in consortiums where institutions wish to leverage combined datasets for analytics or machine learning without compromising student privacy.

Applied studies indicate that MPC-based aggregation can be added to learning-analytics pipelines with only modest runtime and communication overhead when protocols and networking are tuned appropriately [27]. These works find that MPC preserves analytical utility while protecting per-user data, so for our back-end we lean toward MPC (or secure-aggregation variants) for aggregation tasks that would make homomorphic computation too expensive. In both cases we would focus engineering effort on efficient communication patterns and simple validation checks (trusted holdout sets, sanity checks) to keep overheads low and ensure result integrity.

3.5 Synthetic Data Generation

Synthetic data generation is a technique used to create artificial datasets that preserve the utility of real student data while eliminating the presence of actual personal information. By training generative models on real data, one can sample synthetic records that preserve overall statistical properties (such as grade distributions or activity patterns) without including any real student's record. This technique enables sharing of "data" with researchers or third-party developers without exposing PII. For example, Vie *et al.*[10] developed a generative model for educational data that can produce realistic student trajectories while protecting participant privacy . They emphasize that naive approaches (like simply pseudonymizing names) are vulnerable to re-identification, but careful generative models combined with privacy checks can mitigate those risks.

In an educational context, the use of synthetic data enables the supplementation of sparse datasets or the distribution of data for model benchmarking. For instance, an online learning platform could publish a synthetic copy of its user interaction logs for research. Similarly, school districts could share synthetic achievement data to analyze trends without revealing individual records. When properly validated, synthetic data may satisfy regulators that identifiable details are absent, while still providing useful signals for training or analysis. Of course, synthetic generation must itself be done cautiously: the training process should ensure that no single student's record is memorized by the generator, which could lead to privacy leaks. Advanced methods often combine differential privacy with synthetic generation to bound such leakage. Overall, synthetic data is a promising tool to enable privacy-preserving research and development in education, but it is typically used alongside (not in place of) other safeguards like encryption and access control [10].

In the preceding subsections, each privacy-preserving method was discussed in depth. The following table compiles these approaches into a method to deployment stage overview, emphasizing the main advantage and a sample use case to guide implementation choices in educational AI systems.

Privacy Technique	Deployment Stage	Key Benefit	Typical Use Case
Federated Learning	Training	No raw data sharing	Multi-school model collaboration
Differential Privacy	Analytics/Training	Provable privacy guarantee	Publishing aggregated learning analytics
Homomorphic Encryption	Inference	Encrypted computation	Cloud-based predictions on student records
Secure MPC	Joint Computation	Strong confidentiality among peers	Cross-institutional aggregate statistics
Synthetic Data	Benchmarking	Safely share data distribution	Public research dataset release

Table 1. Privacy techniques in EdTech: stages, benefits, and typical use cases.

4 Microservices Architecture for AI-Enhanced Data Protection

Microservices architecture can enhance security by dividing the educational platform into isolated services. For instance, user authentication, data ingestion, model training, analytics, and threat monitoring may be handled by distinct containers or microservices. Each service has a single responsibility and exposes a very small surface area of the data. Inter-service communication takes place between services with explicit APIs protected by mutual authentication and encryption (e.g., HTTPS/TLS). Network's segmentation (service mesh or Kubernetes network policies) restricts lateral movement and if one service is broken into, an attacker can't easily jump to others. Furthermore, each microservice may have its own security policy and may be audited and deployed independently (e.g., a data-processing microservice may mandate in-memory encryption of sensitive fields, whereas a logging microservice might simply log all requests at lower levels).

It is essential to pay close attention when managing APIs because microservices themselves expose numerous APIs. If an endpoint fails to authenticate callers or validate inputs, it could easily become a target for attacks [28]. To prevent this, it's important to implement strict input schemas, enforce rate limiting on all service-to-service communications, and use robust API keys or OAuth tokens. System security is further enhanced by container-focused measures, including limiting container privileges, selecting minimal base images, and incorporating continuous vulnerability scanning tools. The integration of DevOps with security integrity plays an important role and is of great importance. Continuous integration pipelines include automated testing frameworks and security analysis tools (linting) to prevent the introduction of compromised credentials or potentially harmful dependencies. By reducing human error and maintaining security standards at scale, this automated approach ensures seamless execution of security validations throughout all phases of deployment. Figure 1 illustrates

a sample microservices layout: for instance, a secure AI model service might only receive already-encrypted student data from a preprocessing service, and only output encrypted results to a results aggregator. Overall, microservices provide flexibility to embed privacy controls at granular levels, simplifying compliance and reducing the impact of any single service breach.

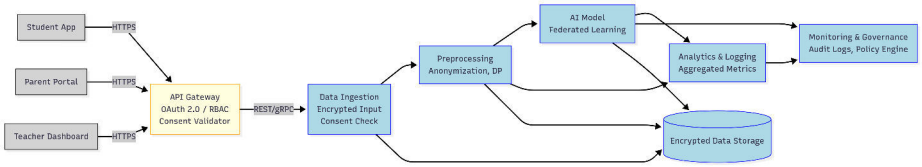

Fig. 1. Illustrative microservices architecture for privacy-preserving educational AI. Each box represents a containerized service (e.g., Data Ingestion, Preprocessing, Analytics, AI Model, Monitoring) communicating over encrypted APIs.

5 Privacy-Aware AI Processing Pipeline

To ensure regulatory compliance and robust protection of student privacy, our AI processing pipeline emphasizes data minimization from the very beginning. During data collection, only essential attributes are captured, and personally identifiable information (PII)—such as names or IDs—is immediately pseudonymized or tokenized [29]. Data ingestion enforces that any PII is removed before storage in encrypted, access-controlled databases using AES-256. In subsequent preprocessing stages, anonymization techniques—like stripping direct identifiers and generalizing demographics—are applied, resulting in datasets suitable for privacy-preserving analytics or training models with federated learning, differential privacy, or synthetic data [29].

For inference, the pipeline secures student inputs and model outputs with end-to-end protection. Student behavior data (e.g., quiz responses, learning patterns) is transmitted through encrypted channels into a Trusted Execution Environment (TEE). Predictions—such as risk scores or resource recommendations—are returned using techniques that minimize disclosure, for example by output masking or limiting results to class labels rather than full probability distributions. This approach safeguards against inference attacks while maintaining actionable feedback. System logs capture only metadata—timestamps or encrypted identifiers—while individual student data remains confidential under strict access control [29].

Governance underpins each transition within the pipeline. Explicit consent flags are verified before any data processing occurs, and automated retention polices purge outdated records. Audit mechanisms meticulously document who accessed which data, when, and for what purpose. By interweaving encryption

(both at rest and in transit), anonymization/sanitization processes, privacy-enhancing techniques (FL, DP, MPC, synthetic data), and strong governance, the pipeline upholds stringent privacy guarantees across all stages—from collection to analytics and inference [29].

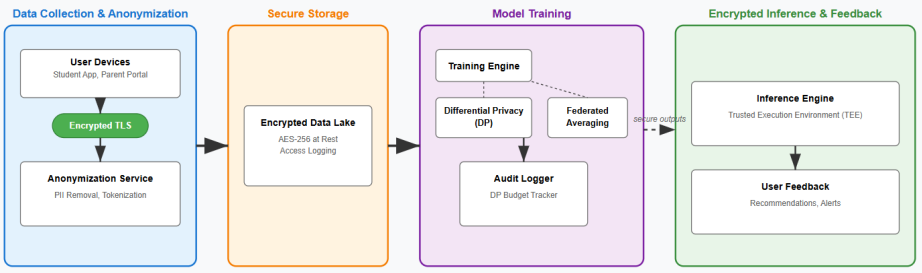

Fig. 2. Example privacy-aware AI pipeline: data collection with anonymization, secure data storage, differential-privacy-protected model training, and encrypted inference (e.g., via a Trusted Execution Environment).

6 Secure AI Model Deployment and Inference

Securing AI model deployment involves safeguarding both the intellectual property encapsulated in model weights and the integrity of runtime inference processes. Model files, which may encode sensitive student data or proprietary logic, should be encrypted at rest and digitally signed to prevent tampering or unauthorized modifications. Containerized deployments must follow security best practices: minimal base operating system images, execution with non-root privileges to avert escalation, continuous vulnerability scanning, and automated patching pipelines to apply updates promptly.

Inference-phase security is equally vital. Exposed models via APIs should enforce authenticated access, such as service accounts or tokens, and implement rate limiting to thwart brute-force or membership-inference attacks. A robust approach leverages Trusted Execution Environments (TEEs). For example, the *Slalom* framework partitions neural network execution by running linear layers within an enclave (e.g., Intel SGX) and offloading heavy computation externally. This protects both the model and input data through enclave memory isolation against OS-level threats [30]. Further safeguards include output perturbation techniques such as MemGuard, which add calibrated noise to confidence scores to mitigate membership inference attempts [31].

A final layer of defense is anomaly detection: profiling deployed models for unusual usage patterns or distributional shifts. Sudden spikes in query volume, particularly for specific student profiles, may signal an attack or exploit attempt. Techniques such as SHAP or LIME can be used to audit model outputs, helping

to identify bias or unexpected behavior, enhancing transparency and trust. By combining encrypted model storage, secure container practices, authenticated service, enclave-based inference, and output monitoring protocols, educational platforms can deliver AI services that balance functionality with strong student data protection.

7 Empirical Evaluation and Comparative Analysis

Research has recently shown that privacy-preserving AI methods are efficient enough to be applied in education, although there are some performance trade-offs involved [32]. Using three significant educational datasets (OULAD, EdNet, and KDD Cup 2015), Van Haastrecht et al. (2024) showed that federated learning maintains full data locality while achieving accuracy levels comparable to centralized methods [32].

The analysis showed that federated learning performed significantly better than local learning approaches, achieving higher accuracy on OULAD and notable AUC gains on EdNet, while introducing the Federated Learning Analytics Metric (FLAME) to capture the privacy–performance balance; FLAME indicated that around 50 clients is a strong operating point for large educational datasets [32]. Building on this direction, Liu et al. (2025) developed DEFLA, a differential privacy framework tailored for learning analytics, and validated it on OULAD with membership inference evaluations to demonstrate privacy protection under different DP configurations [11].

Federated learning experiments further indicate that distributing training across many clients can keep raw data on local devices while incurring only modest utility loss relative to a central model, a trade-off that FLAME makes explicit for educational datasets like OULAD and EdNet [32]. In parallel, DEFLA's experiments on OULAD compared input-, training-, and output-stage differential privacy, finding that input perturbation often preserves utility better while output-stage perturbation can deliver strong privacy–utility trade-offs under conservative settings; privacy leakage was assessed via membership inference risk [11]. Taken together, these results support the view that privacy-preserving methods, paired with modular architectures and encrypted data flows can sustain real-world accuracy needs for EdTech while aligning with data protection requirements [32,11].

8 Conclusion

Integrating AI into education requires a comprehensive backend strategy centered on *privacy by design*. It is not enough to layer technical protections—such as encryption, differential privacy, federated learning, and secure enclaves—without embedding them within strong governance and regulatory frameworks. Modular, microservice-based architectures enable secure APIs to mediate analytics

and inference, while tightly controlled data flows maintain oversight. Privacy-enhancing technologies like secure enclaves and synthetic data generation reduce exposure of real records. Compliance with frameworks such as FERPA, COPPA, and the EU AI Act ensures institutions respect students' legal rights.

Ultimately, safe educational AI demands security at every level, from data collection and model training to deployment and real-time inference. The methods described here provide a blueprint: privacy-preserving machine learning, coupled with cloud-native secure deployment, enables trust, fairness, and accountability in systems designed for young learners. The continued evolution of EdTech underscores the need for close collaboration between technologists, educators, and policymakers to advance these solutions and place student privacy at the forefront.

Acknowledgments

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

References

1. Çela, E., Fonkam, M. M., Vajjhala, N. R., Vedishchev, A.: Artificial Intelligence in Education: Foundations, Trends, and Implications. In: Next-Generation AI Methodologies in Education, pp. 1–20 (2024).
2. Alamleh, H.: Private and Secure Students' Data Sharing in Educational Systems. In: 2020 Sixth International Conference on e-Learning (econf), Sakheer, Bahrain, pp. 158–161 (2020).
3. "FERPA | Protecting Student Privacy." Accessed: Sep. 10, 2025. [Online]. Available: https://studentprivacy.ed.gov/ferpa
4. Children's Online Privacy Protection Rule: Final Rule Amendments (2012-31341), U.S. Federal Trade Commission. `https://www.ftc.gov/system/files/2012-31341.pdf`. Accessed: 2025/06/25.
5. General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, EUR-Lex, `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, Accessed: 2025/06/25.
6. Dhiman, T., Chauhan, V., Kumar, A., Vasantha, M., Kumar, A.: Ethical Crossroads: Navigating Data Privacy, Bias, Accountability and Sustainability in AI-Driven Education. Open Access Journal of Multidisciplinary Research (OAJMR), **1**(3), 69–77 (2025).
7. Shrestha, A., Barthwal, A., Campbell, M., Shouli, A., Syed, S., Joshi, S., Vassileva, J.: Navigating AI to Unpack Youth Privacy Concerns: An In-Depth Exploration and Systematic Review. arXiv preprint arXiv:2412.16369 (2024).
8. Campbell, M., Barthwal, A., Joshi, S., Shouli, A., Shrestha, A.K.: Investigation of the Privacy Concerns in AI Systems for Young Digital Citizens: A Comparative Stakeholder Analysis. (2025).
9. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. (2018).

10. Vie, J.-J., Rigaux, T., Minn, S.: Privacy-Preserving Synthetic Educational Data Generation. In: Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning (EC-TEL 2022), Toulouse, France, September 12–16, 2022, Proceedings, pp. 393–406. Springer, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16290-9_29
11. Liu, Q., Shakya, R., Khalil, M., Jovanovic, J.: Advancing privacy in learning analytics using differential privacy. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25), pp. 181–191. ACM, New York, NY, USA (2025). https://doi.org/10.1145/3706468.3706493
12. Fredrikson, M., Jha, S., Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), pp. 1322–1333. ACM, New York, NY, USA (2015).
13. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership Inference Attacks against Machine Learning Models.
14. Steinhardt, J., Koh, P.W., Liang, P.: Certified Defenses for Data Poisoning Attacks.
15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. arXiv 1412.6572
16. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P.: SoK: Security and Privacy in Machine Learning. In: Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), pp. 399–414. IEEE (2018).
17. EU AI Act: first regulation on artificial intelligence Press release (2023). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2025/06/26.
18. Kizilcec, R. F., & Lee, H. (2020). Algorithmic Fairness in Education.
19. Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications. arXiv:2407.18745.
20. Yu, D., Zhang, X., He, H., Chen, S., Qiao, J., Wang, Y., Cheng, X.: Robust Federated Learning for Edge Intelligence. In: Thai, M. T., Phan, H. N., Thuraisingham, B. (eds.) Handbook of Trustworthy Federated Learning, Springer Optimization and Its Applications, vol. 213 (2025).
21. Latif, E., Zhai, X.: Privacy-Preserved Automated Scoring using Federated Learning for Educational Research.
22. Jalil Piran, F., Chen, Z., Imani, M., Imani, F.: Privacy-Preserving Federated Learning with Differentially Private Hyperdimensional Computing.
23. El Mestari, S. Z., Lenzini, G., Demirci, H.: Preserving data privacy in machine learning systems. Computers & Security, **137**, 103605 (2024).
24. Pan, Y., Chao, Z., He, W., Jing, Y., Li, H., Wang, L.: FedSHE: privacy preserving and efficient federated learning with adaptive segmented CKKS homomorphic encryption. Cybersecurity 7, 40 (2024).
25. Gouert, C., Mouris, D., Tsoutsos, N.G.: SoK: New Insights into Fully Homomorphic Encryption Libraries via Standardized Benchmarks. Proceedings on Privacy Enhancing Technologies (PoPETs) 2023(3), 154–172 (2023).
26. Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., van der Maaten, L.: CRYPTEN: Secure Multi-Party Computation Meets Machine Learning. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp. 11774–11787 (2021).

27. Rodriguez-Garcia, M., Balderas, A., Dodero, J.M.: Privacy Preservation and Analytical Utility of E-Learning Data Mashups in the Web of Data. Applied Sciences 11, 8506 (2021).
28. Narváez, D., Battaglia, N., Fernández, A., Rossi, G.: Designing Microservices Using AI: A Systematic Literature Review. Software, **4**(1), Article 6 (2025).
29. Tsoni, A., Masiello, I., & Martins, R. M. (2021). A Data Pipeline to Preserve Privacy in Educational Settings. Educational Technology & Society.
30. Tramèr, F., & Boneh, D. (2018). Slalom: fast, verifiable and private execution of neural networks in trusted hardware.
31. Jia, J., Salem, A., Backes, M., Zhang, Y., & Gong, N. Z. (2019). MemGuard: defending against black-box membership inference attacks via adversarial examples.
32. van Haastrecht, M., et al.: Federated Learning Analytics: Investigating the Privacy–Performance Trade-off in Educational Data (OULAD, EdNet, KDD Cup 2015). Open-access working paper, Leiden University (2024).

Session 5

Evaluating LLMs on the Extractive Question-Answering Task in Macedonian

Stefan Milev[0009−0007−5623−1915], Monika Simjanoska Misheva[0000−0002−5028−3841], and Kostadin Mishev[0000−0003−3982−3330]

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
stefan.milev.1@students.finki.ukim.mk, monika.simjanoska@finki.ukim.mk, kostadin.mishev@finki.ukim.mk

Abstract. This paper assesses the capabilities of 17 large language models (LLMs) in extractive question-answering (extractive QA) task in the Macedonian language. We evaluate LLMs, from closed-source commercial systems and open-source alternatives to specialized models, using standard machine translation metrics, BLEU, chrF, METEOR, as well as COMET, to assess semantic adequacy and fluency. Our evaluation covers both standard Macedonian and dialectal variants, providing insights into model performance across linguistic variants. Results demonstrate that closed-source models, particularly GPT-4.1 Mini, achieve superior performance with a COMET score of 0.831, while the open-source Llama-4 Maverick achieves competitive results with a COMET score of 0.830. Remarkably, most models show good performance on dialectal variants as well as standard Macedonian, suggesting significant potential for specialized applications in dialectal language processing. Our findings provide crucial insights for practitioners working with low-resource languages and highlight the capabilities of modern LLMs for them.

Keywords: Large Language Models · Macedonian Language · Model Evaluation · Macedonian dialects · Low-Resource Languages

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing [2], demonstrating remarkable capabilities across numerous tasks and languages. However, their performance in low-resource languages, particularly for specialized tasks such as answering questions and extracting information, remains relatively underexplored [7, 8]. This gap is particularly pronounced for South Slavic languages, which, despite their rich linguistic heritage and substantial speaker populations, often lack comprehensive evaluation studies using modern LLM architectures.

The Macedonian, a south Slavic language spoken by approximately 2 million people primarily in North Macedonia, presents an ideal case study for evaluating LLM capabilities in low-resource language settings. The language exhibits

complex morphological features typical of Slavic languages, including rich inflectional systems and relatively free-word order, which pose challenges for language models primarily trained on high-resource languages like English.

The emergence of both closed-source commercial models (such as GPT-4 variants and Gemini models) and open-source alternatives (including Llama [11], Mistral, Qwen, Gemma, and others) has created an opportunity to conduct systematic comparisons across different model architectures, sizes, and training approaches. While similar work exists for other low-resource languages such as Turkish [8] or Tigrinya [14], no equivalent study exists for Macedonian. To support the community, we also release our dataset openly on Hugging Face[1].

This study addresses several critical research questions. Specifically, we ask how different categories of LLMs perform in Macedonian extractive QA, what differences can be observed between closed-source and open-source systems, how model size correlates with performance, and how robust models are with respect to dialectal variation. Although answers are provided to all four questions, we stress that dialectal conclusions are preliminary given the comparatively small sample.

Our contributions can be summarized as follows. We present the first publicly released dataset for Macedonian extractive QA, covering both standard and dialectal variants. We evaluate a wide range of state-of-the-art models, both open and closed, and offer empirical insights into their strengths and weaknesses. Finally, we provide a multi-dimensional analysis of accessibility, scale, and dialectal robustness, contributing important lessons for low-resource NLP.

2 Related Work

2.1 Extractive Question Answering Fundamentals

Extractive question answering, where systems identify text spans within provided contexts that answer posed questions, has become a cornerstone evaluation task for language understanding. The Stanford Question Answering Dataset (SQuAD) established the primary benchmark for this task, consisting of over 100,000 questions created by crowdworkers on Wikipedia articles [1]. BERT-based approaches have dominated extractive QA performance, with models achieving human-level performance on benchmark datasets [2]. The typical approach involves fine-tuning BERT models to predict start and end positions of answer spans within context documents [3].

Recent work has explored improvements to span-based extraction methods. SpanBERT introduced span masking during pre-training and achieved 94.6% and 88.7% F1 on SQuAD 1.1 and 2.0 respectively, demonstrating substantial gains on span selection tasks [4]. These foundational approaches provide the methodological framework for evaluating extractive capabilities across different model architectures.

[1] `https://huggingface.co/datasets/Delemangi/llms-evaluation`

2.2 Multilingual and Cross-lingual Question Answering

The evaluation of question answering systems across languages has received increasing attention, particularly for low-resource languages. The MLQA benchmark offers cross-lingual extractive QA evaluation across 7 languages (English, Arabic, German, Spanish, Hindi, Vietnamese, and Simplified Chinese), providing crucial insights into transfer capabilities between languages [5]. MLQA contains over 12K extractive QA instances in English and over 5K in each target language, with instances being parallel between 4 languages on average.

XLM-R has demonstrated particularly strong cross-lingual transfer capabilities for multilingual question answering. Trained on 2.5TB of data across 100 languages, XLM-R achieved an 8.4% average F1 score improvement on MLQA compared to previous multilingual approaches, with particularly strong performance on low-resource languages such as Swahili and Urdu [6]. Research on multilingual BERT has revealed surprising cross-lingual abilities even when models are trained without explicit cross-lingual objectives [2].

2.3 LLM Evaluation for Low-Resource Languages

Recent comprehensive evaluations have revealed significant challenges in applying large language models to low-resource language tasks. Sengupta et al. investigated extractive question answering with LLMs under domain drift, revealing that LLMs struggle with dataset demands of closed domains and certain models display weaknesses in meeting basic requirements such as discriminating between domain-specific senses of words [7].

Studies on low-resource question answering have demonstrated the feasibility of building effective QA systems using machine-translated datasets and multilingual transfer learning. Research on Turkish QA using SQuAD-TR (a machine translation of SQuAD2.0) achieved 24-32% improvement in Exact Match scores compared to baseline models, demonstrating the viability of adaptation approaches for low-resource languages [8].

Cross-lingual transfer approaches have shown promise for low-resource languages. Zero-shot cross-lingual methods using multilingual models have achieved significant improvements, with fusion-in-decoder techniques showing 3-4.6 point improvements on F1 and EM metrics respectively on cross-lingual datasets [9]. However, performance gaps between high-resource and low-resource languages remain substantial, indicating continued challenges in this domain.

2.4 South Slavic and Macedonian Language Processing

Research specifically focused on South Slavic languages has demonstrated both the challenges and opportunities for NLP in this language family. The SlavNLP workshop series has consistently highlighted the complexity of Slavic languages due to rich inflection, free word order, and derivation phenomena [10]. Specialized models like SlavicBERT, trained on Bulgarian, Czech, Polish, and Russian,

have shown improved performance over general multilingual models for Slavic language tasks [11].

Recent work has investigated transfer learning capabilities within the West Slavic language family, showing that low-resource languages like Upper Sorbian and Kashubian can benefit from models trained on closely related languages [12]. These findings suggest that linguistic similarity within language families can be leveraged to improve performance on underrepresented languages.

For Macedonian specifically, recent efforts have focused on developing open foundation language models and corpora. A comprehensive corpus development project created 1.47 billion high-quality Macedonian words through careful filtering and deduplication processes, along with culturally grounded instruction-tuning datasets [13]. This work demonstrates the growing recognition of the need for dedicated resources for Macedonian language processing.

This body of work establishes the foundation for our evaluation of LLMs on extractive question-answering tasks in Macedonian, providing both methodological frameworks and comparative benchmarks for assessing model performance across different architectures and linguistic variants.

3 Methodology

3.1 Dataset

Our evaluation dataset consists of 194 question-answer pairs in Macedonian, covering both standard Macedonian (168 samples) and its dialects (26 samples). Each question requires extracting specific information from a provided context document.

Dataset Statistics and Composition The dataset exhibits the following characteristics across standard Macedonian and dialectal variants:

Table 1. Dataset Statistics by Component and Dialect

Component	Standard (168 samples)			Dialectal (26 samples)		
	Min	Max	Avg	Min	Max	Avg
Questions (chars)	10	500	64.4	22	138	63.3
Context (chars)	268	1,734	673.9	278	1,734	733.8
Answers (chars)	1	892	135.6	1	322	142.6

The questions are relatively concise, averaging approximately 13 words each, while context documents are more substantial, averaging around 135 words. Answer lengths vary significantly depending on the complexity of the required information, ranging from single-word responses to detailed explanations.

The data was extracted from Discord Q&A conversations and manually annotated where it is official or dialectal, as well as the ground truth from the given context, to ensure high-quality question-answer pairs that reflect real-world information extraction scenarios. All entries in the data are factual and do not require complex reasoning.

The dataset used in this study consists of real-world questions submitted by students and users to the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje. The questions span a range of administrative and academic topics relevant to the faculty, including rules and procedures for studying, institutional regulations, course requirements, information about professors, and other aspects of university life. Each question is paired with a context passage that provides the necessary information for answering, reflecting authentic information-seeking scenarios encountered by students. This domain-specific focus ensures that the dataset closely mirrors practical use cases in higher education administration and student support.

Dialectal Variations The dialectal subset (26 samples) represents linguistic variations found in different regions of North Macedonia. The questions contain elements from various dialects, while the context is always in the official Macedonian language, and the ground truth is also always in the official Macedonian language.

This dialectal component addresses real-world scenarios where language technologies must operate across linguistic variants commonly encountered in North Macedonia, providing valuable insights into model robustness across linguistic diversity.

Dataset Examples Table 2 presents more representative examples, with extra entries added.

Table 2. Dataset Examples

Type	**Question**
Standard	Arhiva do kolku raboti?
Standard	Koj e rokot za prijava na diplomska rabota?
Dialectal	zabravi da prijavu ispiti u rok, treba da platu sg?
Dialectal	koga ce bidit ispitot po matematika?

The dataset along with the context is publicly available on Hugging Face[2].

[2] https://huggingface.co/datasets/Delemangi/llms-evaluation

3.2 Model Categories

We evaluate models across three primary categorizations:

By Accessibility:

- **Closed-source:** GPT-4.1 Mini, GPT-4.1 Nano, O4-Mini (OpenAI); Gemini-2.0 Flash variants, Gemini-2.5 Flash variants (Google)
- **Open-source:** Llama-4 Maverick, Llama-3.3 70B (Meta), Gemma-3 27B (Google), Mistral variants, Qwen-2.5 72B, DeepSeek variants, MKLLM-7B, Domestic Yak-8B

By Model Size:

- **Large (>50B):** Llama-3.3 70B, Qwen-2.5 72B, DeepSeek-R1-Distill-Llama-70B
- **Medium (20-50B):** Gemma-3 27B
- **Small (<20B):** Mistral-7B, MkLLM-7B, Domestic Yak-8B, Mistral-Nemo
- **Proprietary (Unknown Size):** GPT-4.1 Mini, GPT-4.1 Nano, O4-Mini, Gemini-2.0 Flash variants and Gemini-2.5 Flash variants

For proprietary models with undisclosed parameter counts, we create a separate "Proprietary" category to acknowledge that their exact architectural specifications are not publicly available. This classification allows us to analyze performance patterns while being transparent about the limitations in our size-based analysis.

3.3 Evaluation Metrics

We employ both traditional MT metrics (BLEU [16], chrF [18], METEOR [17]) and neural evaluation metrics (COMET [19]) to assess model performance. COMET scores serve as our primary evaluation metric due to their superior correlation with human judgment and their ability to capture semantic similarity beyond surface-level string matching. BLEU and chrF provide insights into n-gram overlap and character-level similarity, while METEOR incorporates stemming and synonymy matching, making it particularly suitable for morphologically rich languages like Macedonian.

4 Results and Analysis

4.1 Overall Performance Ranking

Table 3 presents the overall performance rankings across all 17 models. The results reveal a clear performance hierarchy with closed-source models occupying several top positions.

Table 3. Overall Model Performance (sorted by COMET score)

Model	BLEU	chrF	METEOR	COMET
GPT-4.1 Mini	69.66	72.97	0.732	0.831
Llama-4 Maverick	61.36	68.54	0.691	0.830
Llama-3.3 70B Instruct	49.77	61.24	0.661	0.806
Gemma-3 27B	62.61	68.99	0.682	0.795
Gemini-2.0 Flash	68.79	71.71	0.683	0.791
Gemini-2.0 Flash Lite	56.52	64.89	0.636	0.780
GPT-4.1 Nano	43.20	51.08	0.460	0.726
Gemini-2.5 Flash Preview	41.38	56.41	0.530	0.720
Gemini-2.5 Flash Lite Preview	38.55	52.38	0.501	0.703
MkLLM-7B	34.47	58.57	0.489	0.692
Qwen-2.5 72B Instruct	39.53	50.79	0.516	0.689
Mistral-7B Instruct	34.70	44.09	0.357	0.660
Domestic Yak-8B	34.79	47.65	0.381	0.655
Mistral-Nemo	21.22	35.44	0.336	0.599
DeepSeek-R1	0.49	11.64	0.145	0.481
DeepSeek-R1-Distill-Llama-70B	0.06	7.70	0.084	0.442
O4-Mini	0.00	1.06	0.007	0.395

4.2 Closed-Source vs. Open-Source Performance

Our analysis reveals interesting performance patterns between closed-source and open-source models:

Closed-Source Models: Show varied performance with GPT-4.1 Mini achieving the highest overall score (COMET: 0.831), while O4-Mini performs poorly (COMET: 0.395). Gemini models demonstrate consistent mid-to-high performance ranging from 0.703 to 0.791. This variation within closed-source models suggests that not all proprietary solutions are equal, and specific model capabilities matter more than simply being commercially developed.

Open-Source Models: Demonstrate that open access does not compromise performance, with Llama-4 Maverick achieving exceptional results (COMET: 0.830) that rival the best closed-source models. Large open-source models (>50B parameters) generally outperform their smaller counterparts, though some large models such as DeepSeek-R1-Distill-Llama-70B show surprisingly poor performance, possibly because they were not adequately trained on Macedonian or similar linguistic data during their development process.

4.3 Performance by Model Size

Table 4 shows performance grouped by model size categories. For proprietary models with undisclosed parameters, we classify them separately to maintain analytical transparency while acknowledging the limitation this creates for comprehensive size-based analysis.

Table 4. Performance Analysis by Model Size

Size Category	Model Count	Avg COMET	Best Model
Large (>50B)	3	0.576	Llama-4 Maverick (0.830)
Medium (20-50B)	1	0.795	Gemma-3 27B (0.795)
Small (<20B)	4	0.651	MkLLM-7B (0.692)
Proprietary (Unknown)	9	0.693	GPT-4.1 Mini (0.831)

The performance analysis reveals that model size alone is not a reliable predictor of performance for low-resource languages. The single medium-sized model (Gemma-3 27B) achieves excellent performance, while large models show high variance in results. The poor average performance of the large model category is significantly impacted by the DeepSeek variants, which may lack adequate training on Macedonian-related data. This highlights the importance of training data quality and multilingual capabilities over raw parameter count.

4.4 Dialectal Performance Analysis

Table 5 presents detailed dialectal performance results, showing how models perform on standard Macedonian versus the dialectal variants.

Table 5. Dialectal Performance Comparison (COMET Scores)

Model	Official	Dialects	Difference
GPT-4.1 Mini	0.838	0.854	+0.016
Llama-4 Maverick	0.822	0.847	+0.025
Llama-3.3 70B Instruct	0.798	0.821	+0.023
Gemma-3 27B	0.797	0.829	+0.032
Gemini-2.0 Flash	0.785	0.818	+0.033
Gemini-2.0 Flash Lite	0.770	0.826	+0.056
O4-Mini	0.775	0.808	+0.033
Gemini-2.5 Flash Preview	0.713	0.734	+0.021
GPT-4.1 Nano	0.726	0.766	+0.040
Gemini-2.5 Flash Lite Preview	0.705	0.648	-0.057
MkLLM-7B	0.688	0.723	+0.035
Qwen-2.5 72B Instruct	0.691	0.641	-0.050
Mistral-7B Instruct	0.651	0.679	+0.028
Domestic Yak-8B	0.650	0.657	+0.007
Mistral-Nemo	0.602	0.598	-0.004
DeepSeek-R1	0.754	0.812	+0.058
DeepSeek-R1-Distill-Llama-70B	0.739	0.787	+0.048

Analysis of dialectal performance reveals fascinating patterns. The majority of models (14 out of 17) demonstrate superior performance on the dialectal

variant compared to standard Macedonian. This unexpected finding suggests several possibilities: (1) the dialectal dataset may contain more consistent or focused linguistic patterns, (2) models may have encountered similar dialectal features during training, or (3) the smaller dialectal sample size may lead to more favorable evaluation conditions. A bigger sample size is needed to reach a more precise conclusion.

Notable dialectal performance improvements include Gemini-2.0 Flash Lite (+0.056), DeepSeek-R1 (+0.058), and DeepSeek-R1-Distill-Llama-70B (+0.048). Only three models show better performance on standard Macedonian: Gemini-2.5 Flash Lite Preview (-0.057), Qwen-2.5 72B Instruct (-0.050), and Mistral-Nemo (-0.004).

4.5 Qualitative Error Analysis

In addition to scores, we observed common failure modes: incomplete spans, overlong spans including unrelated clauses, or semantically plausible but unsupported answers. Dialectal queries sometimes caused wrong boundary extraction. Strong multilingual models (GPT-4.1 Mini, Llama-4 Maverick) handled these better, while smaller ones often hallucinated. This complements automatic metrics with a qualitative perspective.

5 Discussion

5.1 Performance Hierarchy

The superior performance of GPT-4.1 Mini and the competitive results of Llama-4 Maverick demonstrate that both closed-source and open-source models can achieve excellent results for Macedonian language processing. The accessibility advantage of open-source models makes Llama-4 Maverick particularly valuable for researchers and practitioners who require local deployment, model customization, or cost-effective solutions for production environments.

The significant performance gap between the best and worst performing models (COMET scores ranging from 0.395 to 0.831) underscores the importance of careful model selection for low-resource language applications. This variation suggests that general-purpose language capabilities do not automatically translate to effective performance on specific languages or tasks.

5.2 Dialectal Processing Capabilities

The observation that most models perform better on dialectal variants is particularly noteworthy and suggests several important implications:

1. Models may have been exposed to diverse linguistic variants during training that share features with the dialectal variants
2. Dialectal variants might contain linguistic patterns that align better with the models' learned representations

3. The smaller dialectal dataset (26 samples) might represent a more coherent subset of the evaluation data
4. Dialectal forms may exhibit linguistic properties that are more universally represented across languages in the training data

This finding has profound practical implications for developing language technologies for dialectal communities and suggests that LLMs may be particularly suited for dialectal language processing tasks. It challenges the common assumption that standard language forms are necessarily easier for AI systems to process.

5.3 Model Size Considerations

The performance analysis reveals that larger models do not always guarantee better performance for low-resource languages. The excellent performance of medium-sized Gemma-3 27B (0.795) compared to some larger models suggests that training methodology, data quality, and architectural choices are crucial factors beyond raw parameter count.

The poor performance of some large models (e.g., DeepSeek variants) indicates that model size must be coupled with appropriate training strategies for multilingual capabilities. This finding is particularly important for resource-constrained environments where computational efficiency is as important as performance.

5.4 Implications for Low-Resource Language Processing

Our results demonstrate that effective language technology for low-resource languages like Macedonian is achievable with current LLM technology. The strong performance of both closed-source and open-source models provides multiple viable paths for implementation, depending on specific requirements regarding cost, customization, and deployment constraints.

The dialectal performance across most models suggests that LLM technology may be particularly valuable for processing non-standard language forms such as dialectal variants as well.

6 Limitations and Future Work

This study has several limitations. The dataset is small, especially dialectal samples (26), limiting generalization. Only extraction-based factual QA is covered, excluding reasoning, multi-hop, or generative tasks. No human evaluation was carried out. Dialectal forms were not sub-annotated by region. Proprietary models with undisclosed parameters complicate fair size-based comparisons. Error categorization was preliminary. Future work should expand datasets, especially dialects, involve human raters, fine-tune models on Macedonian, and compare across South Slavic languages.

7 Conclusion

We presented the first comprehensive evaluation of 17 LLMs on Macedonian extractive QA. GPT-4.1 Mini achieved the best results, closely followed by the open-source Llama-4 Maverick. Model size is not a reliable predictor of performance; training data and methods are key. Many models surprisingly did well on dialects, though evidence is preliminary. Overall, our work highlights both potential and challenges of building effective LLMs for low-resource communities.

Acknowledgments. Supported by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of EMNLP, pp. 2383–2392 (2016)
2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
3. Hugging Face. Question answering with transformers. `https://huggingface.co/docs/transformers/en/tasks/question_answering`, accessed June 30, 2025.
4. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8, 64–77 (2020)
5. Lewis, P., Oğuz, B., Rinott, R., Riedel, S., Schwenk, H.: MLQA: Evaluating cross-lingual extractive question answering. In: Proceedings of ACL, pp. 7315–7330 (2020)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL, pp. 8440–8451 (2020)
7. Sengupta, S., Yin, W., Nakov, P., Ghosh, S., Wang, S.: Exploring language model generalization in low-resource extractive QA. In: Proceedings of COLING, pp. 7106–7126 (2025)
8. Budur, E., Özçelik, R., Soylu, D., Khattab, O., Güngör, T., Potts, C.: Building Efficient and Effective OpenQA Systems for Low-Resource Languages. arXiv preprint arXiv:2401.03590 (2024)
9. Agarwal, S., Tripathi, S., Mitamura, T., Rose, C.P.: Zero-shot cross-lingual open domain question answering. In: Proceedings of MIA Workshop, pp. 91–99 (2022)
10. Piskorski, J., Marcińczuk, M., Nakov, P., Ogrodniczuk, M., Pollak, S., Přibáň, P., Rybak, P., Steinberger, J., Yangarber, R. (eds.): Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023). Association for Computational Linguistics, Dubrovnik, Croatia (2023). `https://aclanthology.org/2023.bsnlp-1.0/`

11. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for named entity recognition on Slavic languages. In: Proceedings of BSNLP Workshop, pp. 89–93 (2019)
12. Torge, S., Politov, A., Lehmann, C., Saffar, B., Tao, Z.: Named Entity Recognition for Low-Resource Languages – Profiting from Language Families. In: Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pp. 1–10. Association for Computational Linguistics, Dubrovnik, Croatia (2023). https://aclanthology.org/2023.bsnlp-1.1/
13. Krsteski, S., Tashkovska, M., Sazdov, B., Gjoreski, H., Gerazov, B.: Towards Open Foundation Language Model and Corpus for Macedonian: A Low-Resource Language. arXiv preprint arXiv:2506.09560 (2025)
14. Gaim, F., Yang, W., Park, H., Park, J.: Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In: Proceedings of ACL, pp. 11857–11870 (2023)
15. Pal, V., Kanoulas, E., Yates, A., de Rijke, M.: Table question answering for low-resourced Indic languages. In: Proceedings of EMNLP, pp. 75–92 (2024)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
17. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of ACL Workshop, pp. 65–72 (2005)
18. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of WMT, pp. 392–395 (2015)
19. Rei, R., Farinha, A.C., Lavie, A., Specia, L.: COMET: A neural framework for MT evaluation. In: Proceedings of EMNLP, pp. 2685–2702 (2020)

Context-Aware Information Retrieval in Workplace Messaging Systems via Retrieval-Augmented Generation and Vector-Based Memory

Ema Pandilova[1,2], Marija Maneva[2], Andrej Petkovikj[2], Ana Markovska[2], Vesna Pop-Dimitrijoska Koteska[2], Pance Ribarski[1,2], and Bojan Ilijoski[1,2]

[1] Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, Skopje, North Macedonia
[2] LOKA, Los Altos, CA 94022, USA ema.pandilova@finki.ukim.mk, marija.maneva@loka.com, andrej.petkovikj@loka.com, ana.markovska@loka.com, vesna@loka.com, pance.ribarski@finki.ukim.mk, bojan.ilijoski@finki.ukim.mk

Abstract. As Large Language Models (LLMs) are increasingly integrated into collaborative platforms like Slack, maintaining contextual relevance across multi-turn, multi-thread conversations remains a significant challenge. This paper presents a Retrieval-Augmented Generation (RAG) system designed to enable persistent, thread-aware, and contextually grounded Slackbot interactions. Our architecture leverages ChromaDB and LlamaIndex for long-term vector-based memory, supporting semantic search over Slack messages enriched with user, thread, and temporal metadata. We introduce a FastAPI-powered pipeline that incorporates keyword filtering and prompt engineering to construct structured, metadata-aware queries, which are then processed by a locally hosted LLM. Through a human-in-the-loop evaluation, we demonstrate the system's ability to retrieve relevant context and generate accurate responses in realistic Slack scenarios. Our results suggest that vector-based memory and metadata-aware prompting significantly enhance LLM usability in dynamic workplace communication environments.

Keywords: Slackbot, Retrieval-Augmented Generation, large language models, vector databases, contextual retrieval, prompt engineering, ChromaDB, LlamaIndex

1 Introduction

Large Language Models (LLMs) are becoming increasingly common in collaborative platforms like Slack, where they offer new possibilities for intelligent assistance and streamlined communication. Despite their capabilities, most LLMs operate without memory or context, they lack thread awareness and cannot easily track multi-turn conversations over

time. This makes it difficult to maintain meaningful context in environments where information is spread across different users, threads, and timestamps[1].
Several frameworks, such as LangChain and ChatGPT Plugins, have attempted to bridge this gap by introducing Retrieval-Augmented Generation (RAG) workflows. Slack itself has released SlackGPT[2] to facilitate AI-driven workplace search. However, most of these implementations still fall short when it comes to metadata-aware retrieval, persistent memory, and fine-grained conversational disambiguation. In particular, thread-specific grounding, temporal context tracking, and user-aware retrieval remain underexplored despite their centrality to team-based knowledge flow.
Recent research has emphasized the growing need for enterprise-grade RAG systems that support context retention and traceability in real-world settings[3]. In response to these challenges, this paper investigates how a vector-based memory system integrated with prompt-aware LLMs can improve information retrieval in Slack-based communication. We explore how thread metadata, user attribution, and message timelines can be leveraged to enhance the accuracy and contextual relevance of generated responses.
The increasing reliance on chat-based tools for organizational knowledge sharing highlights a critical need for conversational agents that can reason over historical exchanges and provide grounded, context-sensitive answers. Prior work in memory-augmented LLMs and conversation disentanglement has shown the importance of capturing not just what was said, but when, by whom, and in what context[4]. Building on these foundations, our work focuses on adapting such techniques to the dynamics of real-world workplace messaging systems.

2 Related Work

Retrieval-Augmented Generation (RAG) has emerged as a widely used method for improving how large language models (LLMs) retrieve and generate context-aware answers. Much of the early research focused on using RAG for tasks like open-domain question answering[3], or adding memory modules to help LLMs handle multi-turn conversations more effectively[4]. Other work has looked at how chunking strategies can be made more adaptive, especially by aligning them with user queries to improve reasoning over long contexts[5].
However, these systems are usually designed for documents or static QA tasks, and they often overlook the more complex structure of workplace messaging systems. Platforms like Slack rely heavily on threads, time-based sequencing, and user attribution. These elements play a critical role in keeping discussions organized and coherent, especially when multiple conversations are happening in parallel. As a result, standard RAG pipelines are often not sufficient for retrieving context in chat-based environments.
More recent work in conversational retrieval has started to include message metadata like user IDs, timestamps, and channel identifiers to improve retrieval accuracy and multi-turn understanding[6,7]. Vector databases

such as ChromaDB support this kind of hybrid search by combining dense vector similarity with structured metadata filtering[8,9], allowing retrieval systems to scale across large archives while staying precise.
Our approach builds on these ideas by applying them specifically to Slack data. We use LlamaIndex for semantically indexing messages while linking them to thread and user information, and ChromaDB for persistent storage with recency-aware filtering. Unlike previous systems that are often tested on synthetic or isolated examples, our implementation is evaluated on real Slack conversations across multiple users and threads, showing how RAG methods can be adapted for realistic, enterprise-level communication tools.

3 System Overview

Slack is a collaboration platform commonly used to support team communication and workflow coordination through persistent messaging and application integration [10]. It supports extensibility through integrations with external services and custom-built applications, enabling teams to centralize discussions, automation, and information access within a single workspace [11].
This work describes the design and deployment of a Slackbot that monitors conversational context, incorporates retrieved knowledge, and produces context-aware responses. The objective is to break down the key architectural components of the system and clarify how their interaction enables context-aware behavior.
We begin by outlining the Slackbot's architectural design, encompassing data storage, transformation, semantic indexing, and language model orchestration. A high-level overview of the system is shown in Figure 1, which illustrates the modular structure and data flow between components.
At the foundation lies robust vector storage via ChromaDB, selected for its ability to handle high-dimensional embeddings efficiently and persistently. We examine its configuration and role in supporting reliable retrieval across sessions.
Next, we detail the query handling mechanism, from initial parsing to the retrieval of contextually relevant information. This includes the algorithms used for filtering and ranking messages based on semantic similarity and metadata.
Finally, we explore the LLM service layer, the core reasoning engine of the Slackbot. We describe how the retrieved context is structured into prompts, passed through a FastAPI gateway, and processed by a local LLM to generate grounded, helpful responses. We highlight how the integration between retrieval and generation enables coherent and thread-aware replies.

3.1 Slackbot Deployment and Permissions

We utilize the Bolt framework to build our Slackbot application. Bolt is a framework that simplifies the process of creation of Slack applications[12].

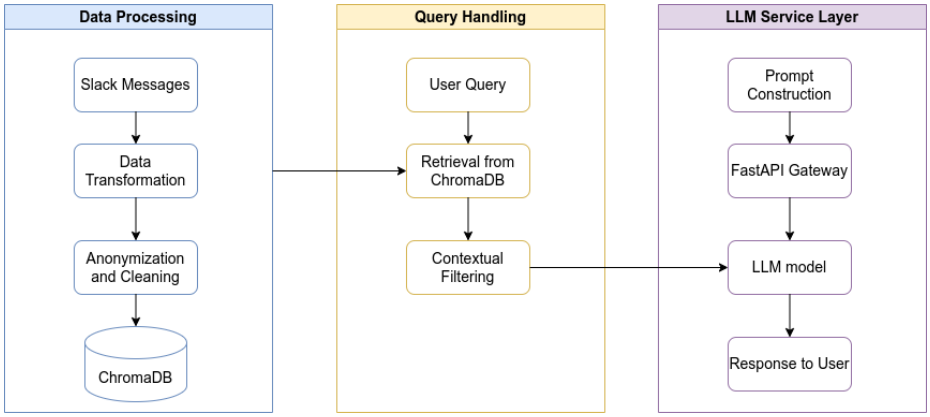

Fig. 1: Overview of the proposed system architecture. The pipeline consists of three modular layers: (1) data processing, where Slack messages are transformed, cleaned, and embedded into ChromaDB; (2) query handling, where user questions trigger contextual retrieval and filtering; and (3) the LLM service layer, which constructs prompts and generates context-aware responses.

Bolt offers two ways of setting up your app either through SocketMode which behind the scenes uses the Events API or setting up a custom app with a public HTTP Request URL. For our Slackbot, for the backend we decided to go with creating a custom application (FastAPI) that will handle the complex logic and streamlining the process a bit better. Before deployment, the Slackbot must be properly registered within the Slack workspace and assigned a unique name, in our implementation, the bot is referred to as *Kenzie*. To enable the required functionality for message processing and interaction, a set of specific permissions (scopes) must be granted through Slack's API configuration interface. These scopes enable the bot to access relevant communication data and perform operations essential to retrieval and response generation. The following OAuth scopes were included:

- `channels:read` - allows the bot to retrieve a list of public channels in the workspace.
- `channels:join` - permits the bot to join public channels automatically if needed.
- `channels:history` - provides access to historical messages within channels, enabling context-aware retrieval.
- `chat:write` - enables the bot to post messages in channels and threads.
- `users:read` - allows access to basic user profile information, supporting user-aware retrieval and metadata enrichment.

Following permission configuration, the Slackbot must be connected to an external service via a public endpoint. In our implementation, a FastAPI-based backend serves as the core application layer that handles request routing, message preprocessing, and error handling. Once

deployed, the bot automatically joins designated channels and becomes capable of responding to user interactions in both channel and thread contexts.
When a new message is sent in a monitored channel, the backend processes the input by transforming it into a structured node that includes key metadata such as sender ID, timestamp, and thread identifier. The message is then embedded into a vector representation and stored in a persistent ChromaDB index. Upon receiving a user query, the system leverages a Large Language Model (LLM) to optionally reformulate the question into a more retrieval-friendly form. The vector database is queried using semantic similarity search to return the top-k relevant message nodes. These results are forwarded to the LLM, which generates a response based on the retrieved context. The final response is then delivered back to the user via Slack.

3.2 Vectore Store and Indexing

LlamaIndex usage LlamaIndex serves as the primary framework for implementing vector-based retrieval and indexing in the Slackbot system. Its main advantage lies in the abstraction layer it provides, allowing seamless integration of storage, embedding, and querying components. The `VectorStoreIndex` class facilitates indexing of message data into vector representations that can later be queried using similarity-based search mechanisms.
To preserve conversational flow, messages are represented as nodes in a directed graph. LlamaIndex maintains message order by linking nodes using the `NodeRelationship.PREVIOUS` attribute, enabling basic thread reconstruction and temporal context tracking [13].
The query engine within LlamaIndex identifies the most relevant chunks in response to a user query through similarity search. A recency-aware postprocessing step (`FixedRecencyPostprocessor`) is applied to prioritize newer messages, improving contextual relevance [5].

ChromaDB collections and persistence strategy The ChromaDB is chosen as a vector storage for persisting through the embedded database. ChromaDB is a specialized vector database designed to store high-dimensional embeddings and support efficient retrieval abilities. This database stands out from the crowd of traditional databases that cannot effectively handle the unique characteristics of feature vectors[9].
The system uses persistence through ChromaDB's PersistentClient architecture. This way, it is ensured that the indexed data is saved and survives beyond application lifecycles. The solution uses ChromaDB to create a persistent layer that saves the vector embeddings and associated metadata across system restarts. The client configurations ensure data durability through disk storage and enable efficient vector similarity querying across a large number of messages.
ChromaDB saves the embedded data in collections using the `get_or_create_collection` method. The data in one collection uses the same embedding function in order to maintain consistency and searchability between the database

and the search query. This design pattern allows for metadata-aware retrieval, which ensures that the search can also benefit from metadata filtering. ChromaDB's collections allow for quick querying through the database and the retrieval of relevant information using its built-in similarity search capabilities.

3.3 Retrieval and Query Handling

The proposed retrieval system uses a process that retrieves the data first and then reads it. This retrieve-then-read pipeline performs vector similarity search to find the top 20 most relevant messages that are stored in the ChromDB database. This retrieval is based on the query that is embedded beforehand by the all-mpnet-base-v2 model[14]. These retrieved results are then sent to the LLM, demonstrating a structure where retrieval precedes generation to increase factual accuracy and context relevance in RAG architectures. This system integrates similarity-based retrieval with metadata filtering to create a hybrid approach that combines semantic understanding with exact filtering. This approach demonstrates an optimization for both relevance and contextual awareness for workplace messaging scenarios[15].

4 Data Processing Pipeline

We developed a pipeline that ingests, transforms, indexes, and exposes Slack messages for retrieval-augmented generation (RAG) using a large language model (LLM), enabling context-aware information retrieval in workplace messaging systems. The system is specifically designed to capture the conversational context inherent in workplace chats and leverage it effectively during both retrieval and response generation The pipeline consists of Slack data transformation, message chunking and vectorization, prompt construction for LLMs, and an API layer for querying and response generation. Each stage plays a critical role in converting raw, unstructured data into a form suitable for semantic search and natural language interaction.

4.1 Dataset Overview

The system was built and evaluated using a publicly available dataset of real Slack conversations collected from open programming communities[16]. The dataset includes discussions from five different Slack channels related to four programming languages: Python, Clojure, Elm, and Racket. These conversations were collected over a two year period, from mid-2017 to mid-2019, and focus mostly on developers asking and answering technical questions. In total, the dataset contains 38,955 conversations and 437,893 individual messages from 12,171 different users.
Each message comes with useful details such as the time it was sent, an anonymous user ID, and the message text. The conversations were grouped using a machine learning method created by the original authors

to help separate different discussion threads that were happening at the same time in the same channel[17]. These Slack channels were public and easy to join, so anyone could participate by simply creating a username and password.
The data was collected daily using special access tokens given by Slack channel administrators. This allowed the researchers to save all messages and avoid losing older ones due to Slack's free-tier limit. During the cleaning process, only messages were kept and things like notifications about users joining or reacting were removed. Usernames were also replaced with fake names to protect privacy. The final version of the dataset is stored in XML format, organized by community and year. It's well-structured and easy to work with, making it a great resource for building systems like ours[16].

4.2 Slack Data Transformation

The original Slack messages in the dataset are provided in XML format, which, while structured, tends to be verbose and less convenient for programmatic manipulation. To enable easier and more efficient processing, we convert all XML files into a structured JSON format. This transformation is carried out using a custom Python script that leverages the xmltodict library[18] to parse XML elements into nested Python dictionaries, which are then serialized into clean, human-readable JSON files.
As part of the transformation process, we also clean the data by removing duplicate messages, which may appear due to inconsistencies in the original exports. The result is a streamlined and consistent dataset that maintains all relevant information such as message content, timestamps, anonymized user identifiers, and conversation identifiers. This structured JSON format is easier to handle and aligns well with the requirements of our vector indexing and prompt generation components.
The xmltodict library has also been adopted in other systems with similar goals of bridging XML and JSON formats. One such example is autoAPI, a web-based tool designed for defining API endpoints that return JSON data from XML sources[19]. In this system, xmltodict plays a central role by converting raw XML inputs into JSON objects, allowing users to interactively select which parts of the data to include. This functionality enables automatic generation of streamlined, simplified JSON outputs tailored to user defined preferences, demonstrating the versatility and robustness of xmltodict for real world data transformation pipelines.

4.3 Message Chunking and Embedding

After converting the Slack data into a structured JSON format, each message undergoes a series of transformations to prepare it for semantic search. At this stage, messages are processed as standalone units, a strategy known as message-level chunking. This avoids the complexity of full thread disentanglement while enabling atomic indexing of communication artifacts.

Inspired by cognitive models of linguistic chunking[20], each message is treated as a semantic unit. It is encapsulated within a `TextNode` structure that also includes associated metadata such as the anonymized user ID, timestamp, message ID, and Slack channel ID. This design enables downstream filtering, contextual analysis, and traceability in conversation flow. When applicable, the system creates directed links between nodes using `NodeRelationship.PREVIOUS` to preserve thread continuity.
Once messages are chunked, they are embedded into high-dimensional vectors using the `all-mpnet-base-v2` sentence-transformer model from HuggingFace. This model is known for high semantic accuracy while maintaining reasonable inference speed[21]. LlamaIndex leverages this embedding model through its `HuggingFaceEmbedding` wrapper, while ChromaDB uses the same model via its `SentenceTransformerEmbeddingFunction`, ensuring consistency between retrieval and storage layers.
The embedding vectors are stored in a persistent ChromaDB index. Chroma's architecture supports fast vector similarity search as well as metadata filtering, enabling hybrid retrieval based on both content and structured attributes. Internally, Chroma employs approximate nearest neighbor (ANN) indexing techniques[22], and its collections are organized by embedding function to maintain consistency.
This architecture enables metadata-aware retrieval with scalable indexing performance. Previous research has shown that combining semantic similarity with structured metadata yields better performance in multi-turn conversational systems compared to vector-only or keyword-only approaches[8].

4.4 Metadata-Aware Retrieval

The system enhances retrieval precision and relevance by applying metadata-aware filtering and recency-based ranking before passing the results to the LLM.
When a user submits a query, the retriever first filters potential candidates based on metadata fields such as sender identity, channel ID, or thread context. These fields are embedded within each message node during preprocessing and enable targeted retrieval aligned with conversational context[7]. User-based filtering supports retrieval grounded in specific authorship, while channel-level separation prevents accidental cross-thread contamination.
In addition to metadata filtering, the system uses a recency-based scoring mechanism to prioritize recent messages in Slack threads. This is implemented using LlamaIndex's `FixedRecencyPostprocessor`, which applies a time decay function that downweights older content based on timestamps. This helps ensure that responses are drawn from the most temporally relevant context, an approach validated in prior research on time-sensitive information retrieval[23,24].
This combination of structured filtering and temporal reranking creates a robust retrieval pipeline. Rather than relying solely on vector similarity, the system leverages hybrid relevance signals that align better with how workplace conversations evolve over time and across users[6].

4.5 Prompt Engineering and API Serving

Prompt Construction Strategy The prompt engineering strategy employed in this implementation is carefully crafted to maximize the usefulness and clarity of the Slack messages retrieved from the vector database. Each message is serialized along with its associated metadata, including the sender's identity, the timestamp, and the channel name, and organized into a structured, JSON-like format. This structured representation provides the language model with essential contextual cues, enabling it to interpret each piece of information effectively.

The prompt begins with an instructional preamble that explicitly explains the significance of each metadata field. This explanation helps the model understand how to utilize the timestamp information to infer the chronological order of the messages, while the sender and channel details offer insight into the participants and the conversational context. Rather than pre-sorting the messages before sending them to the model, the prompt instructs the language model to mentally reorder the messages based on their timestamps. This approach leverages the model's reasoning capabilities to reconstruct the conversation flow without additional preprocessing.

Each message follows a consistent and clear format, which facilitates the model's parsing and comprehension of the content. After presenting the contextualized messages, the user's query is appended following a clearly defined delimiter, accompanied by precise instructions guiding the model to generate responses strictly grounded in the provided context. In situations where the relevant information is not present, the model is encouraged to seek clarification or express uncertainty rather than fabricating unsupported details.

This combination of structured metadata, explicit instructions, and implicit temporal reasoning fosters a robust balance between contextual awareness and cautious fallback behavior, enhancing both the relevance and reliability of the generated answers. Such an approach reflects the growing recognition within the field that prompt engineering is a critical technique to unlock the full potential of large language models, guiding them to produce accurate and contextually appropriate outputs. As highlighted in recent research, prompt engineering plays an essential role in enabling AI systems to interact safely, intuitively, and effectively across various domains, thereby transforming how humans and AI collaborate to solve complex problems[25].

FastAPI Integration The system delivers its core retrieval-augmented generation capabilities through a lightweight, high-performance RESTful API built using FastAPI, a modern web framework designed to make API development in Python both fast and reliable by leveraging Python type hints, async support, and automatic validation through Pydantic[26]. This interface acts as the orchestration layer that ties together user input, semantic retrieval from the vector database, and the response generation process via a locally hosted large language model.

Clients interact with the system by sending POST requests to the root endpoint, submitting a JSON payload containing the user's query. The

interface also supports optional metadata filters such as sender identity. When a query is received, the API creates a retriever instance based on the Chroma vector store to obtain the most semantically relevant messages. These retrieved messages are then serialized into a structured format and embedded within a carefully crafted prompt template.
This complete prompt, which includes both the contextual information and the user's question, is forwarded to a locally hosted language model endpoint. The default model deployed in this system is the *dolphin-2.5-mixtral-8x7b*[27], served through the llama.cpp backend. The integration is designed with full traceability in mind, capturing both the request payload and the response content to support debugging and analysis.
The modular design of this architecture allows for easy adaptation to alternative models or endpoints. It also facilitates experimentation with various retrieval strategies and prompt formats, while maintaining a clear separation between the components responsible for retrieval, prompt construction, and model interaction.

5 Evaluation

We conducted a qualitative and semi-quantitative evaluation to assess the effectiveness of our context-aware Slackbot, focusing on two key components: (1) the capability to retrieve relevant conversational context using a vector database (**ChromaDB** combined with **LlamaIndex**), and (2) the ability to generate accurate, contextually grounded responses using a large language model (LLM).
The evaluation focused specifically on the bot's *thread awareness*, *memory persistence*, and its use of *temporal and user metadata*, all of which are core features supported by our implementation.

5.1 Evaluation Methodology

We used a spreadsheet-based, human-in-the-loop evaluation protocol to reflect realistic interactions with our FastAPI-powered Slackbot. Similar human-in-the-loop approaches have been shown to significantly enhance the performance of RAG chatbots[28]. We started by curating a set of 20 representative queries based on actual Slack conversations, covering different scenarios including single-turn and multi-turn dialogues, thread-specific questions, and time-sensitive topics. The queries were selected to ensure diversity across different Slack channels and user interactions. We included a balance of single-turn and multi-turn conversations, queries that required thread-specific grounding versus cross-thread reasoning, and both straightforward factual requests and more open-ended, time-sensitive questions. This sampling strategy was designed to approximate the range of realistic queries encountered in workplace communication. For each query, an expected answer was defined through human annotation of the Slack discussion logs. Single-turn queries involve a single question–answer exchange, whereas multi-turn queries require the system to track and utilize context across multiple conversational turns or follow-up interactions.

Each query was then run through our system, where the backend retrieved the top-k (k=10) most similar message chunks from the Chroma vector database. Human annotators examined these retrieved nodes to determine whether they included the information necessary to answer the query, and recorded the rank of any relevant context found. We documented both successful retrievals and cases where the context was missing or when irrelevant chunks got pulled in.
The retrieved nodes were formatted as structured context and combined with the user query into a prompt, following the logic implemented in our FastAPI and Slackbot code. The LLM (dolphin-2.5-mixtral-8x7b) generated a response, which was rated by human evaluators on a 1–5 Likert scale for factual correctness and contextual appropriateness (with 5 indicating fully correct and context-aware answers, and 1 indicating hallucinated or irrelevant responses).Our use of a five-point scale for evaluation is consistent with methodologies used in recent user studies of RAG systems[29].
All together, this human-in-the-loop evaluation provided a realistic picture of our system's strengths and weaknesses in actual user scenarios, and helped us identify opportunities for further improvement.

5.2 Results

Over the course of evaluating 20 different queries, the system successfully retrieved at least one highly relevant context node within the top-10 results for 15 queries (75%). Retrieval performance was strongest when the question and answer were part of the same thread, thanks to the way the system leveraged conversation metadata and the structure of Slack threads. Even for more complex cases, like multi-turn exchanges or thread-specific questions, the pipeline generally managed to pull up the key messages or relevant conversation history needed for a good answer. The LLM's answers, scored across all 20 queries regardless of retrieval success, achieved an average human-assigned response score of 4.3 out of 5. When relevant context was retrieved, the LLM was generally able to generate accurate, contextually grounded answers. For straightforward factual queries or those with focused discussion history, the LLM provided concise, precise answers. For more complex or multi-thread scenarios, it demonstrated the ability to synthesize information from multiple retrieved nodes.
A summary of representative evaluation examples across these scenarios is presented in Table 1, illustrating the effectiveness of contextual retrieval and generation in realistic Slack queries.

Table 1: Examples of Retrieval Effectiveness and LLM Answer Quality Across Realistic Slack Queries

Query	**Expected Answer**	**Retrieval Outcome**	**LLM Answer (Score 1–5)**	**Notes**
How is Rona's masters program going?	Graduate a month ago	Most relevant node found	5	Perfect context match and synthesis
How to do sentiment analysis on tweets?	Use TextBlob, MongoDB, real-time aggregation	Top 2 relevant nodes	5	Accurate synthesis from multiple threads
How to 'pluginfy' an application?	Multiple options discussed by Collette	Most relevant node found	5	Several approaches summarized
Does Orpha wrote a flask book?	She is hoping to finish writing a book on flask	Most relevant node found	5	Specific answer provided from thread
Advice a book to study Python 3 for beginners?	"Fluent Python", "Learning Python"	Relevant node not top-ranked	1	LLM provides a generic but not fully accurate answer
Is there any place where you can learn how to import data from pandas DataFrame to MySQL?	Best practices, code examples, or relevant resources for importing pandas DataFrames to MySQL	Multiple related nodes retrieved, discussing code snippets and documentation links	4	LLM synthesized various thread suggestions, provided general guidance, but did not cite a single definitive tutorial
How to reset a password?	Not present in dataset	Irrelevant/low-similarity nodes retrieved	N/A	LLM correctly reported no info available

Our analysis revealed several noteworthy limitations of the system:

- When questions and answers were located in different threads (i.e., different conversation IDs), specifying the conversation ID as a retrieval filter enabled successful retrieval, otherwise, cross-thread linkage was not possible.
- For queries about user-specific actions ("Who asked that question?"), the answer could not be retrieved if user metadata was missing or not explicitly linked within the thread.
- Broad or popular topics (e.g., general Python questions) led to retrieval of diffuse or only partially relevant context, sometimes diluting the LLM's response.
- For queries with no corresponding answer in the dataset, the retriever returned only low-similarity nodes, and the LLM appropriately stated it could not answer, demonstrating low hallucination risk.

6 Conclusion

In this work, we presented a Retrieval-Augmented Generation (RAG) system designed to process and respond to queries over workplace messaging data using Slack conversations. Our system combines structured message transformation, message-level chunking, vector-based retrieval with ChromaDB, and prompt-based response generation using a locally hosted LLM. Through qualitative and semi-quantitative evaluation, the system demonstrated high accuracy in retrieving relevant context and generating grounded, helpful responses.

While our current implementation provides a robust and extensible foundation, it indexes each Slack message in isolation. At this stage, the system does not yet implement advanced chunking strategies such as thread-aware grouping, temporal proximity heuristics, or token-based segmentation. These limitations restrict the model's capacity to preserve long-range conversational context and maintain coherent dialogue flow. In addition, our evaluation relied on a limited set of 20 annotated queries. Although this sufficed for an initial qualitative assessment, future studies should expand the evaluation to a larger pool of queries to strengthen statistical validity and ensure broader generalizability

Future work may explore chunking mechanisms that group semantically related messages into cohesive conversational units. Incorporating contextual and structural signals, such as message threading, temporal proximity, or semantic similarity, could enhance retrieval accuracy and improve the relevance of generated responses.

Progress in chunking strategies, combined with improvements in re-ranking, model adaptation, and failover mechanisms, can contribute to more robust and context-aware RAG systems for workplace communication scenarios.

References

1. Garima Agrawal, Sashank Gummuluri, and Cosimo Spera. Beyondrag: Question identification and answer generation in real-time conversations. *arXiv preprint arXiv:2410.10136*, 2024.
2. Introducing slack gpt. `https://slack.com/intl/en-gb/blog/news/introducing-slack-gpt`, 2023. Accessed: 2025-12-26.
3. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
4. Yujia Zhang, Chenguang Wu, Yichong Chen, et al. Retrieval-augmented generation for conversational memory grounding. *arXiv preprint arXiv:2305.09629*, 2023.
5. Derya Tanyildiz, Serkan Ayvaz, and Mehmet Fatih Amasyali. Enhancing retrieval-augmented generation accuracy with dynamic chunking and optimized vector search. *Orclever Proceedings of Research and Development*, 5(1):215–225, 2024.
6. D. Yaganti. Enterprise-grade conversational intelligence: A domain-aware chatbot framework using gpt-3.5, langchain, and rag with local vector indexing. *International Journal of Advanced Research in Science, Communication and Technology*, pages 625–631, May 2024.
7. Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. Learning implicit user profile for personalized retrieval-based chatbot. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, CIKM '21, page 1467–1477. ACM, October 2021.
8. Simeon Emanuilov and Aleksandar Dimov. Billion-scale similarity search using a hybrid indexing approach with advanced filtering. *arXiv preprint arXiv:2501.13442*, 2025.
9. Jialin Wang and Zhihua Duan. Empirical research on utilizing llm-based agents for automated bug fixing via langgraph. 2025.
10. Heather A Johnson. Slack. *Journal of the Medical Library Association: JMLA*, 106(1):148, 2018.
11. Emanuel Stoeckli, Falk Uebernickel, and Walter Brenner. Exploring affordances of slack integrations and their actualization within enterprises-towards an understanding of how chatbots create value. 2018.
12. Slack Technologies. Bolt for javascript and python – slack api. `https://api.slack.com/bolt`, 2024. Accessed: 2025-06-20.
13. João Pinheiro, Wendy Victorio, Eduardo Nascimento, Antony Seabra, Yenier Izquierdo, Grettel García, Gustavo Coelho, Melissa Lemos, Luiz André P Paes Leme, António Furtado, et al. On the construction of database interfaces based on large language models. In *WEBIST*, pages 373–380, 2023.
14. Ayman Asad Khan, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report, 2024.

15. Xiwei Xu, Hans Weytjens, Dawen Zhang, Qinghua Lu, Ingo Weber, and Liming Zhu. Ragops: Operating and managing retrieval-augmented generation pipelines. *arXiv preprint arXiv:2506.03401*, 2025.
16. Preetha Chatterjee. Software-related slack chats with disentangled conversations. `https://github.com/preethac/Software-related-Slack-Chats-with-Disentangled-Conversations`, 2024. Accessed: 2025-06-20.
17. Preetha Chatterjee, Kostadin Damevski, Nicholas A Kraft, and Lori Pollock. Software-related slack chats with disentangled conversations. In *Proceedings of the 17th international conference on mining software repositories*, pages 588–592, 2020.
18. Martin Blech. xmltodict: Makes working with xml feel like you are working with json. `https://github.com/martinblech/xmltodict`. Accessed: 2025-06-20.
19. Adrian Moore. auto{API} – a web-based tool for specification of an api endpoint to return json data from an xml source. *Journal of Open Research Software*, Aug 2021.
20. Jinbiao Yang, Qing Cai, and Xing Tian. How do we segment text? two-stage chunking operation in reading. *Eneuro*, 7(3), 2020.
21. Carlo Galli, Nikolaos Donos, and Elena Calciolari. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. *Information*, 15(2):68, 2024.
22. Douglas Rolins de Santana and Leonardo Andrade Ribeiro. Approximate similarity joins over dense vector embeddings. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 51–62. SBC, 2023.
23. Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. Tempretriever: Fusion-based temporal dense passage retrieval for time-sensitive questions. *arXiv preprint arXiv:2502.21024*, 2025.
24. Anoushka Gade and Jorjeta Jetcheva. It's about time: Incorporating temporality in retrieval augmented language models. *arXiv preprint arXiv:2401.13222*, 2024.
25. Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 2025.
26. Bill Lubanovic. *FastAPI*. O'Reilly Media, Inc., Sebastopol, CA, 2023.
27. E Hartford. dolphin-2.5-mixtral-8x7b. *URL https://erichartford.com/dolphin-25-mixtral-8x7b. Accessed*, pages 01–02, 2024.
28. Anum Afzal, Alexander Kowsik, Rajna Fani, and Florian Matthes. Towards optimizing and evaluating a retrieval augmented qa chatbot using llms with human in the loop. *arXiv preprint arXiv:2407.05925*, 2024.
29. Septian Eka Ady Buananta, Muhammad Fikri Hasani, Arya Nathan Bara, and Hany Wijaya. User experience analysis when finding information: Comparative study of rag and traditional search engine. In *2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pages 798–803. IEEE, 2024.

LLM Chatbot with SQL Database

Dean Nastevski[1], Lazo Nikoloski[1], Dimitar Kitanovski[1], Zorica Karapancheva[1], Aleksandar Stojmenski[1], Ivan Chorbev[1], and Petre Lameski[1]

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia

Abstract. Accessing and analyzing data from relational databases typically requires knowledge of SQL, which can present a barrier for users without technical backgrounds. To address this challenge, we introduce a web-based application that enables users to retrieve data by expressing their queries in natural language. By integrating Large Language Models (LLMs), the system automatically interprets user requests and translates plain English inputs into correct and meaningful SQL queries. This eliminates the need for manual query writing and lowers the barrier to data access in organizations which are data-driven. The application architecture combines a user-friendly front-end with a back-end pipeline that handles language interpretation, SQL generation, query execution, and result presentation. LLMs play a central role in parsing complex user requests, managing language misunderstandings, and generating safe and effective SQL statements. Through this approach, the platform facilitates faster insights, enhances collaboration between technical and non-technical stakeholders, and supports more inclusive access to data resources.

Keywords: LLM · AI · chatbot · SQL.

1 Introduction

Structured SQL databases play a significant role in the IT world, and having quick and easy access to information is crucial for system users. Typically, retrieving the required information from the tables of these systems is done using SQL queries, which require knowledge of a specific programming language. Users who are unfamiliar with SQL syntax may have difficulties accessing the desired data.

When trying to get that requested data, there is also the problem of having a dedicated application or IDE where an SQL query can be ran. Some of these notable IDEs include PGAdmin [1] and PhpPGAdmin [2]. The IDEs also require additional configuration for properly connecting a database on which the SQL queries will be conducted.

Accessing SQL databases can be simplified using LLM models to create a text-to-SQL system, which generates SQL queries from natural language. This idea has existed for a long time, with one of the first implementations known

as Seq2SQL [3], which revolutionized the way SQL queries are generated from natural language using deep neural networks.

With improved LLM models that can process larger word contexts of up to 128K tokens [4] or are more specifically trained for programming tasks [5], it is possible to develop a system that acts as a mediator for generating SQL queries and displaying responses. The system's purpose is to be simple in nature, where users can easily navigate to desired databases and request data in a more intuitive and conversational manner. This will avoid the additional configuration of IDEs and knowledge of using the SQL query language.

This paper explains the integration of an LLM chatbot with SQL databases through a .NET 8 web application using the *Onion* architecture [6]. Additionally, the paper examines microservice and monolithic implementations of such a system, their advantages and disadvantages, and the obtained results.

2 Related Work

Translating natural language to SQL using large language models (LLMs) right now has become a well-studied area with various architectures and model integration strategies proposed to improve performance, accuracy, and efficiency. In their early stages, systems like Seq2SQL tried making deep learning approaches to make user questions into SQL queries [3].

Recent surveys have classified large language model enhanced text-to-SQL generations into prompt engineering, fine-tuning, pre-trained and Agent groups according to training strategies [7].

Some approaches have tried direct prompting or instruction-tuned large language models for SQL query generation. One such approach is the SQL-PaLM that utilizes few-shot prompting and execution-based error filtering to improve generation quality [8]. Another approach is DAIL-SQL which focuses on prompt optimization and token efficiency [9].

More recent studies have shown significant performance improvements in Text-to-SQL tasks by fine-tuning models on specific datasets. For instance, one study achieved a 61% execution accuracy after fine-tuning the WizardCoder-15B model on the Spider dataset using QLoRa fine-tuning. This highlights that while direct prompting is a viable method, fine-tuning and adaptation to specific data schemas can yield substantial gains in query correctness [10].

Other methods take a more hybrid or modular approach. This means the methods are actually breaking the SQL generation process into smaller stages handled by different services. In this case, there's frameworks like ZeroNL2SQL [11] and CodeS [12] where they use small or pre-trained models for initial query sketching and afterwards give the more complex reasoning tasks to larger models. CHESS [13] and XiYan-SQL [14] go further by including multi-agent and ensemble architectures that decompose the problem into schema linking, candidate generation, and validation tasks. These systems demonstrate improvements in both accuracy and efficiency, often using microservice or modular implementations to implement the various services.

The architectural design—whether monolithic or microservice-based—also plays a key role in performance. Talaei et al. report that their multi-agent microservice design achieved a 2% increase in execution accuracy while reducing LLM calls by 83% and token usage by a factor of five [13]. Similarly, other modular systems demonstrate scalable, efficient performance without sacrificing accuracy, with exact match scores improving by up to 11.8% [15]. While not all studies explicitly state their system architecture, the trend toward modular or pipeline-based designs shows a growing preference in the systems looking out for flexibility and scalability in real-world deployments.

3 Implementation

In this section, we explain the design and components of the web application system. To summarize, there is a web application on which a user can register an account. Within the account, the user can add databases by entering the database's credentials. Finally, they can ask questions and receive answers in a conversational manner for the given database.

3.1 Implementation using a micro-service architecture

The first implementation we did for this SQL chatbot system was using the microservice architecture. Defined by Martin Fowler, the microservice architecture represents creating a system or application of independent connected services [16]. In our case, we can imagine these services as separate modules that implement functionalities towards getting a final answer out of a requested database within the system.

Our approach with the microservice architecture implements 2 separate services that make the final system:

- .NET service consisting of the back-end and front-end;
- Python service consisting of the SQL query generation and answer retrieval.

.NET Service As previously mentioned, the .NET service represents the front-end and back-end of the system using the *Onion* architecture. Specifically, these are the 4 layers this application utilizes of this architecture:

- Domain layer;
- Repository layer;
- Service layer, and
- Web interface layer.

The domain layer contains the definition of the entities (user, database, questions, answers) within the internal Microsoft SQL Server which can be seen in Figure 1. The repository layer contains the logic of communicating with the internal database, while the service layer has the implementation of functionalities for working with the repository layer. Finally, the web interface layer is used to present the data and the other functionalities of working with the system for an end user on a web application.

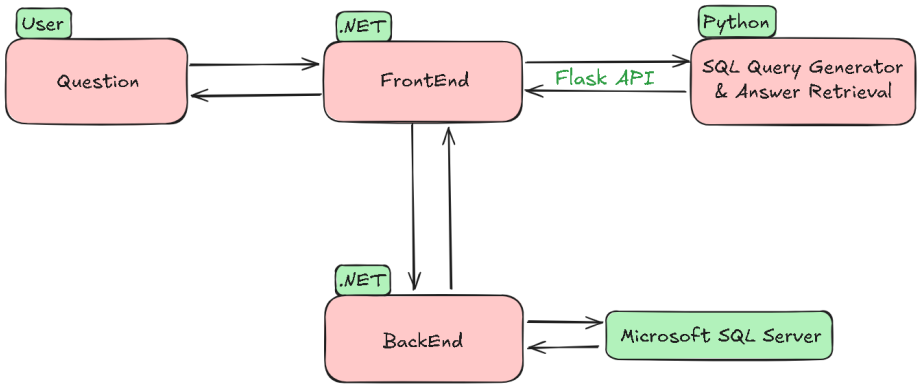

Fig. 1. Microservice architecture implementation of the SQL chatbot

Python Service The Python service generates an SQL query using an LLM. Then the query is run on a given database by the end user to retrieve answers. Since the query retrieves only rows of data from a database, an LLM is used again to reform the final answer in a conversational sentence where it is given back to the user, i.e. back to the .NET service.

The main Python library used in this system for working with LLMs is LangChain [17]. Here are some of the functionalities we are using from the LangChain library:

- Initializing a connection to an LLM;
- Initializing a connection to an SQL database;
- Creating prompts, which are passed to the LLM, and
- Getting the schema of the tables for a given database and 3 example rows from each table, which is passed to a prompt.

There are many LLM API providers, but in our case, we're using HuggingFace's API [18] to connect to the *Mistral-7B-Instruct-v0.2* [19] large language model that's used to generate an SQL query and the answer based on the results of that query as seen in Figure 2.

For the Python service to receive requests from the .NET service, we are using the Flask library [20]. The Flask library opens up API endpoints for the Python service where a request with a question about a database can be obtained and processed using LangChain and the LLM, and then the generated SQL query is executed on the database. The retrieved results are then formatted and returned as a response to the .NET service.

3.2 Implementation using a monolithic architecture

The second implementation for this SQL chatbot system was using monolithic architecture. A monolithic architecture ensures all services are defined and exe-

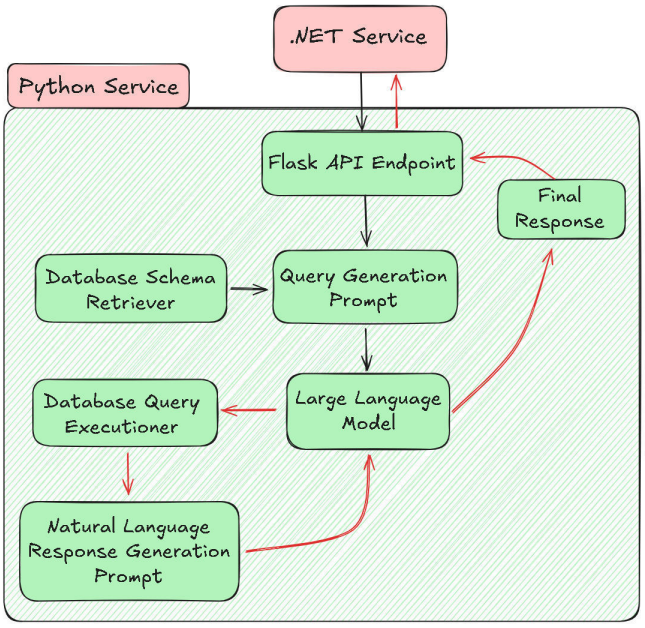

Fig. 2. Python service of the microservice architecture

cuted within a single code base [21, 22]. In our system, the order of implementations is important because this means we adapted the same functionalities we got out of the Python service into the .NET service, ensuring all functionalities of the system are sustained in a single code base.

Since we are using the *Onion* architecture, all these Python service functionalities in Figure 2 were added to the *Service* layer of the .NET service. This includes the prompt generation functions, database schema retrievers, and the API calls to the LLM.

This way, we removed the inter-process communication (IPC) overhead introduced in the microservices architecture and ensured function calls happened directly within the same application context, which helped reduce the overhead associated with inter-service communication [23]. However, we also saw that the monolithic architecture did increase the complexity of the system when it comes to the management of functionalities within a single code base.

The functionalities of the Python service were later used in the *Web Interface* layer of the *Onion* architecture. Since each consecutive layer in the Onion architecture has the previous layers as project dependencies, only the Web Interface layer that has the Service layer as a dependency was able to use these functionalities. This design ensures that the core business logic that is encapsulated in the Domain and Repository layers remains independent and isolated from external concerns like the web interface [24]. As shown in Figure 3, the

Web Interface layer interacts with the Service layer, which contains the Python service functionalities, while the other layers remain unaware of this integration.

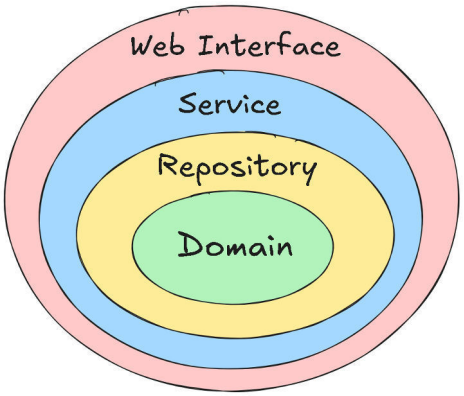

Fig. 3. "Onion" architecture in the .NET service

4 Testing

The system was tested such that we opened a specific API endpoint within the microservice and monolithic versions. The endpoint would simulate the whole process of a user asking a question for a given database through the system, going through every intended service and function. We specifically looked at:

- Successfully ran queries for a given database, and
- Unsuccessfully ran queries for a given database.

For this testing, we used a selection of 350 queries of the lite Mini-Dev dataset [25] from the BIRD benchmark [26]. We tested using the dataset where we only loaded specific tables and their assigned testing queries alongside the tables on which they were dependent on through relations. We are not loading and using the whole dataset with all tables because we are sending its schema and 3 example rows of each table to the LLM, where the prompt can get up to approximately 28095 tokens. It should be noted that our evaluation relied exclusively on the Lite Mini-Dev subset of the BIRD benchmark (approximately 1% of the full dataset). While this choice was motivated by prompt length constraints and computational feasibility, it limits the generalization of our findings. Future evaluations on larger subsets of BIRD, as well as additional benchmarks such as Spider [27] and TPC-DS [28], will be necessary to provide a more comprehensive assessment of system performance.

Testing the SQL chatbot system is a delicate process, as its effectiveness depends on how the system is configured. While we loaded the *Mistral-7B-Instruct-v0.2* large language model using the HuggingFace API, alternative models and

methods for deploying LLMs can be considered for SQL query processing and natural language response generation. Specifically, the HuggingFace API also uses caching and preloading models [29] which can shift results.

For instance, depending on data privacy requirements and user needs, the system may require a locally hosted LLM. System owners might need to comply with emerging privacy laws and standards [30, 31] to ensure that data remains on personal servers rather than being sent through third-party APIs.

Inference time and result accuracy can also vary when running an LLM locally because of computing power [32]. Additionally, the choice of the model and tools or applications running an LLM locally, such as Ollama [33], can further optimize inference performance through hardware and software optimizations. The testing process can be reviewed visually in Figure 4.

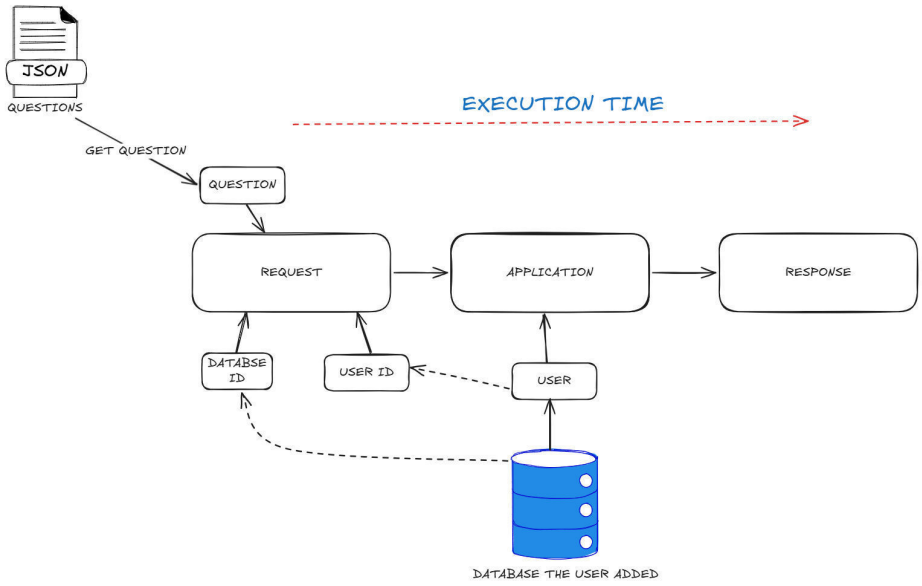

Fig. 4. Diagram that illustrates how we retrieved questions, processed them through the application, and validated responses

5 Results

This section presents a detailed analysis of the system's performance in terms of execution time and success rate of query generation and execution. The evaluation aims to compare the microservice-based and monolithic architectures under the same workload conditions, using the same query difficulty levels.

To estimate the system's performance, two primary metrics were analyzed:

– the average execution time for both valid and invalid queries, and
– the overall success rate of a valid query generated at each difficulty level (Simple, Moderate, Challenging).

These metrics offer insight into the trade-off between performance speed and system reliability in different architectural configurations.

5.1 Results on Time and Query Success Rate

Figure 5 illustrates the execution time distribution and success rates across query difficulty levels using the microservice architecture. We can see that the execution time increases with the complexity of the query as valid queries consistently take longer to execute than invalid ones due to the system's need to make two LLM calls for generating and validating a complete response. The success rate also drops as difficulty increases, with the highest rate being 57.6% for simple queries.

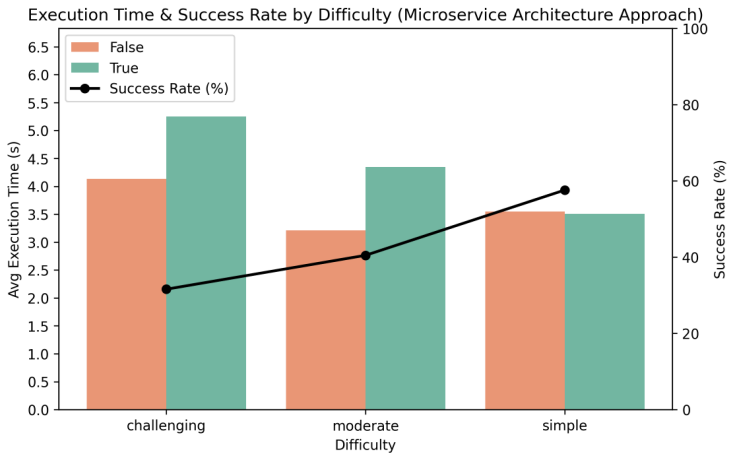

Fig. 5. Execution time distribution across query difficulty levels for the microservice approach.

Figure 6 shows the same metrics for the monolithic approach. Here we can notice that the execution times are shorter across all difficulty levels when compared to the microservice architecture. This indicates that reducing inter-service communication overhead improves responsiveness. But there is a noticeable drop in query success rates, especially for more challenging queries. In our testing the success rate for challenging queries falls to 22.81%, which tells us the system is faster, but the drop in the success rate of generating a valid query indicates the LLM's randomness in this process.

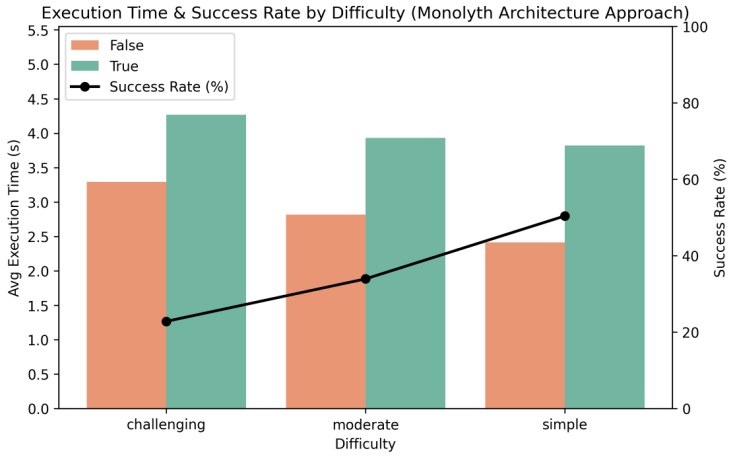

Fig. 6. Execution time distribution across query difficulty levels for the monolithic approach.

Table 1 summarizes the numerical results for both architectures. The microservice architecture yields higher success rates at the cost of longer execution times. On the other hand, the monolithic approach prioritizes speed but struggles a little bit with reliability. These findings highlight a clear trade-off when it comes to the execution time, although valid query generation does largely depend on the LLM.

Table 1. Execution Time and Success Rate by Difficulty: Microservice vs Monolithic

Difficulty	Microservice			Monolithic		
	Avg Time (s)		Success	Avg Time (s)		Success
	Unsuccessful	Successful	Rate (%)	Unsuccessful	Successful	Rate (%)
Challenging	4.135	5.252	31.58	3.289	4.269	22.81
Moderate	3.214	4.344	40.48	2.816	3.930	33.93
Simple	3.550	3.505	57.60	2.412	3.819	50.40

[a]Execution time represents average query processing time for each difficulty level.

These results suggest that while the microservice architecture tends to exhibit longer execution times, it consistently achieves a higher success rate in generating valid queries. However, it's important to note that the success of query generation is influenced primarily by the random nature of the LLM rather than the system architecture itself. So the observed differences in success rates may reflect incidental variation rather than architectural superiority.

Still, in use cases where reliability and valid query completion are critical, such as educational platforms or decision support tools, the microservice approach may offer a slight advantage. On the other hand, monolithic systems, with their faster response times, may be better suited for applications where low latency is a priority. The choice of architecture should be informed by the specific requirements of the target application, especially in balancing accuracy, responsiveness, and system complexity.

5.2 Discussion on Accuracy

For the successful queries that were executed, we did not evaluate whether the returned answers were semantically correct or aligned with the user's intent. This is because the accuracy of SQL query generation is primarily dependent on the capabilities of the underlying large language model (LLM), which operates independently of the system architecture. As such, the evaluation of query accuracy must be conducted through separate benchmarking of the LLM itself.

For example, the BIRD benchmark (as of August 8, 2024) ranks the Mistral baseline model with 123B parameters at an execution accuracy score of 55.84, placing it 48th among tested models [26, 34]. Similarly, another study using the TPC-DS benchmark found that the same model successfully generated 75 out of 99 queries based on natural language descriptions [35]. These results highlight the variability in performance between different datasets and evaluation criteria.

In addition to execution success rate and execution time, future evaluations should incorporate standard text-to-SQL metrics. For instance, Execution Accuracy (ExecAcc) would measure whether the generated query returns the same result set as the gold-standard query, while Exact Match (EM) would check for syntactic identity. Furthermore, we recognize the value of general information retrieval metrics like **precision and recall**, which would capture how many of the returned results are correct and how many of the correct answers were successfully retrieved. Including these metrics would align our evaluation with established practices in the text-to-SQL literature, and provide a deeper understanding of the system's semantic correctness.

Furthermore, execution accuracy depends as much on prompt design as it does on the architecture or size of the model. A recent benchmarking study demonstrated that variations in prompt structure, context granularity, and the inclusion of schema-aware examples can substantially affect the quality of generated SQL [36]. Techniques such as few-shot prompting, schema linking, and context-aware formatting have been shown to improve execution accuracy in LLM-based systems.

This suggests that while system design affects how well it performs and how easy it is to use, the accuracy of the generated queries mostly depends on how the LLM is set up and how the prompts are written. To make the system more reliable in the future, it could use smarter prompt templates, better ways to summarize database structures, and user feedback to improve the results over time.

6 Conclusion and Future Work

This paper presented a system that integrates Large Language Models (LLMs) with SQL databases to enable natural language querying through a chatbot interface. By comparing microservice and monolithic architectures, we highlighted the trade-offs between modularity and performance. The microservice approach demonstrated higher success rates in generating valid SQL queries, especially for complex inputs, while the monolithic design offered faster response times due to reduced communication overhead. Our findings suggest that the choice between these architectures depends on the specific requirements of scalability, maintainability, and performance.

Through testing on the Mini-Dev dataset, we observed that execution times varied depending on architectural decisions, with the monolithic approach showing efficiency gains. However, LLM-generated SQL accuracy remains an open challenge, which requires further research.

Future work will focus on improving query validation techniques, exploring hybrid architectures that combine the strengths of both approaches, and evaluating the system with more diverse datasets. Furthermore, deploying LLMs locally for privacy-sensitive applications and incorporating user feedback loops could further enhance reliability and adaptability. Future work will also expand the evaluation metrics to include semantic measures, which would provide a more complete picture of the system's effectiveness beyond execution success and latency.

References

1. Documentation. pgadmin: Postgresql tools. Accessed: 2025-03-10. [Online]. Available: https://www.pgadmin.org/
2. ——. phppgadmin: phppgadmin. github. Accessed: 2025-03-10. [Online]. Available: https://github.com/phppgadmin
3. V. Zhong, C. Xiong, and R. Socher, "Seq2sql: Generating structured queries from natural language using reinforcement learning," *arXiv preprint arXiv:1709.00103*, 2017.
4. A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo *et al.*, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.
5. B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.
6. S. M. Khan, "Onion architecture used in software development," 2023.
7. X. Zhu, Q. Li, L. Cui, and Y. Liu, "Large language model enhanced text-to-sql generation: A survey," *arXiv preprint arXiv:2410.06011*, 2024.
8. R. Sun, S. Ö. Arik, A. Muzio, L. Miculicich, S. Gundabathula, P. Yin, H. Dai, H. Nakhost, R. Sinha, Z. Wang *et al.*, "Sql-palm: Improved large language model adaptation for text-to-sql (extended)," *arXiv preprint arXiv:2306.00739*, 2023.
9. D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, and J. Zhou, "Text-to-sql empowered by large language models: A benchmark evaluation," *arXiv preprint arXiv:2308.15363*, 2023.

10. R. Roberson, G. Kaki, and A. Trivedi, "Analyzing the effectiveness of large language models on text-to-sql synthesis," *arXiv preprint arXiv:2401.12379*, 2024.
11. J. Fan, Z. Gu, S. Zhang, Y. Zhang, Z. Chen, L. Cao, G. Li, S. Madden, X. Du, and N. Tang, "Combining small language models and large language models for zero-shot nl2sql," *Proceedings of the VLDB Endowment*, vol. 17, no. 11, pp. 2750–2763, 2024.
12. H. Li, J. Zhang, H. Liu, J. Fan, X. Zhang, J. Zhu, R. Wei, H. Pan, C. Li, and H. Chen, "Codes: Towards building open-source language models for text-to-sql," *Proceedings of the ACM on Management of Data*, vol. 2, no. 3, pp. 1–28, 2024.
13. S. Talaei, M. Pourreza, Y.-C. Chang, A. Mirhoseini, and A. Saberi, "Chess: Contextual harnessing for efficient sql synthesis," *arXiv preprint arXiv:2405.16755*, 2024.
14. Y. Gao, Y. Liu, X. Li, X. Shi, Y. Zhu, Y. Wang, S. Li, W. Li, Y. Hong, Z. Luo *et al.*, "Xiyan-sql: A multi-generator ensemble framework for text-to-sql," *arXiv preprint arXiv:2411.08599*, 2024.
15. T. Ren, Y. Fan, Z. He, R. Huang, J. Dai, C. Huang, Y. Jing, K. Zhang, Y. Yang, and X. S. Wang, "Purple: Making a large language model a better sql writer," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 15–28.
16. J. Lewis and M. Fowler. Microservices. Accessed: 2025-03-10. [Online]. Available: https://martinfowler.com/articles/microservices.html
17. O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing langchain: A primer on developing llm apps fast," in *International Conference on Applied Engineering and Natural Sciences*, vol. 1, no. 1, 2023, pp. 1050–1056.
18. Documentation. Huggingface: Serverless inference api. Accessed: 2025-03-09. [Online]. Available: https://huggingface.co/docs/api-inference/en/index
19. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825
20. W. Walecha, "Application of flask with python," *International Journal of Advances in Engineering and Management*, pp. 1665–1669, 2021.
21. C. Richardson. Pattern: Monolithic architecture. Accessed: 2025-03-12. [Online]. Available: http://microservices.io/patterns/monolithic.html
22. O. Al-Debagy and P. Martinek, "A comparative review of microservices and monolithic architectures," in *2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, 2018, pp. 000 149–000 154.
23. M. Ismail and G. E. Suh, "Quantitative overhead analysis for python," in *2018 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2018, pp. 36–47.
24. J. Palermo. The onion architecture : part 1. Accessed: 2025-03-14. [Online]. Available: https://jeffreypalermo.com/2008/07/the-onion-architecture-part-1/
25. Documentation. bird-bench: mini_dev. Accessed: 2025-03-20. [Online]. Available: https://github.com/bird-bench/mini_dev
26. J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo *et al.*, "Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls," *Advances in Neural Information Processing Systems*, vol. 36, pp. 42 330–42 357, 2023.
27. F. Lei, J. Chen, Y. Ye, R. Cao, D. Shin, H. Su, Z. Suo, H. Gao, W. Hu, P. Yin, V. Zhong, C. Xiong, R. Sun, Q. Liu, S. Wang, and T. Yu, "Spider 2.0: Evaluating

language models on real-world enterprise text-to-sql workflows," *arXiv preprint arXiv:2411.07763v2*, 2025.
28. L. Ma, K. Pu, and Y. Zhu, "Evaluating llms for text-to-sql generation with complex sql workload," *arXiv preprint arXiv:2407.19517v1*, 2024.
29. Documentation. Huggingface: Parametersv. Accessed: 2025-03-20. [Online]. Available: https://huggingface.co/docs/api-inference/en/parameters
30. J. Chun, C. S. de Witt, and K. Elkins, "Comparative global ai regulation: Policy perspectives from the eu, china, and the us," *arXiv preprint arXiv:2410.21279*, 2024.
31. C. Novelli, F. Casolari, P. Hacker, G. Spedicato, and L. Floridi, "Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity," *Computer Law & Security Review*, vol. 55, p. 106066, 2024.
32. S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2023, pp. 1–9.
33. K. Arai, "Design of on-premises version of rag with ai agent for framework selection together with dify and dsl as well as ollama for llm." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 12, 2024.
34. Documentation. bird-bench: Bird (big bench for large-scale database grounded text-to-sql evaluation). Accessed: 2025-03-24. [Online]. Available: https://bird-bench.github.io/
35. L. Ma, K. Pu, and Y. Zhu, "Evaluating llms for text-to-sql generation with complex sql workload," *arXiv preprint arXiv:2407.19517*, 2024.
36. B. Zhang, Y. Ye, G. Du, X. Hu, Z. Li, S. Yang, C. H. Liu, R. Zhao, Z. Li, and H. Mao, "Benchmarking the text-to-sql capability of large language models: A comprehensive evaluation," *arXiv preprint arXiv:2403.02951*, 2024.

Session 6

Artificial Intelligence in Medicine in Macedonia

Dragica Bliznakovska Stanchev[1], Smilka Janeska Sarkanjac[1], Sinisha Stanchev[2]
[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
[2] Institute of Immunobiology and Human Genetics, Medical Faculty, Skopje, Macedonia
[2] Institute of Computer Science, Information and Automation, Faculty of Electrical Engineering, Skopje

Abstract. Artificial Intelligence (AI) is redefining the foundations of healthcare systems worldwide, offering transformative potential in diagnostics, operational efficiency, and personalized medicine. However, the integration of AI in healthcare systems in a country transitioning from a state-controlled economy to a liberal one, such as Macedonia, remains limited and fragmented, hindered by infrastructural, educational, and ethical challenges. This study investigates the strategic and ethical readiness of Macedonia's healthcare sector for AI adoption through a mixed-methods approach, combining a systematic literature review and a national survey of 130 healthcare professionals. The results indicate a notable disparity between general awareness (72%) and actual practical experience (19%) with AI tools. While 85% of respondents expressed willingness to undergo formal training, concerns about data privacy (74%), algorithmic bias (42%), and explain ability (33%) persist.
The findings align with international frameworks such as the WHO report on AI governance and the EU's proposed Artificial Intelligence Act, which emphasize the need for transparency, fairness, and human-centric implementation (WHO, 2021; European Commission, 2021). This paper contributes empirical evidence from a South-eastern European perspective and advocates for national-level interventions including the development of legal and ethical frameworks, integration of AI into health education, and the initiation of pilot clinical projects.

Keywords: Artificial Intelligence, Healthcare Policy, Ethics, Transitional Systems, Macedonia, AI Readiness, GDPR, Digital Health.

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative force in modern medicine, significantly advancing diagnostic imaging, genomic analysis, clinical decision support, and patient stratification (Esteva et al., 2017). In high-income countries, AI technologies are actively deployed to improve healthcare outcomes, reduce administrative burdens, and support precision medicine initiatives (Topol, 2019; Rajpurkar et al., 2017).

In contrast, healthcare systems in transitional economies, such as Macedonia, are at the nascent stages of AI adoption. These systems face numerous challenges, including fragmented digital infrastructure, limited institutional capacity, lack of AI-specific regulation, and low public trust. Despite increasing awareness of AI among healthcare professionals, practical experience and systemic support remain limited.

This study investigates the readiness and strategic alignment of Macedonia's healthcare system for AI integration. Employing a mixed-methods approach, it examines ethical, infrastructural, and educational factors influencing AI adoption within a post-socialist, resource-constrained context. The findings contribute empirical evidence to the underexplored domain of AI readiness in South-eastern Europe and respond to international calls for equitable and ethical digital health implementation (World Health Organization [WHO], 2021; European Commission, 2021).

The integration of Artificial Intelligence (AI) in medicine has evolved significantly over the past three decades. Early applications focused primarily on expert systems for rule-based decision support, while recent advances have demonstrated the potential of machine learning (ML) and deep learning (DL) techniques across various clinical domains (Esteva et al., 2017).

AI has shown notable success in specialties such as radiology, dermatology, ophthalmology, and genomics. For example, Esteva et al. (2017) demonstrated dermatologist-level accuracy in skin cancer classification using convolutional neural networks. Similarly, AI has been effectively employed in radiology for anomaly detection in computed tomography (CT) and magnetic resonance imaging (MRI) scans (Rajpurkar et al., 2017), and in ophthalmology for diagnosing diabetic retinopathy through retinal image analysis (Gulshan et al., 2016). These technological advancements underscore AI’s potential to support diagnostic precision, reduce human error, and streamline clinical workflows.

However, the majority of these innovations have been realized in high-resource settings. Research focusing on AI implementation within transitional or resource-constrained healthcare systems, particularly in the Balkan region, remains scarce.

At the global level, institutions such as the NHS AI Lab in the United Kingdom and the France Health Data Hub are advancing large-scale AI integration by providing funding, standardized data platforms, and regulatory support for ethical AI development (NHS AI Lab, 2022; France Health Data Hub, 2023). Furthermore, the World Bank has emphasized the strategic importance of AI in digital health transformation across middle-income countries in its 2023 report, AI for Health: Global Trends and Country Perspectives (World Bank, 2023). These efforts are complemented by the WHO/ITU Focus Group on AI for Health, which provides guidance on validation frameworks, benchmarking, and cross-border collaboration (WHO/ITU, 2023).

Regarding Eastern Europian countries, Estonia and Poland are outstanding examples in this area. Estonia is recognized as a leader in Europe’s digital healthcare transformation. Its National Artificial Intelligence Strategy (Kratt Strategy) envisions integrating AI into public services, including healthcare. The strategy supports pilot projects with dedicated funding and clearly defined legal frameworks to facilitate AI adoption (Ministry of Social Affairs of Estonia, 2020). The AI Leap 2025 program aims to embed AI applications and skills into the education system, extending influence into healthcare

through the development of an AI-supported health information system. This system connects healthcare providers and enables real-time patient data access, improving service delivery and clinical decision-making.

Poland is strengthening primary healthcare through the pilot project PHC Plus (POZ Plus), which integrates additional services such as specialist outpatient care, physical therapy, health screenings, and chronic disease management. A key innovation is the appointment of care coordinators in each facility, ensuring continuous and comprehensive patient care. Poland also invests in telemedicine and e-health initiatives, including telemonitoring for patients with heart failure, aiming to enhance healthcare quality, accessibility, and reduce inequalities (European Commission, 2022; OECD, 2023).

Our study contributes to addressing this regional gap by examining both the readiness and ethical alignment of AI adoption in Macedonia's healthcare sector. By doing so, it not only adds empirical insights but also situates the findings within the broader European digital health landscape.

2 Survey

The survey was administered in early 2025 and included 130 respondents from various healthcare institutions. The questionnaire comprised 25 questions organized into five thematic categories: general AI knowledge, hands-on experience, perceived benefits, ethical and legal challenges, and expectations for institutional support. The survey instruments included Likert scales, multiple-choice items, and open-ended questions. Descriptive statistical analyses were performed using SPSS to summarize the collected data.

Table 1. Summary of Thematic Survey Results

Theme	Key Metric	Value (%)
Awareness of AI	General familiarity	72
Practical experience	Direct AI tool use	19
Training willingness	Support for AI training	85
Data privacy concern	Ranked top ethical issue	74
Algorithmic bias	Identified as concern	42

Table 1 summarizes key thematic survey results regarding healthcare professionals' awareness, experience, and concerns related to AI. The data reveals that a significant majority (72%) of respondents are generally familiar with AI concepts, yet only 19% have practical experience using AI tools. This gap highlights the early stage of AI adoption within the healthcare sector. Encouragingly, a strong 85% of participants expressed willingness to receive formal training, indicating a readiness to engage with AI technologies. Ethical concerns remain prominent, with 74% identifying data privacy as a top issue and 42% noting algorithmic bias as a significant challenge. These findings

underscore the urgent need for comprehensive ethical frameworks and governance structures to support responsible AI integration in healthcare.

Table 2. Distribution of AI Awareness and Experience by Profession

Profession	Familiar with AI (%)	Used AI Tools (%)
Doctors	85	25
Medical Technicians	68	12
Lab Analysts	74	19
Nurses	65	10
Other	60	15

Doctors report the highest familiarity with AI (85%) and the most frequent use (25%), while technicians and nurses report significantly lower engagement. These differences highlight the need for profession-specific training strategies to address knowledge gaps and ensure uniform digital transformation across clinical roles.

Table 3. Ranking of Ethical and Legal Concerns by Perceived Importance

Concern	Identified as Important (%)
Data Privacy and Confidentiality	74
Algorithmic Bias	42
Lack of Explainability	33
Accountability and Liability	28
Informed Consent	25

Participants ranked data privacy as the most critical concern (74%), followed by algorithmic bias (42%) and explainability (33%). Less emphasis on accountability (28%) and informed consent (25%) suggests gaps in understanding legal implications. These findings align with the need for national AI governance frameworks consistent with GDPR and WHO ethical guidelines.

Table 4. Barriers to AI Adoption by Respondent Group

Group	Regulatory Concern (%)	Technical Infrastructure (%)	Training Deficit (%)
Doctors	65	42	71
Lab Analysts	59	50	66
Technicians	52	58	74

Doctors are primarily concerned with regulatory issues (65%) and lack of training (71%). Technicians exhibit the highest training gap (74%), while lab analysts are most concerned with infrastructure (50%). These findings emphasize the importance of tailored policy interventions and educational efforts for different professional roles.

3 Discussion

The findings of this study reveal a paradox common to transitional healthcare systems: high awareness and interest in Artificial Intelligence (AI) contrasted by low levels of practical implementation and institutional support. While 72% of respondents acknowledged AI's diagnostic potential, only 19% reported actual experience with AI tools, underscoring a clear readiness gap. This discrepancy aligns with global trends highlighted by the World Health Organization (WHO, 2021) and the European Commission (2021), emphasizing that enthusiasm alone is insufficient without strategic infrastructure, governance, and educational foundations.

The high willingness (85%) to participate in AI-related training reflects a promising foundation for national policy development. However, without coordinated investment, standardized curricula, and cross-sector collaboration, this potential remains underutilized.

A comprehensive SWOT analysis reveals critical insights: strengths encompass high awareness and receptivity among healthcare professionals, with 85% expressing willingness to engage in AI training; weaknesses involve limited practical experience, absence of dedicated regulatory frameworks, and fragmented healthcare information technology systems; opportunities lie in harmonization with European Union AI regulations and GDPR, access to international support and funding, and potential for public-private partnerships alongside academic collaborations; threats include ethical concerns, diminished public trust, resistance from stakeholders apprehensive about automation, and the potential exacerbation of health inequalities in the absence of inclusive AI policies

Comparative studies from countries like Estonia and Poland demonstrate that pilot projects, public-private partnerships, and dedicated national strategies can significantly accelerate AI adoption in healthcare. In contrast, Macedonia currently lacks systemic initiatives integrating AI into clinical workflows, educational programs, or digital health legislation.

Beyond technical readiness, the ethical and legal dimensions of AI adoption remain underdeveloped. The absence of clear regulatory frameworks generates uncertainty surrounding accountability, data privacy, and patient consent. These issues echo ethical imperatives articulated by Floridi et al. (2018) and the WHO (2021), who advocate for transparency, fairness, and inclusive governance in deploying digital technologies in healthcare.

Overall, these findings underscore the urgent need for a coordinated, ethical, and strategic roadmap for AI in healthcare tailored to transitional system realities.

4 Policy Recommendations

Based on the findings of the national survey and comparative analysis with successful AI integration models in Estonia and Poland, as well as frameworks advocated by the World Health Organization (WHO) and WHO/ITU, we propose the following strategic measures to ensure responsible and effective integration of Artificial Intelligence (AI) in the healthcare system of Macedonia (WHO, 2021; WHO/ITU, 2023):

Develop a National AI-in-Health Strategy. The Ministry of Health, in collaboration with academic institutions and medical associations, should design a comprehensive national roadmap for AI adoption. This strategy must outline short-, medium-, and long-term goals and include provisions for infrastructure, funding, training, and evaluation metrics. Legal and ethical dimensions should be integrated from the outset (Ministry of Social Affairs of Estonia, 2020; European Commission, 2021).

Integrate AI Education in Health and Medical Curricula. Universities and professional training centers should implement formal courses on AI fundamentals, algorithmic bias, ethics, and clinical applications. Curricula should be aligned with international standards such as WHO guidelines and the EU AI Act to prepare future healthcare professionals for digital transformation (WHO, 2021; European Commission, 2021).

Establish Multidisciplinary Ethical Review Boards. National and institutional ethics committees should include experts in medicine, law, informatics, and bioethics. These bodies will oversee AI tool deployment and ensure compliance with principles of transparency, fairness, and accountability. They should also evaluate algorithms for potential bias and harm (Floridi et al., 2018; WHO, 2021).

Launch Pilot Projects and Innovation Hubs. Selected hospitals and health centers should serve as testbeds for AI implementation under controlled and measurable conditions. These pilots should be publicly funded but co-developed with private AI vendors and academic researchers, with transparent evaluation criteria (Ministry of Social Affairs of Estonia, 2020; European Commission, 2022).

Harmonize National Legislation with the EU AI Act and European Health Data Space (EHDS). Data protection laws, medical device regulations, and digital health policies must align with the EU's harmonized framework. Compatibility with the EHDS will enhance trust, facilitate cross-border health services, and support integration into EU-wide AI and data initiatives (European Commission, 2021; European Health Data Space, 2023).

Invest in Infrastructure and Ensure Equitable Access. AI integration must be inclusive, requiring upgrades to ICT infrastructure in under-resourced regions, especially rural and primary care institutions. Technical maintenance support and secure data exchange systems must also be provided (WHO, 2021).

Foster Public Awareness and Stakeholder Engagement. Public campaigns and professional dialogues are essential to build trust and address concerns related to automation and surveillance. Engagement with patients, professional associations, and civil society is crucial for sustainable AI governance.

Implement Monitoring and Evaluation Frameworks. A structured monitoring and evaluation framework should accompany each national AI initiative, featuring clear key performance indicators (KPIs) for effectiveness, bias mitigation, patient safety, and institutional compliance. Periodic public reporting and open access to outcomes will strengthen accountability and public trust (WHO/ITU, 2023).

Adopt an Ethics-by-Design Approach. Ethical principles such as fairness, transparency, explainability, and human oversight should be embedded throughout the design, development, and deployment phases of AI solutions. This proactive model, advocated by WHO and WHO/ITU, is vital for sustainable digital health integration (Floridi et al., 2018; WHO/ITU, 2023).

5 Conclusion

This study reaffirms the structural and ethical contradictions inherent in transitional healthcare systems, such as that of Macedonia, regarding the integration of Artificial Intelligence (AI). Although healthcare professionals demonstrate a high level of awareness and interest, the actual implementation of AI tools remains limited due to insufficient institutional support, regulatory preparedness, and practical experience.

The potential of AI—ranging from enhanced diagnostic accuracy to automation and personalized medicine—is widely acknowledged. Yet, its deployment is hampered by systemic, ethical, and educational obstacles. Addressing these barriers requires comprehensive strategies, including targeted training, robust ethical oversight, and harmonized digital health legislation aligned with frameworks such as the EU AI Act and GDPR (WHO, 2021; European Commission, 2021).

The recommendations provided in this study outline a realistic and ethical roadmap for AI adoption, centered on national policy reform, educational modernization, institutional governance, and controlled pilot testing. Comparative experiences from Estonia and Poland offer transferable models that Macedonia can customize for its healthcare infrastructure and socio-political context (Ministry of Social Affairs of Estonia, 2020; European Commission, 2022).

Participation in initiatives such as the European Health Data Space (EHDS), Horizon Europe, and EU4Health would enable access to funding, technical standards, and collaborative platforms. These programs can amplify national efforts and accelerate AI readiness through cross-border interoperability and shared best practices (European Health Data Space, 2023; European Commission, 2021).

Future research should explore the longitudinal effects of AI deployment, the efficacy of ethical review mechanisms, and public perception of AI in clinical settings. Building a value-based, responsible digital healthcare ecosystem demands not only data and technology but also visionary leadership, interdisciplinary collaboration, and public trust.

In conclusion, transitional healthcare systems like Macedonia's have a critical opportunity to develop and implement a forward-looking, ethically informed framework

for AI integration. This will ensure sustainable, equitable, and responsible digital transformation in medicine.

6 References

1. Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.
2. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
3. World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO, Geneva.
4. Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.
5. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Brussels.
6. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318.
7. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.
8. Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
9. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 1–9.
10. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.
11. Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. Journal of the Royal Society of Medicine, 112(1), 22–28.
12. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94–98.
13. Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current Cardiology Reports, 16(1), 441.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 1–9.
15. Ministry of Social Affairs of Estonia. (2020). Estonian National Strategy for Artificial Intelligence in Healthcare. Retrieved from https://www.sm.ee/en/artificial-intelligence-healthcare
16. European Commission. (2022). Digital Health and Care in Poland: Country Profile. Brussels: European Commission.
17. OECD & European Observatory on Health Systems and Policies. (2023). Estonia: Country Health Profile 2023; Poland: Country Health Profile 2023. OECD Publishing, Paris. Retrieved from https://www.oecd.org/en/publications/2023/12/estonia-country-health-profile-2023_89043bfe.html and https://www.oecd.org/en/publications/2023/12/poland-country-health-profile-2023_80434439.html
18. World Bank. AI for Health: Global Trends and Country Perspectives. World Bank Publications, 2023.

19. NHS AI Lab. Artificial Intelligence in Health and Care Awards. UK Government, Department of Health and Social Care, 2022.
20. France Health Data Hub. Annual Report on Digital Health Innovation. Paris: Ministère des Solidarités et de la Santé, 2023.
21. WHO/ITU. Focus Group on Artificial Intelligence for Health (FG-AI4H): Final Report. World Health Organization and International Telecommunication Union, Geneva, 2022.

Higuchi's Fractal Dimension in EEG signals of Children with Autism and Typical Development

Aleksandar Tenev[1*], Silvana Markovska-Simoska[2] and Igor Mishkovski[1]

[1*]Faculty of Computer Science and Engineering, Sts Cyril and Methodius University, Skopje, 1000, Republic of North Macedonia.
[2]Macedonian Academy of Sciences and Arts, Skopje, 1000, Republic of North Macedonia.

Abstract

Higuchi's method for computing the fractal dimension of a time-series is a powerful nonlinear tool for estimating underlying system complexity, offering advantages in speed, accuracy, and cost compared to traditional linear methods. It's application to brain signals analysis can offer valuable insights to the underlying neural mechanism and corresponding brain dynamics. Autism Spectrum Disorder (ASD) is an increasingly prevalent brain condition that affects social behavior, communication and interaction. In this study we analyze the fractal dimension of EEG signals obtained from 49 children with ASD and 39 children with typical development (TD) during resting state, channel-wise, using the Higuchi's method. Despite the promising results, we believe that the metric alone lacks evidence to detect underlying mechanisms in ASD due to the problem with EEG channel-specific optimization.

Keywords: ASD, Higuchi Fractal Dimension, Quantitative EEG, Complexity

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental condition characterized by persistent difficulties in social communication, along with restricted interests and repetitive behaviors. Although considerable research efforts have been made, the neurobiological mechanisms underlying ASD are still not fully understood, highlighting the need for innovative methods to characterize the altered neural dynamics linked to the disorder. Currently, global diagnostic standards such as the DSM-5 and ICD-11 rely primarily on behavioral assessments. However, the broad and flexible criteria defining ASD permit diverse combinations of symptoms, which increases heterogeneity but diminishes diagnostic precision by overlapping with other conditions, introducing subjectivity, and potentially leading to false positives, higher prevalence rates, and less individualized care [1, 2]. As diagnostic criteria have expanded to include more individuals, there has been a push toward identifying objective biomarkers, or endophenotypes, to delineate subtypes and clarify the disorder's spectrum [3, 4]. Quantitative electroencephalography (QEEG) has emerged as a promising neuroimaging method for probing brain function in ASD [5], offering excellent temporal resolution to investigate disrupted neural oscillations and connectivity patterns in affected individuals.

Traditional QEEG approaches, which usually emphasize spectral power and coherence-based connectivity analyses [6], may fall short in capturing the complex, nonlinear nature of neural dynamics. Critics argue that traditional band-specific power metrics can overlap across various disorders, confounding interpretations of brain function [7]. Given that the brain operates as a complex adaptive system, where emergent properties arise from intricate neural network interactions, linear methods may be insufficient to characterize such complexity, particularly in non-stationary EEG data [8].

Recent advances in complexity science and information theory have introduced sophisticated measures for assessing irregularity, randomness, and structural complexity in EEG signals [9]. Methods derived from nonlinear dynamics and chaos theory, which examine irregular and sensitive patterns in deterministic systems, have been explored to study brain activity, aiming to capture its unpredictable and intricate characteristics. Recent review recognizes Higuchi's Fractal Dimension (HFD) - a measure of signal complexity in time domain - for its speed, accuracy, and cost, as a prominent nonlinear method that stands out from traditional linear methods in neurophysiology, and that has played an important role in the analysis of biological signals for over two decades, helping to reveal the brain's physiological and pathological states [10]. While Higuchi method has been used in many EEG analysis as a feature for machine learning and to analyze the complexity and irregularity of EEG signals and reflect different states of the brain, such as wakefulness, sleep, or seizure [11–13], in the context of ASD, HFD aplication is scarce. One study indicated statistically significant differences between complexity measures such as HFD of brain states and tasks, yet a clear distinction based on ASD risk alone has not yet been established, advocating that further replication with larger sample sizes and inclusion of other complexity measures is needed [14]. Another recent study in the context of ASD made such an extensive analysis, combining multiple time domain features in clinical differentiation study between

excluded its values as a feature in the machine learning models that were build [15].

In this research, we do a sensitivity analysis of the parameter value to estimate the fractal dimension in the EEG of ASD and TD using Higuchi's method, and note the problem with the EEG channel-specific optimization of its parameter value.

2 Materials and Methods

2.1 Data and Procedure

EEG recordings were obtained from 88 children, including 39 typically developing (TD) participants and 49 children diagnosed with Autism Spectrum Disorder (ASD). The mean age of the ASD group was 6.18 years (SD = 1.98), while the TD group had a mean age of 5.35 years (SD = 2.31). Because the TD group consisted exclusively of male participants, sex was not considered in the statistical analyses due to the resulting imbalance between groups. ASD diagnoses were established through a joint assessment by a psychiatrist and a psychologist, with inclusion contingent upon agreement between both professionals. All children in the ASD group met the DSM-V diagnostic criteria, and additional developmental information provided by parents was used to support diagnostic accuracy. EEG data collection was conducted in Skopje, North Macedonia. Written informed consent was obtained from the parents or legal guardians of all participants, and the study protocol was approved by the Ethics Committee of the Faculty of Medicine at Ss. Cyril and Methodius University in Skopje (Ref. No. 03-4953/2).

Considering the behavioral characteristics of children with ASD, EEG signals were recorded during an eyes-open resting-state condition. Electrode placement followed the international 10/20 system using an Electro-Cap (Electro-Cap International). Data were collected from 19 scalp electrodes: Fp1, Fp2, F3, F4, F7, F8, Fz, C3, C4, Cz, T3, T4, T5, T6, P3, P4, Pz, O1, and O2.

Eye movement artifacts were monitored using electrooculogram (EOG) recordings obtained via two 9 mm tin electrodes placed above and below the right eye, referenced to Fpz and Oz. A rejection threshold of 50 µV was applied for artifact detection. The EEG amplifier settings included a band-pass filter from 0.53 to 50 Hz and a notch filter spanning 45–55 Hz. Electrode impedance was maintained below 5 $k\Omega$. Signals were digitized at a sampling frequency of 250 Hz and stored for offline processing using the Mitsar WinEEG software.

For the TD group, continuous EEG recordings of 3 minutes were obtained. In contrast, EEG acquisition in the ASD group varied between 7 and 30 minutes, depending on the child's ability to cooperate and tolerate the procedure, reflecting challenges in sustaining a stable recording environment. All recording sessions were monitored in real time by the second author, who documented observations and visually inspected the signals to identify artifacts and select valid data segments.

Preprocessing procedures were identical for all participants. EEG signals were first visually examined and manually cleaned to remove artifacts. For subsequent time-domain analysis, the longest uninterrupted artifact-free segment during which the child remained calmly seated was selected. To ensure uniformity across participants,

seconds. At a sampling rate of 250 Hz, this yielded 10,750 samples per channel and a total of 204,250 amplitude values per participant across the 19 EEG channels.

2.2 Higuchi Fractal Dimension

The Higuchi fractal dimension (HFD) is a method for estimating the fractal dimension of a time series or function. The fractal dimension is a measure of the complexity or irregularity of a shape and the way it fills space. A higher fractal dimension indicates a more complex shape. For example, a straight line has a fractal dimension of 1, a square has a fractal dimension of 2, and a cube has a fractal dimension of 3. However, some shapes are irregular and have fractional dimensions, such as a coastline, a snowflake, or a fern. These shapes are called fractals.

Within fractal geometry, the Higuchi fractal dimension (HFD) is used to estimate the fractal dimension of the graph representing a real-valued function or time series. [16]. This value is obtained through algorithmic approximation, which is why it is also referred to as the Higuchi method.

The Higuchi algorithm is well explained in the paper [17] and is used to analyze a signal that has been converted into a time series, denoted as $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$, from which new time series $\mathbf{X}_m^k$ are generated, defined as:

$$\mathbf{X}_m^k : \mathbf{x}(m), \mathbf{x}(m+k), \mathbf{x}(m+2k), \ldots, \mathbf{x}\left(m + \left[\frac{N-k}{k}\right] \cdot k\right),$$

where the expression in $[\cdot]$ denotes the integer part of the expression in brackets, $m = 1, 2, \ldots, k$ is the starting point of the newly generated time series, and k is the interval between points and takes values $k = 1, \ldots, k_{\max}$, where $k_{\max}$ is a parameter that must be optimized. This means that for each interval k, there are k sets of new time series. For example, if $k = 10$ and the original series has a length of $N = 1000$, the new series would be:

$$\begin{gathered}
\mathbf{X}_1^{10} : \mathbf{x}(1), \mathbf{x}(11), \mathbf{x}(21), \ldots, \mathbf{x}(991), \\
\mathbf{X}_2^{10} : \mathbf{x}(2), \mathbf{x}(12), \mathbf{x}(22), \ldots, \mathbf{x}(992), \\
\vdots \\
\mathbf{X}_{10}^{10} : \mathbf{x}(10), \mathbf{x}(20), \mathbf{x}(30), \ldots, \mathbf{x}(1000).
\end{gathered}$$

The length of these curves is calculated using:

$$L_m(k) = \left\{ \frac{\sum_{i=1}^{\left[\frac{N-m}{k}\right]} |\mathbf{x}(m+ik) - \mathbf{x}(m+(i-1)k)|}{N-1} \left[\frac{N-m}{k}\right] \cdot k \right\} k.$$

The expression $N - 1 \left[\frac{N-m}{k}\right] \cdot k$ is a normalization factor. The curve length for interval k is the average of all $L_m(k)$:

$$L(k) = L_m(k).$$

fractal with dimension HFD. HFD is determined from the slope of the line that fits the plot of $\ln(L(k))$ versus $\ln(1/k)$, that is:

$$D_{Higuchi} = \frac{\log(L)}{\log(1/k)} \tag{1}$$

where L is the total length of the curve, and k is the interval size.

3 Results

When calculating the fractal dimension value according to the Higuchi method (HFD), it is necessary to select a value for the parameter k_{max}. This value is empirically determined in several ways [17]: (1) a simulation is performed of the dependence of the fractal dimension value on k_{max} presented on a logarithmic scale, and the k_{max} value is selected where a local maximum appears; (2) if no local maximum appears, asymptotic convergence to a boundary value is observed; and (3) if the plot shows neither a local maximum nor an asymptotic value, then a statistical analysis is performed and k_{max} is chosen according to the statistically most significant results [18].

In Figures 1, 2, and 3, graphs of the dependence of HFD on the parameter k_{max} are presented. It should be noted that these figures are only an example, and for the purposes of the paper, such an analysis was performed for each participant across each of the 19 EEG channels and we visually confirmed all the results. From the results of the sensitivity analysis, according to the graphical display of the dependence of HFD on k_{max}, we observed that there are differences between the groups, within the groups, and even differences between different channels for a single participant. In other words, for each participant, and for each EEG channel, the optimal k_{max} value was determined differently. For these reasons, we conducted a statistical analysis with Mann-Whitney U test - a non parametric statistical test for testing statistical significance - across a range of k_{max} values in $[3, 25]$. The results of these tests are presented in Tables 1 and 2, where the p values from the statistical test of significance during the sensitivity analysis are shown for each k_{max} value.

Although there are combination of channels and parameter values for k_{max} that show statistical significance, the channel-wise diversity for parameter optimization represents an obstacle for comparing TD and ASD children based on HFD parameter. In Figure 4, the corresponding error plots are shown, across all channels, for $k_{max} = 24$.

Table 1 Sensitivity analysis for $k_{\max}$. Table values represent p-values from Mann-Whitney U test

$k_{\max}$ value	Fp1	Fp2	F7	F3	Fz	F4	F8	T3	C3
$k_{\max} = 3$	0.0000	0.0197	0.0924	0.0838	0.2388	0.4987	0.4647	0.0008	0.4007
$k_{\max} = 4$	0.0001	0.0374	0.2388	0.1363	0.1622	0.4872	0.8453	0.0026	0.4058
$k_{\max} = 5$	0.0001	0.0374	0.2388	0.1363	0.1622	0.4872	0.8453	0.0026	0.4058
$k_{\max} = 6$	0.0004	0.0942	0.8810	0.2388	0.0838	0.5280	0.8241	0.0178	0.4267
$k_{\max} = 7$	0.0011	0.1056	0.9241	0.2772	0.0872	0.4872	0.6144	0.0326	0.4320
$k_{\max} = 8$	0.0028	0.1488	0.9891	0.3240	0.1096	0.5103	0.5280	0.0438	0.5460
$k_{\max} = 9$	0.0028	0.1488	0.9891	0.3240	0.1096	0.5103	0.5280	0.0438	0.5460
$k_{\max} = 10$	0.0030	0.1915	0.9313	0.3195	0.1339	0.4058	0.5828	0.0593	0.6080
$k_{\max} = 11$	0.0034	0.1514	0.9457	0.1946	0.2280	0.2692	0.7819	0.0522	0.9964
$k_{\max} = 12$	0.0031	0.1462	0.8170	0.1514	0.3611	0.1915	0.8667	0.0326	0.7819
$k_{\max} = 13$	0.0019	0.1180	0.7336	0.1117	0.4536	0.1567	0.9674	0.0245	0.6208
$k_{\max} = 14$	0.0014	0.0855	0.6272	0.0872	0.5339	0.1137	0.9025	0.0157	0.4815
$k_{\max} = 15$	0.0011	0.0686	0.5520	0.0700	0.6337	0.0806	0.8100	0.0100	0.2895
$k_{\max} = 16$	0.0008	0.0568	0.4591	0.0511	0.7404	0.0593	0.7064	0.0069	0.2010
$k_{\max} = 17$	0.0008	0.0568	0.4591	0.0511	0.7404	0.0593	0.7064	0.0069	0.2010
$k_{\max} = 18$	0.0005	0.0468	0.3806	0.0429	0.8667	0.0448	0.6272	0.0039	0.1462
$k_{\max} = 19$	0.0005	0.0350	0.3151	0.0401	0.9819	0.0383	0.5766	0.0027	0.0806
$k_{\max} = 20$	0.0002	0.0202	0.2010	0.0290	0.7819	0.0217	0.4759	0.0015	0.0304
$k_{\max} = 21$	0.0002	0.0145	0.1764	0.0240	0.6796	0.0187	0.3855	0.0008	0.0192
$k_{\max} = 22$	0.0002	0.0117	0.1488	0.0212	0.6017	0.0141	0.3707	0.0006	0.0141
$k_{\max} = 23$	0.0001	0.0089	0.1180	0.0173	0.5581	0.0124	0.3285	0.0004	0.0092
$k_{\max} = 24$	0.0001	0.0092	0.1137	0.0145	0.5642	0.0097	0.3195	0.0003	0.0074
$k_{\max} = 25$	0.0001	0.0092	0.1137	0.0145	0.5642	0.0097	0.3195	0.0003	0.0074

Table 2 Sensitivity analysis for $k_{\max}$. Table values represent p-values from Mann-Whitney U test

$k_{\max}$ value	Cz	C4	T4	T5	P3	Pz	P4	T6	O1	O2
$k_{\max} = 3$	0.2042	0.2461	0.0137	0.1540	0.1946	0.1137	0.6017	0.4482	0.5520	0.3905
$k_{\max} = 4$	0.2652	0.2461	0.0283	0.0568	0.1978	0.1202	0.4929	0.3107	0.4007	0.2652
$k_{\max} = 5$	0.2652	0.2461	0.0283	0.0568	0.1978	0.1202	0.4929	0.3107	0.4007	0.2652
$k_{\max} = 6$	0.4373	0.3563	0.1540	0.0173	0.1292	0.1339	0.3806	0.1462	0.2979	0.1884
$k_{\max} = 7$	0.5520	0.4536	0.2316	0.0117	0.1540	0.1462	0.4427	0.1269	0.2979	0.1793
$k_{\max} = 8$	0.6272	0.5704	0.2812	0.0105	0.1706	0.2176	0.4759	0.1292	0.2853	0.1884
$k_{\max} = 9$	0.6272	0.5704	0.2812	0.0105	0.1706	0.2176	0.4759	0.1292	0.2853	0.1884
$k_{\max} = 10$	0.8311	0.7336	0.3516	0.0141	0.2351	0.2936	0.6208	0.1594	0.3469	0.2280
$k_{\max} = 11$	0.7959	0.9097	0.3240	0.0358	0.6144	0.6863	0.9457	0.3611	0.5220	0.4591
$k_{\max} = 12$	0.6663	0.6467	0.2853	0.0644	0.8810	0.9097	0.6863	0.4815	0.7200	0.5953
$k_{\max} = 13$	0.5460	0.4647	0.1823	0.1056	0.8382	0.6997	0.3707	0.7336	0.9097	0.8524
$k_{\max} = 14$	0.4214	0.2652	0.1315	0.1764	0.5520	0.4267	0.1884	0.9964	0.8810	0.8667
$k_{\max} = 15$	0.3330	0.1540	0.0998	0.3021	0.3285	0.2316	0.0806	0.7611	0.6208	0.5460
$k_{\max} = 16$	0.3021	0.0618	0.0631	0.4320	0.2245	0.1056	0.0358	0.5339	0.4162	0.3285
$k_{\max} = 17$	0.3021	0.0618	0.0631	0.4320	0.2245	0.1056	0.0358	0.5339	0.4162	0.3285
$k_{\max} = 18$	0.2772	0.0358	0.0334	0.5520	0.1315	0.0593	0.0153	0.3330	0.2424	0.1823
$k_{\max} = 19$	0.2772	0.0149	0.0217	0.7611	0.0744	0.0297	0.0051	0.2424	0.1412	0.1056
$k_{\max} = 20$	0.2461	0.0033	0.0114	0.7819	0.0182	0.0049	0.0012	0.1292	0.0489	0.0270
$k_{\max} = 21$	0.2692	0.0014	0.0069	0.6017	0.0108	0.0028	0.0009	0.0714	0.0290	0.0117
$k_{\max} = 22$	0.3021	0.0006	0.0044	0.4320	0.0066	0.0021	0.0005	0.0556	0.0192	0.0057
$k_{\max} = 23$	0.3021	0.0004	0.0035	0.3422	0.0049	0.0019	0.0002	0.0319	0.0089	0.0036
$k_{\max} = 24$	0.3376	0.0002	0.0032	0.2812	0.0023	0.0018	0.0001	0.0251	0.0066	0.0023
$k_{\max} = 25$	0.3376	0.0002	0.0032	0.2812	0.0023	0.0018	0.0001	0.0251	0.0066	0.0023

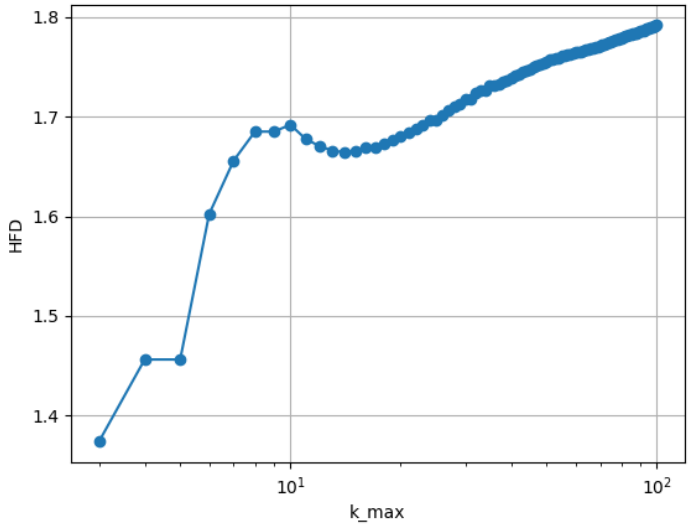

Fig. 1 *Influence of the choice of k_{max} on the value of the Higuchi fractal dimension. Here there is a local maximum at $k_{max} = 10$*

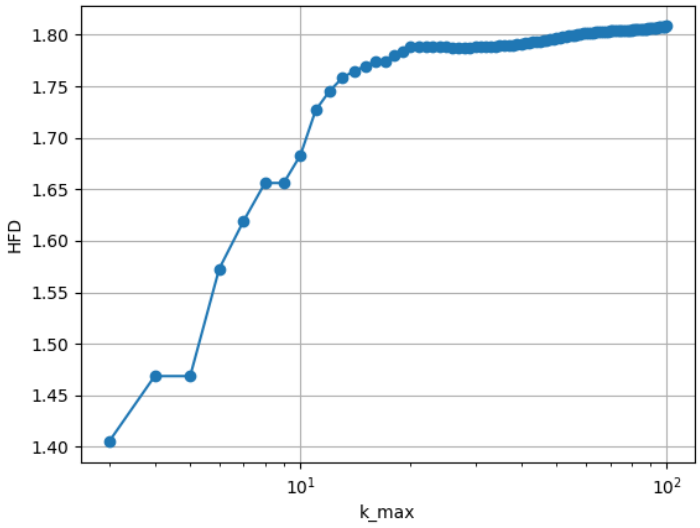

Fig. 2 Influence of the choice of $k_{\max}$ on the value of the Higuchi fractal dimension. Here there is an asymptotic value at $k_{\max} = 20$

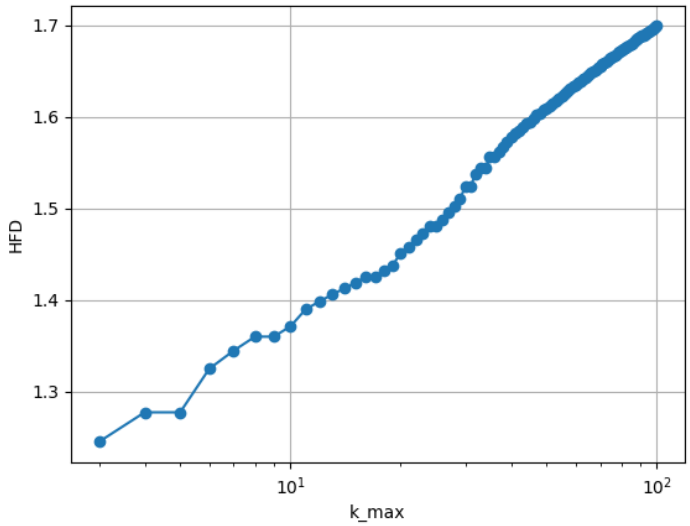

Fig. 3 *Influence of the choice of k_{max} on the value of the Higuchi fractal dimension. Here the curve has neither a local maximum nor an asymptotic value, so a statistical analysis is necessary to determine the value of k_{max}.*

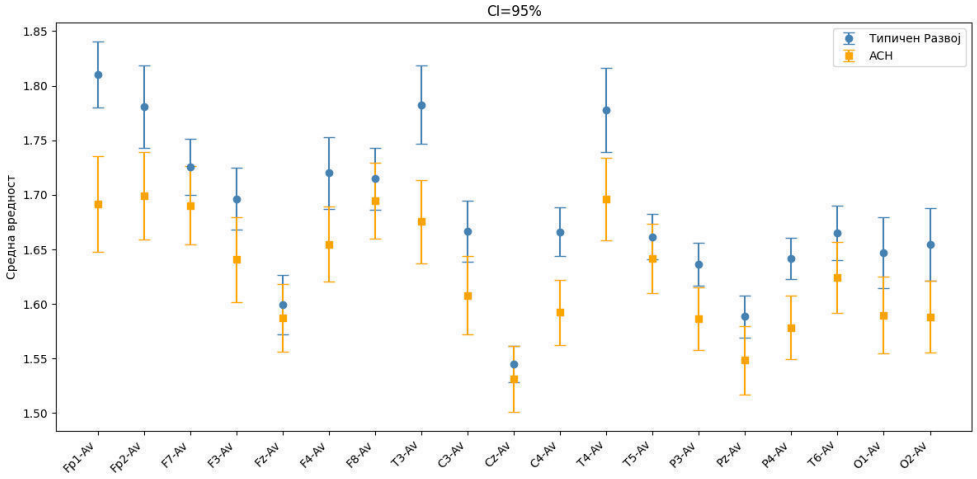

Fig. 4 *Error graphs for the Higuchi fractal dimension across all EEG channels. For this figure, the value $k_{max} = 24$ was used. It can be observed that the fractal dimension values are in the range of 1 to 2, which is expected for the fractal dimension of a time series*

The Higuchi Fractal Dimension (HFD) is a valuable nonlinear metric for quantifying the complexity of EEG signals; however, its application for comparing two groups is severely limited due to challenges in selecting the k_{max} parameter. The value of k_{max} has a critical influence on the estimated fractal dimension, and its optimal value varies not only across different EEG channels but also across individual participants. For each subject and each channel, the sensitivity of HFD to k_{max} may exhibit a local maximum, an asymptotic convergence, or no clear behavior at all, which forces the researcher to choose k_{max} in a subject-specific and channel-specific manner. As a consequence, the same EEG channel across different participants may rely on different methods or criteria for optimizing k_{max}, leading to inconsistent HFD estimates across the group. This lack of standardization makes it impossible to reliably compare group-level differences using HFD, since the extracted values are dependent on individually tuned parameters rather than reflecting a common measurement scale. Therefore, without a consistent and justifiable k_{max} optimization strategy applicable to all subjects and all EEG channels, HFD cannot be meaningfully used for inter-group comparisons.

References

[1] Bach, B., Vestergaard, M.: Differential diagnosis of icd-11 personality disorder and autism spectrum disorder in adolescents. Children **10**(6), 992 (2023)

[2] Kamp-Becker, I.: Autism spectrum disorder in icd-11—a critical reflection of its possible impact on clinical practice and research. Molecular Psychiatry **29**(3), 633–638 (2024)

[3] Waterhouse, L.: Heterogeneity thwarts autism explanatory power: A proposal for endophenotypes. Frontiers in psychiatry **13**, 947653 (2022)

[4] Rabot, J., Rødgaard, E.-M., Joober, R., Dumas, G., Bzdok, D., Bernhardt, B., Jacquemont, S., Mottron, L.: Genesis, modelling and methodological remedies to autism heterogeneity. Neuroscience & Biobehavioral Reviews **150**, 105201 (2023)

[5] Milovanovic, M., Grujicic, R.: Electroencephalography in assessment of autism spectrum disorders: a review. Frontiers in psychiatry **12**, 686021 (2021)

[6] Billeci, L., Sicca, F., Maharatna, K., Apicella, F., Narzisi, A., Campatelli, G., Calderoni, S., Pioggia, G., Muratori, F.: On the application of quantitative eeg for characterizing autistic brain: a systematic review. Frontiers in human neuroscience **7**, 442 (2013)

[7] Newson, J.J., Thiagarajan, T.C.: Eeg frequency bands in psychiatric disorders: a review of resting state studies. Frontiers in human neuroscience **12**, 521 (2019)

[8] Lo, M.-T., Tsai, P.-H., Lin, P.-F., Lin, C., Hsin, Y.L.: The nonlinear and

hilbert–huang transform. Advances in Adaptive Data Analysis **1**(03), 461–482 (2009)

[9] Von Wegner, F., Wiemers, M., Hermann, G., Tödt, I., Tagliazucchi, E., Laufs, H.: Complexity measures for eeg microstate sequences: concepts and algorithms. Brain Topography **37**(2), 296–311 (2024)

[10] Kesić, S., Spasić, S.Z.: Application of higuchi's fractal dimension from basic to clinical neurophysiology: a review. Computer methods and programs in biomedicine **133**, 55–70 (2016)

[11] Anderson, K., Chirion, C., Fraser, M., Purcell, M., Stein, S., Vuckovic, A.: Markers of central neuropathic pain in higuchi fractal analysis of eeg signals from people with spinal cord injury. Frontiers in Neuroscience **15**, 705652 (2021)

[12] Klonowski, W.: Fractal analysis of electroencephalographic time series (eeg signals). The fractal geometry of the brain, 413–429 (2016)

[13] Cukic, M., Pokrajac, D., Stokic, M., Radivojevic, V., Ljubisavljevic, M., et al.: Eeg machine learning with higuchi fractal dimension and sample entropy as features for successful detection of depression. arXiv preprint arXiv:1803.05985 (2018)

[14] Wolfson, S.S., Kirk, I., Waldie, K., King, C.: Eeg complexity analysis of brain states, tasks and asd risk. In: The Fractal Geometry of the Brain, pp. 733–759. Springer, ??? (2024)

[15] Tenev, A., Markovska-Simoska, S., Müller, A., Mishkovski, I.: Entropy, complexity, and spectral features of eeg signals in autism and typical development: a quantitative approach. Frontiers in Psychiatry **16**, 1505297 (2025)

[16] Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D: Nonlinear Phenomena **31**(2), 277–283 (1988)

[17] Wanliss, J., Wanliss, G.E.: Efficient calculation of fractal properties via the higuchi method. Nonlinear Dynamics **109**(4), 2893–2904 (2022)

[18] Olejarczyk, E.: Analysis of eeg signals using fractal dimension. Warsaw: PhD Thesis, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences (2003)

Identifying medical ontologies in MIMIC-IV dataset

Zorica Karapancheva[1], Mirjana Sadikovikj[2], and Goran Velinov[1]

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia
[2] Technische Universität Wien (TU Wien), Vienna, Austria

Abstract. The effective reuse of critical care Electronic Health Records (EHRs) for secondary analyses and clinical decision-support applications depends on the use of standardized data models and ontologies. The Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset provides a comprehensive, modular representation of intensive care units (ICUs) and hospital data, serving as a foundational resource for such research. To support effective data integration, the dataset's embedded coding systems must be carefully mapped to established medical ontologies. This paper investigates the presence, structure, and implementation of key medical ontologies, including SNOMED CT, LOINC, ICD, RxNorm, CPT4, and NDC — within MIMIC-IV. We review the native use of ICD for diagnoses and procedures, the systematic mapping of laboratory tests to LOINC, and the incorporation of NDC and generic sequence numbers for medications, which can be resolved to RxNorm identifiers. Through detailed identification and contextual mapping of ontologies in MIMIC-IV, this paper establishes a foundation for improved semantic interoperability, allowing better patient categorization and integration of machine learning models in critical care research.

Keywords: MIMIC-IV · SNOMED CT · LOINC · ICD · RxNorm · CPT4 · NDC.

1 Introduction

Intensive care units (ICUs) generate large amounts of clinical data through continuous monitoring and complex interventions [1]. However, establishing clear links between clinical decisions and patient outcomes remains challenging - connections that are essential for enabling large language models (LLMs) to support decision-making in critical care. The complexity is primarily attributed to inconsistencies in data architectures, individualized treatment decisions, and the wide-ranging characteristics of patients in intensive care. Despite the constant flow of data in ICUs, the evidence used by critical care physicians to guide decisions remains limited [2].

The main **research question** of this study is: How can established medical ontologies be identified and conceptually mapped within the MIMIC-IV dataset

to enhance semantic interoperability and support downstream clinical research and machine learning applications?

With growing volumes of clinical data from electronic health record (EHR) systems and evolving data mining capabilities, reusing existing data has a potential for advancing clinical practice [3,4]. Access to critical medical data improves clinical decision making, precision medicine, and early detection of complications, as well as efficient recruitment for large-scale multicenter studies at minimized costs [5]. With respect to the variety of healthcare datasets in use, the Medical Information Mart for Intensive Care IV (MIMIC-IV) represents a significant advancement offering expanded patient coverage, data granularity and research utility compared to previous versions and other datasets [6].

MIMIC-IV adopts a modular structure designed to reflect the complexity and diversity of clinical data. As mentioned in [7], "Database modeling is the process of determining how data are to be stored in a database". It provides a structured and standardized approach for defining resource properties, such as data types, constraints, and metadata, which ensures consistent representation and relationship modeling in data-driven applications. Research indicates that implementing a Common Data Model (CDM) supports large - scale research and exploitation of rare diseases or rare events, which accelerates scientific progress through the sharing of methodologies, source code and analytical tools [8, 9]. Through the formal specification of medical concepts and their relations, ontologies contribute significantly to improving semantic consistency and interoperability across varied clinical datasets [10].

AI-driven clinical research and secondary analyses increasingly rely on routinely collected EHRs. However, the lack of standardization in data schemas and semantic models presents a major challenge. This highlights the urgent need for adopting common ontologies and harmonized data structures. In this context, our study aims to identify and map established ontologies from the medical domain within the MIMIC-IV dataset, with the objective of improving future research efficiency through the incorporation of standardized and widely accepted clinical concepts.

The remainder of this paper is organized as follows: Section 2 presents the literature review on clinical data standardization and semantic integration. Section 3 provides an overview of the MIMIC-IV database, including its model and structure. Section 4 discusses medical ontologies and details their identification within MIMIC-IV. Finally, Section 5 offers a discussion of the findings, highlighting the challenges, implications, and potential directions for future work.

2 Literature Review

The MIMIC database family, with its most recent MIMIC-IV iteration, has become one of the most widely used open-access critical care datasets, providing a robust relational schema that encapsulates patient demographics, ICU stays, laboratory measurements, microbiology, billing codes, and prescriptions [6]. Its relational model and rich multi-table structure have been extensively documented

and visualized, such as by Barrios who offers comprehensive ER diagrams highlighting the foreign key relationships fundamental to its design [11]. In related work, De Nicola et al. examine ontology alignment for clinical diagnosis using SNOMED CT and ICD-10-CM within knowledge graphs, emphasizing issues and usage scenarios applicable to semantic representation and reasoning over EHRs [12].

Significant research has focused on improving the semantic interoperability of MIMIC data by integrating it with widely accepted data models. The work of Paris and Parrot, and Paris et al. represents a foundational effort to transform MIMIC (originally structured largely around ICD-9-CM for diagnoses and local codes for laboratories and medications) into the OMOP Common Data Model (CDM) [9,13]. Through this intermediate structure, they achieved notable mapping rates, with approximately 78% of source concepts aligned to standard ontologies including SNOMED CT for clinical conditions, LOINC for lab tests, RxNorm for drugs, and CPT4 for procedures. These transformations not only standardized data semantics but also enabled the execution of tens of thousands of reproducible queries in collaborative data analysis settings, underscoring the practical feasibility of ontology-driven harmonization.

Complementing these efforts, recent studies have demonstrated the portability of the MIMIC-IV schema beyond its original US hospital context. For example, Giesa et al. leveraged MIMIC-IV's relational structure to store EHR data from a German ICU system, successfully mapping more than 154 million records while navigating discrepancies such as German billing code systems not directly translatable to US-centric ICD-9-CM or CPT4 ontologies [14]. The study underscored the strengths and limitations of applying the MIMIC schema in international environments. It emphasized the importance of developing context-specific mapping strategies to maintain semantic accuracy and relevance.

In the broader context of ontological infrastructure, tools like the EMBL-EBI Ontology Lookup Service (OLS) significantly improve access to diverse biomedical ontologies [15]. These include widely used standards such as SNOMED CT, LOINC, and RxNorm, which in turn improve consistent annotation and enhance interoperability in datasets like MIMIC. Further, datasets derived from MIMIC-IV continue to explore the distribution of ICD-coded diagnoses and procedures across patient populations. For example, the work by Mathew et al. in Nature Scientific Data demonstrates practical applications of these coding systems in both epidemiological characterization and the development of machine learning models [16,17].

Collectively, the literature underscores a clear trajectory: integrating clinical datasets like MIMIC-IV with standardized ontologies is crucial to enabling robust, scalable, and internationally interoperable research. By enabling reproducible research across institutions, this approach also advances methods for grouping patient populations, detecting time-based clinical patterns, and performing complex temporal analyses—key components for the development of AI-driven healthcare applications.

Beyond ontology alignment, prior efforts have structurally harmonized MIMIC into community standards such as the OMOP Common Data Model and HL7 FHIR. Transformations to OMOP demonstrate substantial standardization and concept mapping rates that enable scalable, reproducible analytics across sites [18]. Complementarily, MIMIC-IV on FHIR exposes resources for interoperable applications and SMART-on-FHIR tooling [19]. Our work is complementary: rather than proposing a full CDM conversion, we document and analyze ontology-level mappings within the native MIMIC-IV schema to support semantic enrichment and downstream tasks.

3 MIMIC-IV

MIMIC-IV is an extensive, de-identified clinical dataset collected from the Beth Israel Deaconess Medical Center in Boston, providing comprehensive patient-level data for research, as reported by Johnson et al. [16]. Spanning the years 2008 to 2019, the MIMIC-IV dataset offers a rich foundation for critical care research by capturing detailed clinical information across a broad patient population. It includes data from more than 65,000 intensive care unit (ICU) admissions and over 200,000 emergency department visits. This comprehensive dataset includes diverse data types such as vital signs, laboratory results, medications, and clinical notes, enabling multifaceted analyses in epidemiology, outcome prediction, and the development of machine learning models for critical care decision support.

MIMIC-IV is architected into modules that encapsulate specific domains of medical data. The **Core** Module contains fundamental entities such as patient demographics, hospital admissions, and intrahospital transfers, serving as primary identifiers and relational anchors. The **Hospital** Module, also referred to as the *hosp* module, contains granular clinical process data, including laboratory measurements, medication administration records, and procedure codes, facilitating detailed longitudinal analyses. The **ICU** Module provides high-resolution time-series data capturing critical care interventions, encompassing physiologic monitoring parameters, ventilator settings, and therapeutic device usage. To ensure compliance with privacy regulations, all direct identifiers have been de-identified per HIPAA Safe Harbor standards, and temporal data are uniformly time-shifted on a per-subject basis to preserve event chronology while mitigating re-identification risk. The dataset is distributed in CSV format, optimizing interoperability with diverse analytical frameworks.

Given the nature of the dataset, our ontology identification efforts focused specifically on the *hosp* module, which includes 22 tables representing various categories of clinical data:

- **Patient tracking** - patients, admissions, transfers, provider
- **Administration** - services, poe, poe_detail
- **Billing** - diagnoses_icd, d_icd_diagnoses, drgcodes, hcpcsevents, d_hcpcs, procedures_icd, d_icd_procedures
- **Measurement** - microbiologyevents, labevents, d_labitems, omr
- **Medication** - emar, emar_detail, pharmacy, prescriptions

3.1 Data structure

Since the MIMIC-IV dataset is originally available in CSV format, restructuring it into a normalized relational schema was a necessary step for efficient ontology-based concept extraction. Unlike its previous version, MIMIC-III, the latest version lacks an official diagram or formally documented schema, which made the task of understanding table relationships more challenging. As a result, we manually identified primary and foreign keys and examined functional dependencies between tables to reconstruct an accurate relational model for subsequent ontology mapping and semantic integration.

As part of this effort, particular attention was given to the **hosp** module, which contains the majority of hospital-related events and patient interactions during inpatient stays. Most tables in this module include the column *subject_id*, a unique identifier assigned to each patient, and *hadm_id*, which uniquely identifies each hospital admission and maps back to a single patient. Although hadm_id alone can be used to link records to patients, both identifiers are commonly used as a composite key. This design aligns with MIMIC's modeling principles and optimization of the data structure for patient-centered research. In tables that contain multiple records per hospital admission - such as procedures, diagnoses, or transfers — an additional column is often included to ensure row-level uniqueness (e.g., *seq_num*, *drg_type*, or *transfer_id*). These columns help define composite primary keys, which are essential for ensuring consistent relationships between tables and accurate data integration in further analyses. A comprehensive overview of all tables within the hosp module, along with their identified primary key columns, is presented in Table 1.

Having identified the primary keys of the tables, the next step was to determine the foreign keys in the context of managing inter-table relationships. To facilitate this process, the tables were first grouped into functional categories based on the type of data they contain. As previously noted, the hosp module is inherently patient-oriented, with a strong emphasis on hospital admissions.

At the core of this schema are three foundational tables - **patients**, **admissions** and **omr** (online medical record) - which are highlighted in red in Figure 1. These form the structural backbone of the module. Surrounding them is a layer of secondary tables containing detailed information for each hospital admission. These tables are typically linked to the admissions table through a composite foreign key consisting of subject_id and hadm_id. Most of these tables, marked in blue in the diagram, store multiple records per admission. For instance, the procedures_icd table captures all procedures billed during a patient's stay and thus includes a third column—such as seq_num—to complete its composite primary key and ensure row-level uniqueness.

Additionally, several of these detailed tables form many-to-one dictionary relationships with reference tables, creating a referential data structure indicated in green. These relationships provide semantic meaning to coded data through standardized definitions. For example, the hcpcsevents table (which stores hospital billing codes) links to the d_hcpcs dictionary table. While many hcpcsevents entries can refer to the same d_hcpcs code, each code appears only once in the

Table 1. Database tables in MIMIC-IV and their primary keys (PKs)

Table name	Primary Key(s)
patients	subject_id
omr	subject_id, chartdate, seq_num, result_name
admissions	subject_id, hadm_id
services	subject_id, hadm_id, transfertime, curr_service
diagnoses_icd	subject_id, hadm_id, seq_num, icd_version
procedures_icd	subject_id, hadm_id, seq_num, icd_version
hcpcsevents	subject_id, hadm_id, seq_num
drgcodes	subject_id, hadm_id, drg_type
emar_detail	emar_id, emar_seq, parent_field_ordinal
d_icd_diagnoses	icd_code, icd_version
d_icd_procedures	icd_code, icd_version
transfers	transfer_id
d_hcpcs	code
labevents	labevent_id
d_labitems	itemid
microbiologyevents	microevent_id
pharmacy	pharmacy_id
prescriptions	pharmacy_id, drug_type, drug
emar	emar_id
poe	poe_id
poe_detail	poe_id, field_name
provider	provider_id

dictionary, where it is annotated with descriptive metadata. This design allows for the consistent interpretation of standardized codes across events.

Another subset of the schema (highlighted in orange) involves one-to-many hierarchical relationships, forming parent-child data structures. A clear example is the relationship between the poe (provider order entry) and poe_details tables. A single medical order (poe_id) may have multiple associated detail records—such as components for fluid type, administration rate, and duration. These details are stored in the field_name and field_value columns of the poe_details table, allowing for a flexible representation of structured order information.

Finally, there is the provider table, which appears unlinked in the relational diagram highlighted in purple. This table contains only a list of deidentified provider codes without accompanying metadata. However, it serves as a reference for multiple other tables containing columns with the suffix provider_id (e.g.,

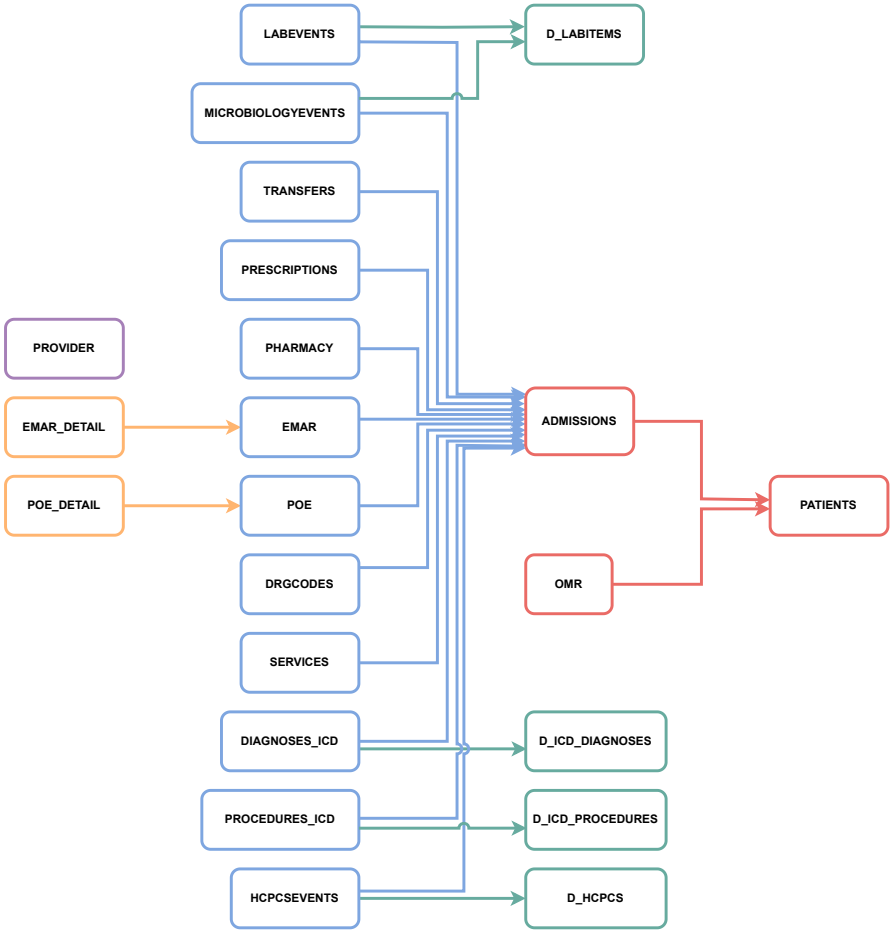

Fig. 1. MIMIC-IV diagram schema

order_provider_id in prescriptions table and admit_provider_id in admissions table). Although not explicitly linked in the schema, these connections can be inferred and used to enrich provider-level analyses when needed.

4 Medical Ontologies and Implementation in MIMIC-IV

By formally organizing clinical concepts and their interrelations, medical ontologies ensure semantic consistency across healthcare systems and enable advanced data analysis, integration, and interoperability [20]. Among the most widely adopted are SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), ICD (Inter-

national Classification of Diseases), RxNorm, CPT4 (Current Procedural Terminology) and NDC (National Drug Code). The following subsections provide an in-depth overview of each ontology, detailing their structure, scope, and specific applications within clinical data standardization and research contexts, along with a description of how each has been mapped to the MIMIC-IV database to enhance semantic interoperability and support downstream analytics.

4.1 SNOMED CT

SNOMED CT is a comprehensive, multilingual clinical terminology that provides standardized codes, terms, synonyms, and definitions used to represent clinical content in electronic health records (EHRs) [21]. It is organized as a hierarchical ontology, enabling detailed representation of diseases, findings, procedures, microorganisms, pharmaceuticals, and other clinical concepts. SNOMED CT supports complex querying and reasoning over clinical data by defining relationships such as *is a*, *part of*, and *associated with*, which facilitates semantic interoperability and advanced clinical decision support.

The analysis focused on a subset of SNOMED CT hierarchies that are particularly applicable to hospital-based clinical data. These include the following concepts of the ontology - Clinical Finding, Procedure, Substance and Pharmaceutical/Biological Product, Observable Entity, Organism, and Environment / Geographical Location. Other hierarchies, such as Social Context or Specimen, were not considered due to their limited or implicit representation in the available data.

To support this conceptual mapping, relevant tables from the hosp module were reviewed. These include *diagnoses_icd* and *procedures_icd* for diagnosis and procedure codes, *prescriptions* and *pharmacy* for medication-related data, *labevents* and *d_labitems* for laboratory test results, and *microbiologyevents* for microbiological findings. Additional contextual information was drawn from the *drgcodes* and *services* tables. Each of these tables contains structured clinical information that could potentially be aligned with corresponding SNOMED CT concepts.

The proposed mapping strategies combine both code-based and text-based approaches. For example, ICD-9/10 codes found in diagnosis and procedure tables can be mapped to SNOMED CT using resources such as the UMLS Metathesaurus or publicly available crosswalks [22]. For fields containing free-text descriptions — such as drug names or lab test labels — semantic annotation tools like the BioPortal Annotator may be employed [23]. In cases where intermediary vocabularies such as RxNorm or LOINC are available, these can serve as bridges to SNOMED CT, particularly for medications and laboratory tests.

The mapping process focuses primarely on columns that contain semantically rich clinical information, such as icd_code, drug, itemid, and test_name. Columns related to administrative metadata or identifiers would generally be excluded from the ontology alignment process, as they do not contribute directly to clinical concept representation.

4.2 LOINC

LOINC is a universal standard for identifying laboratory and clinical observations, including lab tests, measurements, and clinical assessments [24]. It provides unique codes linked to detailed metadata specifying the test method, specimen type, and measured property. This granular coding enables consistent interpretation and exchange of laboratory data across diverse healthcare information systems, supporting interoperability in clinical data exchange, public health reporting, and research applications.

The focus of this analysis was on identifying MIMIC-IV tables that contain data corresponding to LOINC's core concept types, especially laboratory test results and microbiological observations. The most relevant SNOMED CT concept hierarchy in this context - Observable Entity - is also closely aligned with LOINC's scope, but LOINC provides more granular and standardized identifiers for lab tests and clinical observations.

The primary tables identified for LOINC mapping include *labevents* and *d_labitems*, which contain laboratory test results and metadata, respectively. These tables provide test identifiers (itemid), test names (label), and associated measurement values. Additionally, the microbiologyevents table contains structured information about microbiological tests, including organism names, antibiotic susceptibility results, and test types, which may also be mapped to LOINC categories where applicable.

The proposed mapping strategy involves associating MIMIC-IV test identifiers and names with corresponding LOINC codes. This can be approached through text-based matching of test names to LOINC descriptions, supported by tools such as the LOINC Search API or the Regenstrief LOINC mapping assistant [25]. Where available, mappings from MIMIC-IV itemid values to LOINC codes can be leveraged to improve accuracy. In cases where exact matches are not available, approximate or hierarchical mappings may be considered based on test category, specimen type, and measurement units.

The mapping would primarily rely on fields such as itemid, label, fluid, and category in d_labitems, and test_name and org_name in microbiologyevents. These fields contain the semantic content necessary for identifying the nature of the observation and aligning it with a LOINC concept. Limitations include the potential ambiguity of test names, the absence of standardized units in some cases, and the challenge of distinguishing between similar tests without additional context.

4.3 ICD

ICD, maintained by the World Health Organization (WHO), is a globally recognized diagnostic tool for epidemiology, health management, and clinical purposes [26]. The ICD ontology organizes diseases, symptoms, abnormal findings, and external causes of injury into a hierarchical classification system. Versions such as ICD-9 and ICD-10 have been widely used for billing, morbidity, and

mortality statistics, while ICD-11 introduces enhanced digital compatibility and increased clinical detail.

The ICD ontology is divided into two main branches relevant to this analysis: ICD-9-CM and ICD-10-CM for diagnoses, and ICD-9-PCS and ICD-10-PCS for procedures [27]. These classifications provide hierarchical groupings of clinical conditions and interventions, enabling standardized reporting and analysis. In the context of MIMIC-IV, these codes are already embedded in the dataset, primarily within the *diagnoses_icd* and *procedures_icd* tables.

The diagnoses_icd table contains diagnosis codes assigned during hospital admissions, while the procedures_icd table includes procedural codes. Both tables are accompanied by dictionary tables (d_icd_diagnoses and d_icd_procedures) that provide textual descriptions of the codes. These resources allow for direct mapping to the ICD ontology without the need for intermediary vocabularies or external annotation tools.

The conceptual mapping focuses on the use of fields such as icd_code and icd_version, which indicate the specific ICD code and whether it belongs to the ICD-9 or ICD-10 system. These fields are semantically rich and directly correspond to entries in the ICD classification. Other columns in the tables, such as patient identifiers or timestamps, were not considered part of the ontology mapping process.

Unlike SNOMED CT or LOINC, where mappings often require translation or semantic annotation, the use of ICD in MIMIC-IV is native and explicit. However, challenges still exist. For example, the transition from ICD-9 to ICD-10 introduces differences in code structure and granularity. Additionally, while ICD provides a robust framework for diagnoses and procedures, it does not cover other clinical domains such as medications, lab tests, or observations, limiting its scope in comprehensive ontology integration.

4.4 RxNorm

RxNorm is a standardized nomenclature for clinical drugs produced by the U.S. National Library of Medicine [28]. It provides normalized names and unique identifiers for medications, including active ingredients, strengths, dose forms, and brand names. RxNorm facilitates interoperability between pharmacy management systems, EHRs, and clinical research databases by enabling consistent drug identification and mapping across different vocabularies.

RxNorm is particularly relevant for representing medications at various levels of specificity, including ingredients, clinical drugs, and branded products. In the context of MIMIC-IV, medication data is primarily found in the *prescriptions* and *pharmacy* tables, which contain structured information about drug names, dosages, routes of administration, and coded identifiers such as NDC (National Drug Code) and GSN (Generic Sequence Number).

The conceptual mapping focuses on aligning these medication records with RxNorm concepts. Fields such as drug, ndc, and gsn serve as potential entry points for mapping. For example, NDC codes can be translated to RxNorm Concept Unique Identifiers (RxCUIs) using publicly available crosswalks or the

RxNorm API [29]. In cases where structured codes are unavailable, drug names may be normalized and matched to RxNorm terms using text - based tools such as the RxNorm Normalized String Search or the BioPortal Annotator [23].

The mapping strategy emphasizes the use of semantically meaningful fields while excluding administrative or metadata columns that do not contribute to the clinical representation of the medication. The goal is to identify the most specific RxNorm concept that accurately reflects the prescribed or administered drug, including its form, strength, and route when available. Challenges include the variability in drug naming conventions, incomplete or missing identifiers, and the need for determining when multiple RxNorm concepts may correspond to a single drug entry.

4.5 CPT-4

CPT4 is a medical ontology maintained by the American Medical Association (AMA) that describes medical, surgical, and diagnostic services [30]. It is widely used for billing and documentation of clinical procedures in the United States. The ontology's hierarchical structure supports classification of procedures into categories and subcategories, allowing precise representation of healthcare services for administrative, research, and quality reporting purposes.

Although MIMIC-IV primarily uses ICD-9 and ICD-10 procedure codes, the CPT-4 ontology offers an alternative and often more granular representation of outpatient and physician-performed procedures. As such, this conceptual mapping explores the potential of translating MIMIC-IV procedural data into CPT-4 terms, particularly for use cases that require alignment with U.S. billing standards or integration with systems that rely on CPT coding.

The primary source of procedural data in MIMIC-IV is the *procedures_icd* table, which contains ICD-coded procedures along with associated metadata. While these codes are not CPT-4 by default, crosswalks exist between ICD-9/10-PCS and CPT-4, which can be used to approximate mappings [31]. These mappings are not always one-to-one, and differences in coding granularity and intent (clinical vs. billing) must be considered. Fields such as icd_code and icd_version are central to this process. In cases where direct mappings are unavailable, text-based descriptions from the *d_icd_procedures* dictionary table may be used to support semantic matching with CPT-4 concepts.

Limitations of this approach include the potential for mismatches due to differences in coding systems, the lack of CPT-specific data in MIMIC-IV, and the uncertainty in mapping between classification systems designed for different purposes. Nevertheless, this conceptual alignment provides a foundation for integrating MIMIC-IV data with CPT-4, particularly in contexts where procedural standardization or billing-related analysis is required.

4.6 NDC

The National Drug Code system is a universal product identifier for medications approved by the U.S. Food and Drug Administration (FDA) [32]. It en-

codes manufacturer, product, and package information into a unique numeric code. The NDC ontology enables tracking and standardization of pharmaceutical products in healthcare supply chains, pharmacy systems, and clinical research databases. In MIMIC-IV, medication data is primarily found in the *prescriptions* and *pharmacy* tables. These tables include fields such as ndc, drug, gsn, and formulary_drug_cd, which provide structured and semi-structured information about prescribed and administered medications. Among these, the ndc field directly corresponds to the NDC system and serves as a key element for mapping.

The conceptual mapping focuses on leveraging the ndc field to align MIMIC-IV medication entries with the NDC ontology. This process involves validating and standardizing NDC codes, which may appear in various formats (e.g., 10-digit or 11-digit representations). Tools such as the FDA's NDC Directory or RxNorm APIs can be used to verify and enrich NDC entries with additional metadata, such as drug names, manufacturers, and packaging details [33].

While the presence of NDC codes in MIMIC-IV allows for direct mapping, challenges may arise due to missing or malformed codes, inconsistencies in formatting, or the use of deprecated identifiers. In such cases, supplementary fields like drug or gsn may be used to identify the correct NDC or to support mapping through intermediary vocabularies such as RxNorm. The mapping strategy emphasizes the use of clinically meaningful fields while excluding administrative or non-semantic columns. The goal is to ensure that each medication entry is accurately represented by a valid NDC code, enabling downstream applications such as drug utilization studies and interoperability with pharmacy systems.

5 Discussion

The integration of clinical data with standardized medical ontologies is a critical step toward enabling semantic interoperability, enhancing data quality, and supporting advanced analytics in healthcare research. This paper explored the conceptual mapping of the MIMIC-IV hosp module to several widely used ontologies, including SNOMED CT, LOINC, ICD, RxNorm, CPT-4, and NDC. Each of these ontologies serves a distinct purpose in the representation of clinical knowledge, and together they provide a comprehensive framework for structuring and interpreting electronic health record (EHR) data. A summary of the conceptual mappings between MIMIC-IV hosp module tables and major medical ontologies, including the relevant fields and mapping strategies, is presented in Table 2.

The analysis demonstrated that MIMIC-IV is compatible with some ontologies, such as ICD and NDC, which are already embedded in the dataset through structured codes. Other ontologies, such as SNOMED CT and LOINC, require additional mapping efforts, either through intermediary vocabularies (e.g., ICD to SNOMED CT, LOINC to SNOMED CT) or through semantic annotation of free-text fields. RxNorm and CPT-4 mappings are also possible, though they may involve more complex translation processes due to differences in coding granularity and system design.

Mapping to SNOMED CT enables a richer semantic representation of diagnoses, procedures, and clinical findings, supporting use cases such as phenotyping and decision support. LOINC provides standardized identifiers for laboratory and clinical observations, which are essential for harmonizing test results across institutions. ICD remains a foundational ontology for disease classification and epidemiological analysis, while RxNorm and NDC offer complementary views of medication data - RxNorm for clinical drug concepts and NDC for product-level specificity. CPT-4, although not natively present in MIMIC-IV, offers a valuable perspective for aligning procedural data with billing and outpatient care standards.

Future work should include systematic evaluation using metrics such as mapping coverage (proportion of source fields/rows aligned to target vocabularies), alignment accuracy (manual review or gold-standard agreement), and semantic consistency. We also acknowledge partial support for some hierarchies (e.g., incomplete SNOMED CT coverage for certain clinical contexts; CPT-4 approximations via crosswalks), which can limit semantic integrity and should be made explicit when reporting results.

The conceptual mapping process also highlighted several challenges. Differences in coding systems, variations in data granularity, and the presence of ambiguous or incomplete entries can complicate the alignment process. Moreover, not all SNOMED CT hierarchies are represented in MIMIC-IV, and some ontologies, such as CPT-4, may only partially overlap with the dataset's structure and content. These limitations underscore the importance of careful validation, and context - aware mapping strategies.

Despite these challenges, the conceptual mappings outlined in this paper provide a structured foundation for future work involving the semantic improvement of MIMIC-IV. By aligning clinical data with standardized ontologies, researchers can unlock new opportunities for cross-dataset comparisons, multi-center studies, and the development of interoperable clinical decision support tools. Furthermore, ontology-based integration enhances the interpretability and reproducibility of research findings, contributing to the broader goals of transparency in biomedical informatics.

Conclusion and Future Work

The integration of standardized medical ontologies into clinical datasets like MIMIC-IV is crutial for advancing the semantic interoperability, precise analysis, and generalizability of health informatics research. Through the detailed mapping of MIMIC-IV's hosp module tables to established ontologies — including SNOMED CT, LOINC, ICD, RxNorm, CPT-4, and NDC — this study has demonstrated both the opportunities and complexities essential for harmonizing diverse clinical data. The analysis highlighted that while MIMIC-IV is structured around some ontologies, like ICD for diagnoses and NDC for medication products, incorporating other ontologies like SNOMED CT and LOINC requires specific strategies like intermediary vocabularies and text-based semantic anno-

Table 2. Conceptual mapping of MIMIC-IV `hosp` tables to medical ontologies

MIMIC-IV Table	Description	Relevant Ontologies	Mapping Approach	Key Fields Used
`diagnoses_icd`	Diagnosis codes (ICD-9/10)	ICD, SNOMED CT	Code-based (ICD → SNOMED CT)	`icd_code`, `icd_version`
`procedures_icd`	Procedure codes (ICD-9/10-PCS)	ICD, SNOMED CT, CPT-4	Code-based (ICD → CPT-4/SNOMED)	`icd_code`, `icd_version`
`prescriptions`	Medication orders	RxNorm, SNOMED CT, NDC	Code/text-based (NDC → RxNorm)	`ndc`, `gsn`, `drug`
`pharmacy`	Medication administration	RxNorm, SNOMED CT, NDC	Code/text-based	`ndc`, `gsn`, `drug`
`labevents`	Lab test results	LOINC, SNOMED CT	Text-based (label → LOINC)	`itemid`, `value`, `flag`
`d_labitems`	Lab test metadata	LOINC, SNOMED CT	Text-based	`itemid`, `label`, `fluid`
`microbiology-events`	Microbiology test results	SNOMED CT, LOINC	Text-based	`test_name`, `org_name`
`drgcodes`	Diagnosis-related groups	ICD, SNOMED CT	Text-based (description → ICD)	`drg_code`, `description`
`services`	Hospital service units	SNOMED CT	Text-based (unit → SNOMED CT)	`curr_service`, `prev_service`

tation. These mappings enable more sophisticated analyses, such as phenotype construction, and the development of interoperable AI-driven decision support systems.

The work also emphasizes several challenges, including differences in code system granularity, incomplete or ambiguous records, and the necessity of context-aware mapping approaches to preserve clinical meaning. Finally, the structured alignment presented here provides a robust foundation for future efforts to semantically improve MIMIC-IV and similar datasets. This will contribute to cross-institutional studies, enhance reproducibility, and support the scalable integration of ML models in critical care research. By embedding MIMIC-IV more deeply within widely recognized medical ontologies, this work contributes to higher-impact healthcare analytics. As a next step, future work will focus on the development of a dedicated ontology that is based on the MIMIC table schema, integrates the identified medical ontologies, and is optimized for the seamless integration of new laboratory data.

In future work, we will extend the proposed mappings by incorporating applied examples—such as phenotype extraction, and ontology-backed decision-support queries—to demonstrate the practical impact of semantic enrichment on analytic validity and portability. Complementing OMOP and FHIR conversions [18, 19], this work will clarify when ontology-level enrichment within the native schema is sufficient and when a full CDM transformation may be more appropriate. This ontology will serve as a dynamic and extensible framework to support evolving clinical research needs and data interoperability.

References

1. D. C. Angus, M. A. Kelley, R. J. Schmitz, A. White, and J. Popovich, "Current and projected workforce requirements for care of the critically ill and patients with pulmonary disease," *Jama*, vol. 284, no. 21, pp. 2762–2770, 2000.
2. J.-L. Vincent, "Is the current management of severe sepsis and septic shock really evidence based?" *PLoS medicine*, vol. 3, no. 9, p. e346, 2006.
3. M. Ross, W. Wei, and L. Ohno-Machado, ""big data" and the electronic health record," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 97–104, 2014.
4. Y. Zhang, S.-L. Guo, L.-N. Han, and T.-L. Li, "Application and exploration of big data mining in clinical medicine," *Chinese Medical Journal*, vol. 129, no. 06, pp. 731–738, 2016.
5. S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C. U. Lehmann, "Clinical data reuse or secondary use: current status and potential future progress," *Yearbook of medical informatics*, vol. 26, no. 01, pp. 38–52, 2017.
6. D. Chrimes and C. Kim, "Comparison of mimic-iii and mimic-iv for big data analytics of health informatics," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 6128–6130.
7. M. G. Kahn, D. Batson, and L. M. Schilling, "Data model considerations for clinical effectiveness researchers," *Medical care*, vol. 50, pp. S60–S67, 2012.
8. R. Platt and T. Lieu, "Data enclaves for sharing information derived from clinical and administrative data," *JAMA*, vol. 320, no. 8, pp. 753–754, 2018.
9. N. Paris, A. Lamer, and A. Parrot, "Transformation and evaluation of the mimic database in the omop common data model: development and usability study," *JMIR Medical Informatics*, vol. 9, no. 12, p. e30970, 2021.
10. F. Zeshan and R. Mohamad, "Medical ontology in the dynamic healthcare environment," *Procedia Computer Science*, vol. 10, pp. 340–348, 2012.
11. J. I. B. Arce. (2023, Apr.) Diagrama relacional base de datos mimic-iv de physionet. Ciencia de Datos, Informática Médica, Inteligencia Artificial. [Online]. Available: https://www.juanbarrios.com/diagrama-relacional-base-de-datos-mimic4-de-physionet/
12. A. D. Nicola, R. Zgheib, and F. Taglino, "Toward a knowledge graph for medical diagnosis: issues and usage scenarios," in *Semantic Models in IoT and eHealth Applications*, S. Tiwari, F. O. Rodriguez, and M. A. Jabbar, Eds. Academic Press, 2022, pp. 129–142.
13. N. Paris and A. Parrot, "Mimic in the omop common data model," *medRxiv*, pp. 2020–08, 2020.
14. N. Giesa, P. Heeren, S. Klopfenstein, A. Flint, L. Agha-Mir-Salim, A. Poncette, F. Balzer, and S. Boie, "Mimic-iv as a clinical data schema," in *Challenges of Trustable AI and Added-Value on Health*. IOS Press, 2022, pp. 559–560.
15. S. Jupp, T. Burdett, C. Leroy, and H. E. Parkinson, "A new ontology lookup service at embl-ebi." *SWAT4LS*, vol. 2, pp. 118–119, 2015.
16. A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.
17. M. A. U. Zaman, "Assessing different diagnoses in mimic-iv v2. 2 and mimic-iv-ed datasets," *Archives of Proteomics and Bioinformatics*, vol. 4, no. 1, pp. 1–5, 2024.
18. N. Paris, A. Lamer, and A. Parrot, "Transformation and evaluation of the mimic database in the omop common data model: Development and usability study," *JMIR Medical Informatics*, vol. 9, no. 12, p. e30970, 2021. [Online]. Available: https://medinform.jmir.org/2021/12/e30970/

19. A. M. Bennett, A. E. Johnson, T. J. Pollard *et al.*, "Mimic-iv on fhir: converting a decade of in-patient data to an accessible fhir dataset," *JAMIA Open*, vol. 6, no. 1, p. ooac115, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10018258/
20. M. Ivanović and Z. Budimac, "An overview of ontologies and data resources in medical domains," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5158–5166, 2014.
21. D. Lee, N. de Keizer, F. Lau, and R. Cornet, "Literature review of snomed ct use," *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e11–e19, 2014.
22. P. L. Schuyler, W. T. Hole, M. S. Tuttle, and D. D. Sherertz, "The umls metathesaurus: representing different views of biomedical concepts," *Bulletin of the Medical Library Association*, vol. 81, no. 2, p. 217, 1993.
23. N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute *et al.*, "Bioportal: ontologies and integrated data resources at the click of a mouse," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W170–W173, 2009.
24. C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook *et al.*, "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.
25. D. J. Vreeman, J. Hook, and B. E. Dixon, "Learning from the crowd while mapping to loinc," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1205–1211, 2015.
26. O. WHO, "International classification of diseases," *WHO [Internet]*, 1992.
27. J. Hirsch, G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle, and L. Manchikanti, "Icd-10: history and context," *American Journal of Neuroradiology*, vol. 37, no. 4, pp. 596–599, 2016.
28. S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "Rxnorm: prescription for electronic drug information exchange," *IT professional*, vol. 7, no. 5, pp. 17–23, 2005.
29. L. Peters and O. Bodenreider, "Using the rxnorm web services api for quality assurance purposes," in *AMIA Annual Symposium Proceedings*, vol. 2008, 2008, p. 591.
30. P. L. Elkin and S. H. Brown, "Current procedural terminology," in *Terminology, Ontology and their Implementations*. Springer, 2023, pp. 367–370.
31. T. Faciszewski, R. Jensen, and R. L. Berg, "Procedural coding of spinal surgeries (cpt-4 versus icd-9-cm) and decisions regarding standards: a multicenter study," *Spine*, vol. 28, no. 5, pp. 502–507, 2003.
32. J. J. Guo, M. C. Diehl, B. G. Felkey, J. T. Gibson, and K. N. Barker, "Comparison and analysis of the national drug code systems among drug information databases," *Drug information journal: DIJ/Drug Information Association*, vol. 32, pp. 769–775, 1998.
33. N. Kathe, V. Vivek, N. Agrawal, and R. Aparasu, "Rwd51 comparing the food and drug administration national drug code directory and redbook to identify prescription records," *Value in Health*, vol. 25, no. 7, p. S585, 2022.